

https://dspace.org/

Bioinformatics-driven development of a queryable cardiometabolic database and its application in a biological setting

Hendry, Liesl Mary

2017

http://hdl.handle.net/10539/23508

Hendry, Liesl Mary (2017) Bioinformatics-driven development of a queryable cardiometabolic database and its application in a biological setting, University of the Witwatersrand, Johannesburg, <http://hdl.handle.net/10539/23508>

Downloaded from a DSpace repository. https://dspace.org/

BIOINFORMATICS-DRIVEN DEVELOPMENT OF A QUERYABLE CARDIOMETABOLIC DATABASE AND ITS APPLICATION IN A BIOLOGICAL SETTING

Liesl Mary Hendry

Supervisor: Dr Zané Lombard

A thesis submitted to the Faculty of Science, University of the Witwatersrand, Johannesburg in fulfilment of the requirements for the degree of Doctor of Philosophy.

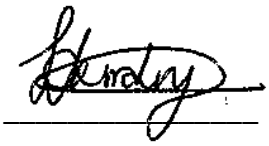
June 2017, Johannesburg

DECLARATION

I, **LIESL MARY HENDRY** (**590448**), am a student registered for the degree of Doctor of Philosophy in the academic year 2017.

I hereby declare the following:

- I am aware that plagiarism (the use of someone else's work without their permission and/or without acknowledging the original source) is wrong.
- I confirm that the work submitted for assessment for the above degree is my own unaided work except where explicitly indicated otherwise and acknowledged.
- I have not submitted this work before for any other degree or examination at this or any other University.
- The information used in the thesis has not been obtained by me while employed by, or working under the aegis of, any person or organisation other than the University.
- I have followed the required conventions in referencing the thoughts and ideas of others.
- I understand that the University of the Witwatersrand may take disciplinary action against me if there is a belief that this is not my own unaided work or that I have failed to acknowledge the source of the ideas or words in my writing.

Signature:  **2nd** day of **June 2017**

ABSTRACT

As sequencing and genotyping technologies are advancing, larger and more complex sets of biological data are being produced. Databases can be used to efficiently store and manage the data. Typically, publicly available datasets are accessed through web browsers that offer a user-friendly interface to a database, making complex queries simple to execute. However, research project-specific data are not commonly stored in this way. In this research, a database (designed in MySQL) and accompanying interface (developed using PHP, HTML and CSS) has been designed for the storage and querying of the quality controlled data from the current project using Metabochip-genotyped Birth to Twenty (Bt20) cohort participants and their female caregivers. Users can easily access the data to generate summary statistics on the phenotype data and download phenotype, single nucleotide polymorphism (SNP) annotation and association analysis data that match user-supplied criteria.

Some of the data from the database was used to investigate the genetics of blood pressure (BP) in black South African individuals. Hypertension is a major risk factor for cardiovascular diseases (CVDs). BP variation is known to have a genetic component, but genetic studies in indigenous Africans have been limited. Association analysis, carried out in a merged sample of caregivers and participants, pointed to novel regions of interest in the *NOS1AP* (DBP and SBP), *MYRF* (SBP) and *POC1B* (SBP) genes and two intergenic regions (*DACH1*|*LOC440145* (DBP and SBP) and *INTS10*|*LPL* (SBP)). Two SNPs in the *MYRF* gene met the calculated "array-wide" significance threshold ($p<6.7x10^{-7}$ for the merged dataset) for multiple testing.

Genotype imputation is a useful addition to association studies to increase the SNP panel for association testing. An investigation into the efficiency of imputation in this dataset using a mixed population reference panel was carried

out. Imputation was achieved with high confidence in all genes, but a more detailed view of the region was only seen in *NOS1AP* (DBP and SBP in both the merged and female caregiver datasets) and *POC1B* (Bt20 participant dataset only).

Overall, the research contributed a useful tool for the efficient management of project-specific biological data. The analysis and genotype imputation, which is a promising tool in future studies in this or other African datasets, also provided some insight into the genetics of blood pressure in black South Africans with further functional and replication studies in larger samples required to confirm and explain the findings.

To my family and friends who have supported me on this journey and in memory of my late Nana who was always interested in the lives of her grandchildren

ACKNOWLEDGEMENTS

I wish to express my sincere gratitude and appreciation to the following people and institutions that have made this PhD possible:

To my family and friends for their constant love, support and guidance throughout my studying years and for always taking interest in what I was doing, even if they didn't understand any of it. Thank you especially to my Mom (Gill) and Dad (Keith) for putting up with having a student under their roof for so many years, for never discouraging me from studying further and for giving me the space to follow my passion and reach my goals. Thank you also to my Mom for the stats advice, afternoon chats over tea and for always being a listening ear and source of advice. To my brother (Neil) and sister-in-law (Kendra) for their constant encouragement and moral support and for always being incredible role models. I also acknowledge our Heavenly Father for my God-given talents and abilities and for the peace and guidance through the good and bad times of the past three years.

To my supervisor, Zané Lombard, for her guidance and support through my research, for always having the best interests of her students at heart, for offering invaluable advice and feedback, for enabling me to see other parts of the world over the last few years, for helping me to grow as a young researcher and for believing in me and giving me the chance to get into the field when I first arrived in Joburg.

To Michèle Ramsay and the rest of the special people (too many to mention!) at the Sydney Brenner Institute for Molecular Bioscience for providing such a wonderful working environment, for the friendships formed and for all the good conversations and fun times over Friday tea, lunch and other social events. To

Ananyo Choudhury, Phelelani Mpangase, Scott Hazelhurst and Shaun Aron for their “technical” and Bioinformatics-related help, big and small, and for the patience while offering the help! To students, past and present, for the many interesting conversations, the fun, laughter and banter and for riding this journey as a team. A special thank you to:

- Venesa Sahibdeen, my “Metabochip co-worker” and Bristol travel buddy, for all the initial work she did on the Metabochip study, for the quality control and other work we were able to achieve together and for the help and support when it was needed.
- Andrew Ndhlovu, my programming buddy, for the invaluable advice and guidance given when I knew nothing about PHP and designing a web interface, for sharing in the excitement of my database and for often believing in me more than I did!
- Richard Munthali, my MCB buddy, for enduring the things we didn’t want to do together and for the help and advice when needed.
- Thandiswa Ngcungcu, for her friendship and sharing in this journey first as Masters students and then as PhD students.

To Shane Norris and colleagues at the Developmental Pathways for Health Research Unit for the use of the phenotype data in this study and to the participants of the Birth to Twenty study for their ongoing participation in the study.

To experts from around the world for sharing their knowledge and expertise and for allowing me to learn all that I’ve learnt along the way. A special mention to the people at the MRC Integrative Epidemiology Unit of the University of Bristol (especially Tom Gaunt and Nic Timpson) for their help during our GEoCoDE research exchange, to the trainers of the Wellcome Trust Advanced Course that I attended in 2014 (Jonathan Marchini, Andrew Morris and Heather Cordell) for teaching me so much and to Ele Zeggini and Daniel Suveges from the Wellcome

Trust Sanger Institute for the advice and help on the post-analysis QC. The willingness of almost total strangers to offer their expertise so freely has been humbling.

To the National Research Foundation (NRF) for financial support in the form of an Innovation Doctoral Scholarship. The financial assistance of the NRF towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at, are those of the author and are not necessarily to be attributed to the NRF.

RESEARCH OUTPUTS

Original article

Hendry, L.M. et al., 2016. Insights into the genetics of blood pressure in black South African individuals: the Birth to Twenty cohort. *Journal of Hypertension*. Under review.

Conference outputs (including research days)

Oral presentations:

Conference	Date (Place)	Presentation	Authors
16th Biennial Congress of the Southern African Society for Human Genetics (SASHG)	16-19 August 2015 (Centurion, SA)	The genetics of blood pressure in black South African individuals from the Birth to Twenty cohort	Liesl M Hendry, Venesa Pillay, Shane A Norris, Zané Lombard
Molecular Biosciences Research Thrust Research Day 2015	3 December 2015 (Wits, Johannesburg, SA)	Association of genetic variation with systolic and diastolic blood pressure in black South Africans	Liesl M Hendry, Venesa Pillay, Shane A Norris, Michele Ramsay, Zané Lombard (of the AWI-Gen study and as members of the H3Africa Consortium)

Poster presentations:

Conference	Date (Place)	Presentation	Authors
7th H3Africa Consortium Meeting	11-14 October 2015 (Washington DC and Bethesda Maryland, USA)	Association of genetic variation with systolic and diastolic blood pressure in black South Africans	Liesl M Hendry, Venesa Pillay, Shane A Norris, Michele Ramsay, Zané Lombard (of the AWI-Gen study and as members of the H3Africa Consortium)
7th Cross-Faculty Symposium	1-2 March 2016 (Wits, Johannesburg, SA)	Association of genetic variation with systolic and diastolic blood pressure in black South Africans	Liesl M Hendry, Venesa Pillay, Shane A Norris, Michele Ramsay, Zané Lombard (of the AWI-Gen study and as members of the H3Africa Consortium)
Molecular Biosciences Research Thrust Research Day 2016	8 December 2016 (Wits, Johannesburg, SA)	Bioinformatics-driven development of a queryable cardiometabolic database	Liesl M Hendry, Zané Lombard

CONTENTS

DECLARATION ii

ABSTRACT iii

ACKNOWLEDGEMENTS vi

RESEARCH OUTPUTS ix

CONTENTS xi

LIST OF FIGURES xv

LIST OF TABLES xviii

LIST OF SYMBOLS xix

ABBREVIATIONS xx

PREFACE xxxv

The current study and its significance xxxv

Aims and objectives xxxviii

Outline of the thesis xxxix

Chapter 1: INTRODUCTION 1

1.1 The “omics” and “big data” 1

1.1.1 Advancements in DNA sequencing and chip technologies 1

1.2 Data management: benefits and challenges 2

1.3 Databases 3

1.3.1 Online databases 3

1.4 Biological focus: Cardiovascular diseases and hypertension 7

1.4.1 Ethnic differences in hypertension prevalence 8

1.4.2 Non-genetic risk factors 10

1.4.3 Genetic risk factors 11

BP genetics in African individuals 24
1.5 Genotype imputation 26
1.5.1 Benefits of imputation 27
1.5.2 IMPUTE and IMPUTE2 28
1.5.3 Reference panel 31
Mixed/"cosmopolitan" reference panels 31
HapMap reference panels 32
1000 Genomes reference panels 33
1.5.4 Factors affecting imputation performance and accuracy 34
Reference panel size 34
Study sample size and SNP density 35
LD and MAF 35
Other factors 36
1.5.5 Imputation in African populations 36
1.5.6 Pre-phasing 38
Chapter 2: STUDY PARTICIPANTS AND DATA QUALITY CONTROL 39
2.1 Study participants 39
2.2 Phenotyping 39
2.3 DNA sample preparation and genotyping 40
2.4 Data quality control 41
2.4.1 Genotype data quality control 41
Initial QC steps 41
SNP QC 42
Sample QC 44
2.4.2 Phenotype data quality control 46

2.5 Merging of datasets and update to Build 37 46
2.6 Post-quality control data characteristics 47
2.6.1 Descriptive statistics 47
2.6.2 Population structure 47
Chapter 3: DESIGN OF A QUERYABLE CARDIOMETABOLIC DATABASE 51
3.1 Database construction 51
3.1.1 Database content 52
3.1.2 Database design evaluation 53
3.2 Web interface 55
3.2.1 Access and security 55
3.2.2 Function 56
Summary statistics 56
Data download 63
Genotype data 74
Additional features 75
3.3 Expandability of the database 76
3.4 Use by other research groups 77
3.5 Discussion 77
Chapter 4: INSIGHTS INTO THE GENETICS OF BLOOD PRESSURE IN BLACK SOUTH AFRICANS 82
4.1 Post-analysis quality control 83
4.2 Manuscript 84
Chapter 5: AN INVESTIGATION INTO GENOTYPE IMPUTATION OF METABOCHIP DATA IN BLACK SOUTH AFRICANS USING A MIXED REFERENCE PANEL 127
5.1 Methodology 127

5.1.1 Preparation of files for imputation 127
5.1.2 Pre-phasing 128
5.1.3 Imputation 128
5.1.4 Association analysis 128
5.1.5 Result visualisation 129
5.2 Results 130
5.2.1 Overall assessment of accuracy and yield 130
5.2.2 Merged dataset 132
5.2.3 Individual datasets 136
5.3 Discussion 138
Chapter 6: GENERAL CONCLUSIONS 144
REFERENCES 154
WEB REFERENCES 170
APPENDIX A – Ethics certificates, relevant agreements/ letters and consent forms 171
APPENDIX B – Script for conversion of GenomeStudio forward report files into transposed PLINK files 185
APPENDIX C – PLINK, SMARTPCA GEMMA, SHAPEIT, IMPUTE2, SNPTEST and R commands for chapters 2, 4 and 5 188
APPENDIX D – MySQL code for creation of database tables 203
APPENDIX E – Python code for input of data into MySQL tables 210
APPENDIX F – MetaboBTT README 221
APPENDIX G – Post-analysis QC: Q-Q plots and genomic inflation factors 226

LIST OF FIGURES

Figure 2.1 The steps involved in the genotype data QC process 43

Figure 2.2 PCA plot of population structure in the female caregivers and other African populations 49

Figure 2.3 PCA plot of population structure in the Bt20 participants and other African populations 50

Figure 3.1 The relational model for MetaboBTT 54

Figure 3.2 Home page of the MetaboBTT Database web interface. 56

Figure 3.3 Summary statistics, data download and genotype data landing pages. 57

Figure 3.4 Layout of the basic count page. 58

Figure 3.5 An example of the input and output of a basic count query. 58

Figure 3.6 Layout of the complex count page. 59

Figure 3.7 An example of the input and output of a complex count query. 60

Figure 3.8a Layout of the average/minimum/maximum page where the user can specify all/male/female individuals. 61

Figure 3.8b Layout of the average/minimum/maximum page where the user can specify specific individuals by uploading a file of Individual IDs..... 62

Figure 3.9 An example of the input and output of an average/minimum/maximum query..... 62

Figure 3.10a Layout of the female caregiver phenotype data download page where the user can specify all individuals 64

Figure 3.10b Layout of the female caregiver phenotype data download page where the user can specify specific individuals by uploading a file of Individual IDs..... 65

Figure 3.11a Layout of the Bt20 participant phenotype data download page where the user can specify all/male/female individuals 66

Figure 3.11b Layout of the Bt20 participant phenotype data download page where the user can specify specific individuals by uploading a file of Individual IDs..... 67

Figure 3.12 An example of the input and output of a phenotype data download query 68

Figure 3.13 Layout of the Metabochip data download page. 69

Figure 3.14a Example one of the input and output of a Metabochip data download query. 70

Figure 3.14b Example two of the input and output of a Metabochip data download query. 71

Figure 3.15 Layout of the association analysis data download page. 72

Figure 3.16 An example of the input and output of an association analysis data download query. 73

Figure 5.1 The proportion of well-imputed SNPs (info metric ≥ 0.4) in different MAF bins for each imputation scenario considered. 131

Figure 5.2a Imputation in the merged dataset resulted in a more detailed view/enrichment of the region where an association signal was observed before imputation for *NOS1AP* (SBP and DBP). 132

Figure 5.2b Imputation in the merged dataset didn't result in a more detailed view/enrichment of the region where an association signal was observed before imputation for *MYRF* or *POC1B*.. 133

Figure 5.3a Imputation in the individual datasets resulted in a more detailed view/enrichment of the region where an association signal was observed before imputation for *NOS1AP* (SBP and DBP) in the female caregivers 136

Figure 5.3b Imputation in the individual datasets resulted in a very slight enrichment of the region where an association signal was observed before imputation for *POC1B* (SBP) in the Bt20 participants and no enrichment for *POC1B* (SBP) in the female caregivers 137

LIST OF TABLES

Table 1.1 A selection of useful internet accessed databases related to human genes and diseases..... 5

Table 1.2 Findings from some of the main BP/hypertension genetic studies in non-Africans 14

Table 2.1 Descriptive statistics of the individuals remaining after QC in the individual datasets. 48

Table 5.1 Significance thresholds for post-imputation analysis. 129

Table 5.2 Number of SNPs pre- and post-imputation and with info metric ≥ 0.4 for each imputation scenario considered. 131

Table 5.3a Several imputed SNPs in the *NOS1AP* gene associated with DBP in the merged and female caregiver datasets 134

Table 5.3b Several imputed SNPs in the *NOS1AP* gene associated with SBP in the merged and female caregiver datasets 135

Table 5.4. A few imputed SNPs in the *POC1B* gene associated with SBP in the Bt20 participant dataset 138

LIST OF SYMBOLS

β	beta
cm	centimetre
g	gram
kg	kilogram
$kg.m^{-2}$	kilogram per metre squared
m^2	metres squared
mmHg	millimetres of mercury
$ng.\mu l^{-1}$	nanogram per microliter
p	p-value

ABBREVIATIONS

ABCC10	ATP binding cassette subfamily C member 10
ABLIM3	actin binding LIM protein family member 3
ACE	angiotensin-converting enzyme
ADAMTS5	ADAM metallopeptidase with thrombospondin type 1 motif 5
ADAMTS9	ADAM metallopeptidase with thrombospondin type 1 motif 9
ADM	adrenomedullin
ADRB1	adrenoceptor beta 1
AF	allele frequency
AGEN-BP	Asian Genetic Epidemiology Network Blood Pressure
AGT	angiotensinogen
ALDH2	aldehyde dehydrogenase 2 family (mitochondrial)
AMH	anti-Mullerian hormone
ANP	atrial natriuretic peptide
ARHGAP24	Rho GTPase activating protein 24
ARHGEF12	Rho guanine nucleotide exchange factor 12
ARL6IP6	ADP ribosylation factor like GTPase 6 interacting protein 6
ARRDC3	arrestin domain containing 3

ATG7	autophagy related 7
ATP2B1	ATPase plasma membrane Ca2+ transporting 1
ATXN2	ataxin 2
AWI-Gen	African Wits-INDEPTH Partnership for the GENomic study of body composition and cardiometabolic risk
BAT2	proline rich coiled-coil 2A
BAT2D1	proline rich coiled-coil 2C
BAT5	abhydrolase domain containing 16A
BLK	BLK proto-oncogene, Src family tyrosine kinase
BMI	body mass index
BOC	BOC cell adhesion associated, oncogene regulated
BP	blood pressure
BSO	black Sowetans
Bt20	Birth to Twenty
Bt20_CG	female caregivers
Bt20_yr1718	Bt20 participants
C3orf17	chromosome 3 open reading frame 17
C5orf23	chromosome 5 open reading frame 23
C10orf107	chromosome 10 open reading frame 107
C10orf114	chromosome 10 open reading frame 114

C17orf82	chromosome 17 open reading frame 82
C18orf1	chromosome 18 open reading frame 1
C21orf91	chromosome 21 open reading frame 91
C21orf94	chromosome 21 open reading frame 94
CA	California
CACNA1D	calcium voltage-gated channel subunit alpha1 D
CACNA1H	calcium voltage-gated channel subunit alpha1 H
CACNB2	calcium voltage-gated channel auxiliary subunit beta 2
CAD	coronary artery disease
CAPZA1	capping actin protein of muscle Z-line alpha subunit 1
CASZ1	castor zinc finger 1
CDC123	cell division cycle 123
CDH13	cadherin 13
CDH17	cadherin 17
CDK6	cyclin dependent kinase 6
CHARGE	Cohorts for Heart and Aging Research in Genomic Epidemiology
CHIC2	cysteine rich hydrophobic domain 2
CHST12	carbohydrate sulfotransferase 12
CMD	cardiometabolic disease

CNNM2	cyclin and CBS domain divalent metal cation transport mediator 2
COGENT	Continental Origins and Genetic Epidemiology Network
CRYAA	crystallin alpha A
CSNK1G3	casein kinase 1 gamma 3
CSK	c-src tyrosine kinase
CSV	comma separated values
CVD	cardiovascular disease
CXADR	coxsackie virus and adenovirus receptor
CYB5R2	cytochrome B5 reductase 2
CYP1A1	cytochrome P450 family 1 subfamily A member 1
CYP1A2	cytochrome P450 family 1 subfamily A member 2
CYP11B2	cytochrome P450 family 11 subfamily B member 2
CYP17A1	cytochrome P450 family 17 subfamily A member 1
CYP19A1	cytochrome P450 family 19 subfamily A member 1
CYP21A2	cytochrome P450 family 21 subfamily A member 2
DACH1	dachshund family transcription factor 1
DBP	diastolic blood pressure
DBH	dopamine beta-hydroxylase
DNA	deoxyribonucleic acid

DNAH10	dynein axonemal heavy chain 10
DOT1L	DOT1 like histone lysine methyltransferase
DRD1	dopamine receptor D1
DUSP6	dual specificity phosphatase 6
DXA	dual-energy X-ray absorptiometry
EAGLE	Early Genetics and Lifecourse Epidemiology
EBF1	early B-cell factor 1
EDN3	endothelin 3
ELAVL3	ELAV like RNA binding protein 3
EMBL-EBI	European Bioinformatics Institute
EML2	echinoderm microtubule associated protein like 2
ENaC	epithelial sodium channel
eNOS	endothelial nitric oxide synthase
ENPEP	glutamyl aminopeptidase
EVX	even-skipped homeobox
EVX1	even-skipped homeobox 1
FAM5C	BMP/retinoic acid inducible neural specific 3
FAM19A5	family with sequence similarity 19 member A5, C-C motif chemokine like
FAT3	FAT atypical cadherin 3

FBN1	fibrillin 1
FES	FES proto-oncogene, tyrosine kinase
FGD5	FYVE, RhoGEF and PH domain containing 5
FGF5	fibroblast growth factor 5
FIGN	fidgetin, microtubule severing factor
FLJ13197	hypothetical FLJ13197
FLJ32810	hypothetical FLJ32810
FLJ46257	hypothetical FLJ46257
FMO4	flavin containing monooxygenase 4
FURIN	furin, paired basic amino acid cleaving enzyme
GATA4	GATA binding protein 4
GBPGEN	Global Blood Pressure Genetics
GIPR	gastric inhibitory polypeptide receptor
GNAS	GNAS complex locus
GOSR2	golgi SNAP receptor complex member 2
GPR98	G protein-coupled receptor 98
GRB14	growth factor receptor bound protein 14
GRIK4	glutamate ionotropic receptor kainate type subunit 4
GRK4	G protein-coupled receptor kinase 4
GUCY1A3	guanylate cyclase 1 soluble subunit alpha

GUCY1B3	guanylate cyclase 1 soluble subunit beta
GWAS	genome-wide association study/studies
H19	H19, imprinted maternally expressed transcript (non-protein coding)
H3Africa	Human Heredity and Health in Africa
HC	hip circumference
HDAC9	histone deacetylase 9
HDL	high density lipoprotein
HFE	hemochromatosis
HIVEP3	human immunodeficiency virus type I enhancer binding protein 3
HLA-DQB1	major histocompatibility complex, class II, DQ beta 1
HMG20A	high mobility group 20A
HMM	hidden Markov model
HOTTIP	HOXA distal transcript antisense RNA
HOXA	homeobox A cluster
HOXC	homeobox C cluster
HRH1	histamine receptor H1
HWE	Hardy-Weinberg equilibrium
IBD	identity by decent
IBS	identity by state

ICBP	International Consortium for Blood Pressure
IGFBP1	insulin like growth factor binding protein 1
IGFBP3	insulin like growth factor binding protein 3
INDEPTH	International Network for the Demographic Evaluation of Populations and their Health in low and middle income countries
INSR	insulin receptor
INTS10	integrator complex subunit 10
IPO7	importin 7
ITGA9	integrin subunit alpha 9
ITGA11	integrin subunit alpha 11
JAG1	jagged 1
KCNK3	potassium two pore domain channel subfamily K member 3
KCNQ1	channel, voltage gated KQT-like subfamily Q, member 1
LD	linkage disequilibrium
LDL	low density lipoprotein
LFNG	LFNG O-fucosylpeptide 3-beta-N-acetylglucosaminyltransferase
LITAF	lipopolysaccharide induced TNF factor
LMIC	low- and middle-income countries

LPL	lipoprotein lipase
LRRC10B	leucine rich repeat containing 10B
LSP1	lymphocyte-specific protein 1
LWK	Luhya in Webuye, Kenya
MA	Massachusetts
MAF	minor allele frequency
MAN2B2	mannosidase alpha class 2B member 2
MAP	mean arterial pressure
MAP4	microtubule associated protein 4
MCMC	Markov chain Monte Carlo
MDM4	MDM4, p53 regulator
MECOM	MDS1 and EVI1 complex locus
MED13L	mediator complex subunit 13 like
MI	myocardial infarction
miRNA	micro-ribonucleic acid
MKK	Maasai in Kinyawa, Kenya
MOV10	Mov10 RISC complex RNA helicase
MTHFR	methylenetetrahydrofolate reductase
MYRF	myelin regulatory factor
NAT1	N-acetyltransferase 1

NAT2	N-acetyltransferase 2
NCAPH	non-SMC condensin I complex subunit H
NCBI	National Centre for Biotechnology Information
NCD	non-communicable disease
NFAT5	nuclear factor of activated T-cells 5
NHLS	National Health Laboratory Service
NO	nitric oxide
NOS1AP	nitric oxide synthase 1 (neuronal) adaptor protein
NOS3	nitric oxide synthase 3
nNOS	neuronal nitric oxide synthase
NPPA	natriuretic peptide A
NPPB	natriuretic peptide B
NPR3	natriuretic peptide receptor 3
NRF	National Research Foundation
NT5C2	5'-nucleotidase, cytosolic II
NUCB2	nucleobindin 2
OR	odds ratio
OSR1	odd-skipped related transciption factor 1
PAGE	Population Architecture using Genomics and Epidemiology
PC	principal component

PCA	principal component analysis
PDE1A	phosphodiesterase 1A
PDE3A	phosphodiesterase 3A
PIK3CG	phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit gamma
PLCD3	phospholipase C delta 3
PLCE1	phospholipase C epsilon 1
PLEKHA7	pleckstrin homology domain containing A7
PLEKHG1	pleckstrin homology and RhoGEF domain containing G1
PLEKHH2	pleckstrin homology, MyTH4 and FERM domain containing H2
PLEKHJ1	pleckstrin homology domain containing J1
PMS1	PMS1 homolog 1, mismatch repair system component
PNPT1	polyribonucleotide nucleotidyltransferase 1
POC1B	POC1 centriolar protein B
PP	pulse pressure
PRDM6	PR/SET domain 6
PRKAG2	protein kinase AMP-activated non-catalytic subunit gamma 2
PSD3	pleckstrin and Sec7 domain containing 3
PSMC3	proteasome 26S subunit, ATPase 3

PSMD5	proteasome 26S subunit, non-ATPase 5
QC	quality control
Q-Q	quantile–quantile
RAAS	renin-angiotensin-aldosterone system
RAB8B	RAB8B, member RAS oncogene family
RAPSN	receptor associated protein of the synapse
RELA	RELA proto-oncogene, NF-kB subunit
RGS5	regulator of G-protein signaling 5
RSPO3	R-spondin 3
SBP	systolic blood pressure
SCARB1	scavenger receptor class B member 1
SCL4A7	solute carrier family 4 member 7
SD	standard deviation
SEB	southeastern Bantu-speakers
SELE	selectin E
SETBP1	SET binding protein 1
SF3A2	splicing factor 3a subunit 2
SGK269	pseudopodium enriched atypical kinase 1
SH2B3	SH2B adaptor protein 3
SH3TC2	SH3 domain and tetratricopeptide repeats 2

SIB	Swiss Institute for Bioinformatics
SIK1	salt inducible kinase 1
SIPA1	signal-induced proliferation-associated 1
SKAT	SNP-set (Sequence) Kernel Association Test
SLC4A5	solute carrier family 4 member 5
SLC4A7	solute carrier family 4 member 7
SLC7A1	solute carrier family 7 member 1
SLC22A7	solute carrier family 22 member 7
SLC22A18	solute carrier family 22 member 18
SLC24A4	solute carrier family 24 member 4
SLC39A8	solute carrier family 39 member 8
SLC39A13	solute carrier family 39 member 13
STK33	serine/threonine kinase 33
STK39	serine/threonine kinase 39
SNP	single nucleotide polymorphism
SOX6	SRY-box 6
SQL	Structured Query Language
SSA	sub-Saharan Africa
ST7L	suppression of tumorigenicity 7 like
SUB1	SUB1 homolog, transcriptional regulator

SWB	southwestern Bantu-speakers
SYNPO2L	synaptopodin 2 like
SYT7	synaptotagmin 7
T2D	type 2 diabetes
TAX1BP1	Tax1 binding protein 1
TBC1D1	TBC1 domain family member 1
TBX2	T-box 2
TBX3	T-box 3
TBX5	T-box 5
TMEM133	transmembrane protein 133
TNNT3	troponin T3, fast skeletal type
TRAFD1	TRAF-type zinc finger domain containing 1
TRIM36	tripartite motif containing 36
TRPV4	transient receptor potential cation channel subfamily V member 4
TTBK1	tau tubulin kinase 1
UC	University of California
UK	United Kingdom
ULK3	unc-51 like kinase 3
ULK4	unc-51 like kinase 4

UMOD	uromodulin
URL	Uniform Resource Locator
USA	United States of America
VCL	vinculin
VNN1	vanin 1
WC	waist circumference
WHR	waist-to-hip ratio
WNK1	WNK lysine deficient protein kinase 1
YRI	Yoruba in Ibadan, Nigeria
YWHA7	tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein 7
ZC3HC1	zinc finger C3HC-type containing 1
ZNF318	zinc finger protein 318
ZNF652	zinc finger protein 652

PREFACE

The current study and its significance

The current study forms part of a larger project focused on investigating the genomic and environmental risk factors for cardiometabolic disease (CMD) in Africans (Lombard et al., 2012; Pillay et al., 2015; Ramsay et al., 2016). The huge burden of chronic non-communicable diseases (NCDs) and specifically cardiometabolic or cardiovascular diseases (CVDs) in black or African individuals provides motivation for these investigations. The study makes use of a dataset consisting of individuals from the Birth to Twenty (Bt20) cohort genotyped using the Metabochip.

The Bt20 cohort forms the basis of the largest longitudinal study on child and adolescent health and development in Africa. The cohort initially enrolled 3273 babies born as single births to women residing in Soweto, Johannesburg during a 7-week enrolment period between April and June 1990 (Richter et al., 2007). The race groups represented in the cohort are an approximate representation of the South African population (78% African, 6% White, 12% Coloured and 4% Indian) (Richter et al., 2007), although only individuals of African ancestry are included in this study. Data, including body composition and cardiometabolic data, have been collected regularly at various stages since then, with a decrease in individuals from the initial number largely due to migration and resultant loss of follow up (Richter et al., 2007).

The Metabochip is a custom genotyping array that allows for the genotyping of almost 200,000 single nucleotide polymorphisms (SNPs) known to influence cardiometabolic traits. The chip was designed using contributions from several large consortia and contains SNPs specific to various traits of primary interest including type 2 diabetes (T2D), fasting glucose, coronary artery disease and

myocardial infarction (CAD/MI), low density lipoprotein (LDL) cholesterol, high density lipoprotein (HDL) cholesterol, triglycerides, body mass index (BMI), systolic blood pressure (SBP) and diastolic blood pressure (DBP), QT interval, and waist-to-hip ratio (WHR) adjusted for BMI. Also contained on the chip are SNPs specific to traits of secondary interest, namely fasting insulin, 2-hour glucose, glycated hemoglobin, T2D age of diagnosis, early onset T2D, waist circumference (WC) adjusted for BMI, height, body fat percentage, total cholesterol, platelet count, mean platelet volume, and white blood cell count. The SNPs on the chip are mostly replication (to follow up top independent association signals for each of the traits) and fine-mapping SNPs (to fine-map 257 loci associated at genome-wide significance SNPs in preliminary analyses for one or more of the traits). (Voight et al., 2012)

The data being used and produced in this project includes Bt20 phenotype (from multiple data collection time points), Metabochip SNP annotation and resulting association analysis outputs. This constitutes a significant volume and has, up until now, been stored in basic Excel spreadsheets. This data needs to be organised in a more user-friendly format that is easily accessible and usable by relevant individuals, and also queryable for useful information. Relational databases are one such tool that can be used to achieve this. Publically available data is usually accessed via a web interface linked to a database, but project-specific data is not often stored in this way. A major part of the current study therefore involved the development of a queryable cardiometabolic database, and accompanying user interface, which stores the longitudinal phenotype, SNP annotation and association analysis data. Having the data organised in a database will facilitate access to the data and extraction of specific data and useful information/summary statistics by individuals within the research group. The stored association analysis data will also be a useful reference to inform future studies. In addition, the database will provide a model/framework for the development of other similar databases.

This study also involved an investigation into one aspect of CMD risk using the data stored in the database, namely the genetics of blood pressure (BP) or hypertension. Important advances are being made in genomic research to discover the genetics of BP/hypertension, but the studies have largely been conducted in non-African populations. This is, in fact, the case in most genetic association studies. African populations are genetically more diverse than non-African populations and tend to have a greater SNP density and lower linkage disequilibrium (LD) between SNPs (Remm & Metspalu, 2002; Tishkoff & Verrelli, 2003). The level of genetic diversity is also known to be greatly decreased as one moves further away from Africa (Tishkoff et al., 2009). Genetic association studies involve the identification of either a SNP that has a direct association with the phenotype in question (i.e. it is the causal variant) or a SNP that has an indirect association and is in LD with the causal variant (Lewis & Knight, 2012). Due to the differences in LD structure between African and other populations and the fact that different SNPs may be in LD with the causal SNP in the different populations, markers identified in other populations as being associated with disease susceptibility may in fact have no significant association in an African population. Indeed, in the few studies on various diseases that have been conducted in Africans, the associated SNPs are often different to those reported in other populations. Despite this, the African population still remains largely under-represented in genetic association studies. Some studies on blood pressure and hypertension have been reported in individuals of "African ancestry", but have mainly been conducted in African-Americans. African-Americans are, however, also genetically very different to native Africans, due to the European admixture, and therefore cannot be accurately used as a reference for what is found in "true African" individuals (i.e. those individuals of pure African ancestry who have not undergone migration and been influenced genetically by non-African populations). Studies have also suggested that there is great genetic diversity between subgroups within the African population itself (Tishkoff et al., 2009), highlighting the importance of not only conducting genetic

association studies in separate populations, but in distinct subgroups too. This study will hopefully provide some insight into the genetics of BP/hypertension in a black population of “true African” ancestry in South Africa.

The Metabochip was developed from data obtained from studies in European populations. The set of SNPs available for testing therefore might not be ideal or adequate for use in an African population and associations or exact causal variants may be undetectable with the dataset as it is. Genotype imputation has become a useful addition to genetic association studies to recover some of the missing or ungenotyped data. African populations can be challenging to impute due to their genetic diversity and lower levels of LD (Howie et al., 2011; Huang et al., 2011). An investigation was therefore deemed necessary in this study to assess the effectiveness of genotype imputation in this African dataset. This serves as an extension to the investigation into the genetics of BP/hypertension in our black South African population. Successful imputation in identified regions of interest may help to identify actual causal variants. In addition, imputation in the rest of the dataset, including regions where no signals are initially identified, may allow for identification of additional signals associated with the phenotype under investigation.

Aims and objectives

Minor aim:

(1) Perform quality control on the raw genotype and associated phenotype data for use in each subsequent step.

Major aims:

(1) Develop a queryable cardiometabolic database using the current project-specific data.
    - Arrange the data into structured tables in MySQL.
    - Design a web interface for users to access and query the data.
    - Set up a model/framework for use by other researchers to develop other similar databases.

(2) Use the data stored in the database to identify genetic markers for SBP and DBP in black South African individuals and record the findings in the developed database.

(3) Investigate the effectiveness of genotype imputation in this black South African dataset and possibly provide a more detailed view of identified association signals.

Outline of the thesis

Chapter 1 of this thesis gives an overall background to and literature review of the three main parts of the study, namely the generation of "big data", data management and databases in the biological field; CVD and hypertension prevalence and the non-genetic and genetic risk factors of BP/hypertension; and various aspects of genotype imputation. A description of the participants, the associated data and the quality control of the data used in all aspects of the research is presented in **Chapter 2**. The three major aims of the study are dealt with separately in **Chapters 3 (Database), 4 (BP genetics) and 5 (genotype imputation)**. General conclusions pertaining to all areas of the research are presented in **Chapter 6**.

All program commands/scripts used throughout the thesis are recorded in **Appendix C**, unless otherwise stated. Due to the length of the PHP/HTML/CSS code used for the web interface described in Chapter 3, it is accessible in a separate file (**Web_interface.pdf**) along with the general database model code (**General_database_model.pdf**) at https://github.com/LieslH/Liesl-Hendry-PhD-Code.

Chapter 1: INTRODUCTION

1.1 The “omics” and “big data”

Researchers in biology are constantly striving to characterise biological molecules and translate the information into the structure, function and dynamics of organisms. High-throughput experiments and multi-disciplinary research have contributed to this endeavour and have resulted in a huge growth in the amount and diversity of biological data being generated in recent years (Hirschman et al., 2012). The broad field of biology dedicated to collectively exploring and analysing large amounts of data representing an entire set of some kind is referred to as the “omics”. Genomics specifically involves research into deoxyribonucleic acid (DNA) and its structure and function. It is through the field of genomics and advancements in DNA technologies that have helped researchers to understand the underlying genetic basis of various traits and to better characterise the relationship between a particular phenotype and the genomic pattern for possible prediction of disease risk in high-risk individuals (Merelli et al., 2010).

1.1.1 Advancements in DNA sequencing and chip technologies

To date, great advancements have been made in DNA sequencing technologies resulting in vast amounts of sequencing data being produced. Common techniques build on the Sanger process that was developed in the 1970’s (Adams, 2008) and which lead to the success of the Human Genome Project. Over the years, sequencing has been improved by developing much faster and more automated technologies which has allowed for a subsequent decrease in the cost associated with sequencing (Adams, 2008). Next generation sequencing has also become a popular sequencing technology. This technique parallelises

sequencing, thus increasing the number of sequence reads per run and in turn lowering costs even further (Buermans & den Dunnen, 2014).

Advancements have also been made in the various genotyping arrays (also known as chips) that have already been produced and continue to be produced. These arrays have allowed for the analysis of millions of SNPs per individual in a relatively simple, cost-effective manner. Advancement has been seen in genome-wide association study (GWAS) arrays, as well as arrays for higher resolution SNP genotyping, comprising dense SNP representation at specific loci already identified to be associated with specific trait(s) of interest (Voight et al., 2012). One such chip that has been developed is the Metabochip.

1.2 Data management: benefits and challenges

The volume and complexity of the biological data being produced, together with the pace at which it is being produced, poses many challenges to researchers who need to analyse the data efficiently (Topaloglou, 2004). In order to do this, researchers need to find efficient ways to store and manage the data and make it more accessible. This will allow researchers to process large amounts of data quickly and design experiments with more insight (Howe et al., 2008).

Several challenges have arisen, however, when dealing with the management of biological data. As biological data are broad, diverse and forever evolving, existing data management technology has often not been adequate or suitable (Topaloglou, 2004). Another barrier has been the cost associated with implementing some data management systems or paying data managers (Anderson et al., 2007) and, very often, little funding is set aside for this (Gross, 2011). The limited extra time available for researchers to implement their own data management systems and change their current practices into more efficient ones has also been a challenge (Anderson et al., 2007).

Many researchers generally use some form of electronic organisation, but instead of using specialised applications suited for their data, they have in the past, and still to this day, used general-purpose applications like spreadsheets. Spreadsheets have become a common tool for storing data, but are limited by processing power, the complexity of queries that can be run and in the size of data that can be stored. As data becomes more complex and increases in size, spreadsheets become a less feasible option for storage. Spreadsheets, however, remain a popular choice due to their ease of use, simplicity, familiarity and little or no cost associated with using them. (Anderson et al., 2007)

1.3 Databases

The challenges associated with data management and some past and current practices has lead to the need to develop formal databases to efficiently store and manage data (Larranaga et al., 2006). The organisation of relevant data into databases aims to make the data more accessible and in such a format that allows for easy extraction of useful information from the data (Larranaga et al., 2006). Central to database development is biocuration. Biocuration can be defined as "the activity of organising, representing and making biological information accessible to both humans and computers" (Howe et al., 2008). Although biocuration is considered to be a vital "tool" in making data accessible, it tends to lag behind data generation in terms of available funding and the development and recognition of the tool as an important step in biological research (Howe et al., 2008).

1.3.1 Online databases

Online databases, which are those accessible via the internet as opposed to being stored locally, have become popular tools for publishing biological data (Howe et al., 2008), with the number of databases increasing every year (Hirschman et al., 2012). The Nucleic Acids Research online Molecular Biology

Database Collection currently lists 1685 available databases (Rigden et al., 2016). These databases make up a selection of resources provided by three major bioinformatics centres: the United States National Centre for Biotechnology Information (NCBI), the European Bioinformatics Institute (EMBL-EBI) and the Swiss Institute for Bioinformatics (SIB) (Rigden et al., 2016).

Many of the new and updated databases for 2016 are dedicated to the genetics of disease and drug research (Rigden et al., 2016). The existing databases fall into several categories with some being specific to certain populations, diseases/traits or genetic loci. Some are also more general. Examples of some of these general databases related to human genes and diseases and which are relevant to our area of research are listed in **Table 1.1**.

Table 1.1 A selection of useful internet accessed databases related to human genes and diseases

Database	Brief description	URL
ClinVar	Groups information about genomic variation and its relationship to human health.	www.ncbi.nlm.nih.gov/clinvar/
dbGaP	Database of genotypes and phenotypes. Archives and distributes the data and results from studies investigating the interaction of genotypes and phenotypes in humans.	www.ncbi.nlm.nih.gov/gap
dbSNP and dbSNP-Q	Database of SNPs and multiple small-scale variations including insertions/deletions, microsatellites, and non-polymorphic variants. dbSNP-Q is a web application for querying dbSNP.	www.ncbi.nlm.nih.gov/projects/SNP/ cgsmd.isi.edu/dbsnpq/
DECIPHER (DatabasE of genomiC varIation and Phenotype in Humans using Ensembl Resources)	Incorporates a suite of tools designed to aid the interpretation of genomic variants.	decipher.sanger.ac.uk/
F-SNP	Provides integrated information about the functional effects of SNPs obtained from 16 bioinformatics tools and databases.	compbio.cs.queensu.ca/F-SNP/
GenAtlas	Provides information on the structure, expression and function of genes, gene mutations and their consequences on diseases.	genatlas.medecine.univ-paris5.fr/
GWAS Catalog	A curated collection of all published genome-wide association studies.	www.ebi.ac.uk/gwas/

Table 1.1 (continued)

Database	Brief description	URL
GWAS Central	Provides a centralized compilation of summary level findings from genetic association studies, both large and small.	www.gwascentral.org/
GWASdb	Human genetic variants identified by genome wide association studies.	jjwanglab.org/gwasdb/
PhenomicDB	Integrates public genotype/phenotype data from a wide range of model organisms and humans.	www.phenomicdb.de/
SNPedia	Wiki on SNPs and genome annotation.	snpedia.com/
SNPlogic	Provides a comprehensive interactive SNP annotation, selection and prioritization system for focused genotyping projects and/or analysis and interpretation of SNP data.	www.snplogic.org/
VarySysDB	Contains various types of human gene polymorphisms.	h-invitational.jp/varygene/home.htm
VaDE (VarySysDB Disease Edition)	Provides genomic polymorphisms associated to diseases, traits, and pharmacogenomics.	bmi-tokai.jp/VaDE/

The multitude of online databases contain publicly available data for use by multiple researchers to obtain important information relevant to their studies. Biological research can be further advanced by moving project-specific data from basic spreadsheets into databases and making the data available to and queryable by all members of a research group via web interfaces similar to those that exist for the public databases.

1.4 Biological focus: Cardiovascular diseases and hypertension

CVDs are the leading cause of NCD deaths globally ahead of cancers, respiratory diseases and diabetes (World Health Organization, 2014b). Research on communicable or infectious diseases has gained much of the focus until now, particularly in Africa where they are a main cause of morbidity and mortality. Chronic NCDs are, however, gaining increasing interest due to their burden becoming as significant. In the latest global status report on NCDs, it was reported that in 2012 approximately 38 million (63%) deaths globally were due to NCDs (World Health Organization, 2014a). It is projected that NCD deaths will increase to 52 million by 2030 (Mathers & Loncar, 2006). Low- and middle-income countries (LMIC) are the most affected, with about 28 million of the NCD deaths occurring in these countries and the NCD death rate being 625 and 673 per 100 000 in low-income and lower-middle-income countries, respectively, compared to 397 per 100 000 in high-income countries (World Health Organization, 2014b). In addition, 82% of premature deaths (deaths before the age of 70) occur in LMIC.

CVDs were responsible for approximately 17.5 million (46%) NCD deaths in 2012, with more than 80% occurring in LMIC (Abegunde et al., 2007). This figure is expected to increase to about 22.2 million in 2030, with approximately 85% occurring in LMIC (World Health Organization, 2014b; Mathers & Loncar, 2006). In LMIC, serious and fatal CVD-related events usually occur in females, and commonly those that are pregnant, and younger individuals (Sliwa et al., 2014).

A major risk factor contributing to CVDs, and the leading cause of morbidity and mortality worldwide (Guwatudde et al., 2015), is hypertension or raised BP. In 2014 the global prevalence of raised BP was approximately 22% in adults aged 18 years and older, with the highest prevalence reported in Africa at 30% for all adults combined (World Health Organization, 2014b). In 2010, raised BP was estimated to have caused 9.4 million deaths globally (World Health Organization,

2014b). Hypertension itself is considered a major risk factor for strokes, MI, cardiac failure, dementia, renal failure and blindness. Specifically, hypertension is responsible for about 45% and 51% of deaths due to heart disease and stroke, respectively (Singh et al., 2016).

Two main forms of hypertension exist: primary (or essential) hypertension, which is the most prevalent form and has no single known cause, and secondary hypertension, which results from some or other underlying condition and include the monogenic forms of hypertension. The focus of this research and the main subject for the remainder of this chapter is primary hypertension. SBP and DBP (both continuous traits) are common measurements of hypertension (a dichotomised disease category), which is characterised by sustained high BP, with average SBP and DBP readings in adults of 140mmHg and 90mmHg or greater, respectively (Rosamond et al., 2007). This is in contrast to the normal or accepted values of less than 120mmHg and 80mmHg for SBP and DBP, respectively.

1.4.1 Ethnic differences in hypertension prevalence

There is a difference in the prevalence and clinical presentation of hypertension among individuals of different ethnicity, with African or black individuals in many cases fairing much worse (Cooper & Rotimi, 1994, 1997; Takeuchi et al., 2010; Egan et al., 2010). Hypertension in black Africans usually presents more frequently, is more severe (Salako et al., 2007) and is more resistant to treatment (Addo et al., 2007). The 2005-2006 National Health and Nutrition Examination Survey reported that 27% of African American adults suffer from hypertension compared to only 17% of Caucasians (Redmond et al., 2011). A similar trend was seen in a study in an adult urban population in Durban, South Africa with the prevalence in black, white and Indian individuals being 25%, 17.2% and 14.2%, respectively (Seedat, 1999).

The increased prevalence of and predisposition to hypertension in black individuals may be due to differences in the handling of potassium by the kidney, with black individuals usually having a lower urinary excretion of potassium (Aviv et al., 2004). In addition, they have increased sodium reabsorption by the kidneys leading to a more expanded plasma volume (Tu & Pratt, 2013). As a result, black individuals usually develop salt-sensitive hypertension (Weinberger et al., 1986).

The total number of individuals with hypertension in developing countries, and particularly in Africa, is high. A major contributing factor to this is the inability of these individuals to afford treatment, compared to those in developed countries (Nissinen et al., 1988), or the lack of treatment facilities (Opie & Seedat, 2005). Poor education and a low understanding of the severity of hypertension also play a role in inadequate control of BP (Opie & Seedat, 2005). There has also been a reported difference in prevalence in rural compared to urban areas, with individuals in urban areas being at a higher risk than those in rural areas (Opie & Seedat, 2005), mainly due to dietary differences. This difference becomes less apparent, however, following urban exposure in rural areas.

Hypertension was once considered rare in sub-Saharan Africa (SSA), but has become more prevalent in recent years, mainly due to migration of people from rural to urban areas and subsequent changes in lifestyle (Ogah & Rayner, 2013). In 2008, the prevalence of hypertension in SSA was 16.2%, with approximately 74.7 million affected individuals. This figure is expected to increase to about 125.5 million by 2025 (Ogah & Rayner, 2013). A recent study also reported an increased severity in South African blacks compared to African Americans where SBP was 9.7mmHg higher in South African blacks (Cooper et al., 2015). As with the rest of Africa, the awareness, treatment and control of hypertension are generally low in SSA (Ogah & Rayner, 2013).

1.4.2 Non-genetic risk factors

Several studies have reported obesity, or an increased BMI, and salt intake as two of the main risk factors associated with raised BP and hypertension (Forrester, 2004; Yang et al., 2012). A study conducted in individuals of African ancestry from West Africa, the Caribbean and North America showed that these two risk factors accounted for about 70% of the prevalence differences across these populations. In addition, they are risk factors common to populations worldwide (Forrester, 2004). Approximately 1.7 million CVD-related deaths globally are due to excessive sodium intake (World Health Organization, 2014b).

Age and sex are also known risk factors for BP and hypertension. For example, men have a higher SBP and DBP than women at a younger age (Roger et al., 2012), but, following menopause, women have higher BP than men (Reckelhoff, 2001). An early study in South African men and women revealed that women aged between 35 to 40 are at a greater risk for hypertension than men (Seedat, 1983). Additionally, an assessment of hypertension in four countries in SSA (Burkina Faso, Ghana, Kenya and South Africa) showed that the prevalence of hypertension was significantly higher in women than in men and increased with age in males and females separately and combined (Gómez-Olivé et al., under review).

Other risk factors include insulin resistance (Banerjee, 2013), physical inactivity (Council on High Blood Pressure Research, 2003), alcohol (Council on High Blood Pressure Research, 2003) and coffee (Chalmers et al., 1999) intake, psychosocial stress (Council on High Blood Pressure Research, 2003), consumption of high-fat foods (Douglas et al., 2003; Boutin-Foster et al., 2007), smoking (Chalmers et al., 1999), dyslipidemia (Tchelougou et al., 2015) and HIV infection (Bärnighausen et al., 2007).

1.4.3 Genetic risk factors

Primary hypertension is a multifactorial disorder with genetics playing a role in its aetiology in addition to the various environmental or non-genetic risk factors. Familial studies have shown that BP/hypertension is roughly 30-50% heritable (Munroe et al., 2013). It has also been suggested that differences in prevalence of the disorder among various ethnic groups is due to the underlying genetics of the individuals (Grim & Robinson, 1996). Although much has been elucidated about the rare monogenic forms of hypertension, the underlying genetic basis of primary hypertension remains poorly understood, with small contributions from multiple genes believed to affect the aetiology of the disorder (Doris, 2002). This makes identifying loci/variants associated with BP/hypertension quite challenging.

Attempts to identify the genetic determinants of BP/hypertension have been made by conducting genome-wide linkage analyses, candidate gene studies and, more recently, GWAS. Probable candidate genes for investigation were first chosen based on their known biochemical or physiological function and link to BP regulation.

As the kidney plays a role in the long-term regulation of BP, genes influencing renal salt handling were of particular interest (Singh et al., 2016). Mutations in genes involved in BP control by the kidneys have already been identified for the monogenic forms of hypertension (Lifton, 2004). It has been hypothesised that causes of hypertension include low plasma renin levels, sodium sensitivity and cellular abnormalities, epithelial sodium channel (ENaC) changes, altered genes regulating the renin-angiotensin-aldosterone system (RAAS) or increased peripheral vascular resistance (Opie & Seedat, 2005). Some of the first genes studied for primary hypertension were those involved in the RAAS including, among others, the angiotensin-converting enzyme (*ACE*) and angiotensinogen (*AGT*) genes (Norton et al., 2010). As reviewed by Norton and colleagues (2010),

despite the role of the protein products of *ACE* and *AGT* in BP regulation, studies have generally been inconclusive about the impact of variants in these genes on BP, with studies both supporting and refuting an association. In black South Africans, an association was found between a polymorphism in the promoter region of *AGT* and a greater than expected increase in SBP (Tiago et al., 2002). In a study in Burkina Faso, West Africa, an association was found between the DD genotype of the *ACE* gene and hypertension susceptibility (Tchelougou et al., 2015). Another association was found between the aldosterone synthase gene (*CYP11B2*) and higher initial SBP in previously untreated black South Africans (Tiago et al., 2003).

Other candidate genes of interest are those encoding the subunits of the ENaC. It is a plausible candidate for salt sensitivity as it is the final regulator of sodium balance in the kidney (Jones et al., 2012). The first report of SNPs in ENaC genes associated with hypertension in blacks was in a set of individuals from London where an association was found with the T594M SNP of the β-chain (Baker et al., 1998). This finding could not, however, be replicated in African Americans or South African blacks (Nkeh et al., 2003; Hollier et al., 2006). A novel mutation was, however, found between R563Q and low-renin, low-aldosterone hypertension in black and mixed ancestry South Africans (Rayner et al., 2003). The G protein-coupled receptor kinase 4 (*GRK4*) gene has also been suggested as a possible candidate for hypertension. Variants in *GRK4*, especially Ala142Val, could account for the low-renin, low-aldosterone phenotype seen in black individuals (Rayner & Spence, 2017).

Early studies were relatively unsuccessful and produced inconsistent findings (Zhao et al., 2013) and replication in linkage analyses and candidate gene studies has proven to be challenging (Zheng et al., 2015b). Introduction of SNP genotyping arrays for large numbers of variants and the formation of large BP consortia has advanced the discovery of BP/hypertension variants to a certain

extent (Zhao et al., 2013; Padmanabhan et al., 2015). In many cases, however, although multiple BP/hypertension loci have been found, few reach genome-wide significance ($p<5x10^{-8}$), raising concern about the effectiveness of GWAS in studying BP/hypertension (Levy et al., 2009; Takeuchi et al., 2010). In addition, the identified variants/loci are very often not the causal variants/loci, but rather tag the causal variants, making fine-mapping studies a necessity (Wang et al., 2011). Nevertheless, GWAS, and meta-analysis across many studies, have shed some light on possible genetic variants or loci linked to BP for further study.

To date, however, most of the large scale studies that have been published have been carried out in individuals of European Ancestry, with fewer conducted in individuals of Asian or African ancestry. Significant findings from some of the main studies conducted in non-African individuals are summarised in **Table 1.2**. In addition, several associations not at genome-wide significance were reported in smaller studies in Europeans, with some of these findings then confirmed in the larger studies or following replication. The identified loci included *STK39* (SBP and DBP) (Wang et al., 2009), upstream of *CDH13* (DBP and hypertension) (Org et al., 2009) and *NPPA/NPPB* (SBP, DBP and hypertension) (Newton-Cheh et al., 2009b).

Table 1.2 Findings from some of the main BP/hypertension genetic studies in non-Africans

Population [Cohort/Consortium]	Type of study	Replication/follow-up performed?	Genes/Variants (Association) *(At genome-wide, array-wide, replication or study significance level)*	Year	Reference
Europeans [Global BP Genetics (GBPGEN) consortium]	GWAS meta-analysis	Yes	Variants in/near: *CYP17A1* (SBP) *CYP1A2* (DBP) *FGF5* (DBP) *SH2B3* (DBP) *MTHFR* (SBP) *C10orf107* (DBP) *ZNF652* (DBP) *PLCD3* (SBP)	2009	(Newton-Cheh et al., 2009a)
Europeans [Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) consortium]	GWAS meta-analysis	Yes	Variants in/near: *ATP2B1* (SBP, DBP, hypertension) *C18orf1* (SBP) *CASZ1* (SBP) *SH2B3* (DBP) *ATXN2* (DBP) *TRAFD1* (DBP) *TBX3/TBX5* (DBP) *PLEKHA7* (DBP) In silico comparison with GBPGEN: *ATP2B1* (SBP, DBP and hypertension) *SH2B3* (DBP) *TBX3/TBX5* (DBP) Joint meta-analysis of CHARGE and GBPGEN: *CYP17A1* (SBP) *PLEKHA7* (SBP) *ATP2B1* (SBP, DBP, hypertension) *SH2B3* (SBP, DBP)	2009	(Levy et al., 2009)

Table 1.2 (continued)

Population [Cohort/Consortium]	Type of study	Replication/follow-up performed?	Genes/Variants (Association) *(At genome-wide, array-wide, replication or study significance level)*	Year	Reference
			ULK4 (DBP) *CACNB2* (DBP) *TBX3/TBX5* (DBP) Locus adjacent to *CSK/ULK3* (DBP)		
Japanese	Multistage replication study (of European signals)	No	Variants in/near: *CASZ1* (SBP, DBP, hypertension) *MTHFR*(SBP) *ITGA9* (hypertension) *FGF5* (SBP, DBP, hypertension) *CYP17A1/CNNM2* (SBP, DBP, hypertension) *ATP2B1* (SBP, DBP, hypertension) *CSK/ULK3* (DBP, hypertension)	2010	(Takeuchi et al., 2010)
Europeans	GWAS	Yes	Variant upstream of *UMOD* (hypertension)	2010	(Padmanabhan et al., 2010)
East Asians [Asian Genetic Epidemiology Network Blood Pressure (AGEN-BP) consortium]	GWAS meta-analysis (3-stage including replication)	Yes	Novel loci: *ST7L-CAPZA1 (DBP)* *FIGN-GRB14 (SBP)* *ENPEP (DBP)* *NPR3 (SBP)* Novel association near known locus (*TBX3)* East Asian specific association at *ALDH2* Previous European loci replicated in East Asians: *CASZ1* (DBP) *FGF5* (SBP, DBP) *CYP17A1* (SBP, DBP) *ATP2B1* (SBP, DBP)	2011	(Kato et al., 2011)

Table 1.2 (continued)

Population [Cohort/Consortium]	Type of study	Replication/follow-up performed?	Genes/Variants (Association) *(At genome-wide, array-wide, replication or study significance level)*	Year	Reference
Europeans [International Consortium for BP-GWAS (ICBP-GWAS)]	GWAS meta-analysis (including discovery and follow-up data)	Yes	Novel loci: *MOV10* (SBP, DBP) *SCL4A7* (DBP) *MECOM* (SBP, DBP) *SLC39A8* (SBP, DBP) *GUCY1A3-GUCY1B3* (DBP) *NPR3-C5orf23* (SBP, DBP, hypertension) *EBF1* (SBP, DBP) *HFE* (SBP, DBP, hypertension) *BAT2-BAT5* (SBP, DBP, hypertension) *CACNB2(5')* (SBP, DBP) *PLCE1* (SBP, hypertension) *ADM* (SBP) *FLJ32810-TMEM133* (SBP, DBP, hypertension) *FURIN-FES* (SBP, DBP) *GOSR2* (SBP) *JAG1* (SBP, DBP) *GNAS-EDN3* (SBP, DBP, hypertension) Confirmed loci: *MTHFR-NPPB* (SBP, DBP, hypertension) *ULK4* (DBP) *FGF5* (SBP, DBP) *CACNB2(3')* (SBP, DBP, hypertension) *C10orf107* (SBP, DBP, hypertension) *CYP17A1-NT5C2* (SBP, DBP) *PLEKHA7* (SBP, DBP) *ATP2B1* (SBP, DBP, hypertension) *SH2B3* (SBP, DBP)	2011	(Ehret et al., 2011)

Table 1.2 (continued)

Population [Cohort/Consortium]	Type of study	Replication/follow-up performed?	Genes/Variants (Association) *(At genome-wide, array-wide, replication or study significance level)*	Year	Reference
			TBX5-TBX3 (DBP) *CYP1A1-ULK3* (SBP, DBP, hypertension) *ZNF652* (SBP, DBP) 9 loci replicated in East Asians and 6 in South Asians		
Europeans (multiple cohorts)	Meta-analysis using a gene-centric array (HumanCVD BeadChip from Illumina)	Yes	Novel loci: *LSP1/TNNT3* (mean arterial pressure (MAP)) *MTHFR-NPPB* (DBP) Confirmed loci: *ATP2B1* (hypertension) *AGT* (hypertension) Combined discovery and follow-up data: *MTHFR-NPPB* (DBP) *AGT* (hypertension) *NPR3* (SBP) *HFE* (DBP) *NOS3* (DBP) *LSP1/TNNT3* (MAP) *SOX6* (MAP) *ATP2B1* (hypertension)	2011	(Johnson et al., 2011)
Europeans [HYPERGENES project part of the European Network for Genetic-Epidemiological Studies]	GWAS meta-analysis (of discovery and validation phase)	Yes	Novel locus: *NOS3* (hypertension)	2012	(Salvi et al., 2012)

Table 1.2 (continued)

Population [Cohort/Consortium]	Type of study	Replication/follow-up performed?	Genes/Variants (Association) *(At genome-wide, array-wide, replication or study significance level)*	Year	Reference
Europeans (multiple cohorts)	Meta-analysis using a gene-centric array (HumanCVD BeadChip from Illumina)	Yes	Novel loci: *HRH1* (SBP) *SOX6* (SBP) *[previously MAP]* *MDM4* (DBP) Confirmed loci: *ADRB1* (MAP) *ATP2B1* (SBP, MAP) *SH2B3/ATXN2* (SBP, DBP, MAP) *CSK* (SBP, DBP, MAP) *CYP17A1* (SBP, pulse pressure (PP)) *FURIN* (SBP, DBP, MAP) *HFE* (DBP) *LSP1* (SBP) *MTHFR* (SBP, DBP, MAP)	2013	(Ganesh et al., 2013)
Europeans (multiple cohorts)	Meta-analysis using a gene-centric array (HumanCVD BeadChip from Illumina)	Yes	Novel loci: *PDE1A* (DBP, MAP) *HLA-DQB1* (DBP) *VCL* (DBP, MAP) *PRKAG2* (SBP) *H19* (SBP, MAP) *NUCB2* (SBP, MAP, PP) *SIPA1* (SBP) *HOXC* complex (SBP) *RELA* (MAP) *CDK6* (PP) *FBN1* (PP) *NFAT5* (PP) Confirmed 27 previous loci	2014	(Tragante et al., 2014)

Table 1.2 (continued)

Population [Cohort/Consortium]	Type of study	Replication/follow-up performed?	Genes/Variants (Association) *(At genome-wide, array-wide, replication or study significance level)*	Year	Reference
Chinese	GWAS meta-analysis	Yes	Novel loci: *CACNA1D* (DBP) *CYP21A2* (SBP, DBP, hypertension) *MED13L* (SBP, DBP) Confirmed loci: *CASZ1* (SBP, hypertension) *MOV10* (SBP, hypertension) *FGF5* (SBP, DBP, hypertension) *CYP17A1* (SBP, DBP, hypertension) *SOX6* (SBP, DBP) *ATP2B1* (SBP, DBP, hypertension) *ALDH2* (DBP) *JAG1* (SBP, DBP) *SLC4A7* (new variant; SBP)	2015	(Lu et al., 2015)
Trans-ethnic (East Asian, South Asian, European)	GWAS meta-analysis	Yes	Novel loci: *IGFBP3* (PP) *KCNK3* (MAP) *PDE3A* (DBP) *PRDM6* (SBP) *ARHGAP24* (SBP) *OSR1* (PP) *SLC22A7/TTBK1/ZNF318* (SBP) *TBX2/c17orf82* (MAP) *ABLIM3/SH3TC2* (DBP) *HDAC9* (PP) *LRRC10B/SYT7* (MAP) *AMH/DOT1L/PLEKHJ1/SF3A2* (PP)	2015	(Kato et al., 2015)

Table 1.2 (continued)

Population [Cohort/Consortium]	Type of study	Replication/follow-up performed?	Genes/Variants (Association) *(At genome-wide, array-wide, replication or study significance level)*	Year	Reference
Various populations (East Asian, African, Caucasian)	Meta-analysis of case-control studies	No	Variants in: *GRK4* (hypertension in Caucasians) *DRD1* (hypertension in East Asians)	2015	(Zhang et al., 2015)
Europeans [Early Genetics and Lifecourse Epidemiology (EAGLE) consortium]	GWAS meta-analysis	Yes	Novel loci: *ITGA11* (pre-puberty SBP) rs872256 in unknown locus (puberty SBP)	2016	(Parmar et al., 2016)
Europeans [BELHYPGEN cohort]	Snapshot genotyping of a few SNPs	No	Confirmed loci: *STK39* (SBP, hypertension) *WNK1* (SPB, hypertension)	2016	(Persu et al., 2016)
Europeans (multiple studies)	Meta-analysis using the Metabochip	Yes	Novel loci: *HIVEP3* (SBP) *PNPT1* (DBP) *FGD5* (SBP, DBP) *ADAMTS9* (DBP) *TBC1D1–FLJ13197* (SBP, DBP) *TRIM36* (SBP, DBP) *CSNK1G3* (DBP) *CHST12–LFNG* (SBP, DBP) *ZC3HC1* (SBP, DBP) *PSMD5* (SBP) *DBH* (SBP, DBP) *RAPSN/PSMC3/SLC39A13* (SBP, DBP) *LRRC10B* (SBP, DBP) *SETBP1* (SBP, DBP) *INSR* (SBP, DBP) *ELAVL3* (DBP) *CRYAA–SIK1* (SBP, DBP)	2016	(Ehret et al., 2016)

Table 1.2 (continued)

Population [Cohort/Consortium]	Type of study	Replication/follow-up performed?	Genes/Variants (Association) *(At genome-wide, array-wide, replication or study significance level)*	Year	Reference
			Confirmed loci: *CASZ1* (SBP, DBP) *MTHFR–NPPB* (SBP, DBP) *ST7L–CAPZA1–MOV10* (SBP, DBP) *MDM4* (DBP) *AGT* (SBP, DBP) *KCNK3* (SBP, DBP) *NCAPH* (DBP) *FIGN–GRB14* (SBP, DBP) *HRH1–ATG7* (SBP) *SLC4A7* (SBP) *ULK4* (DBP) *MAP4* (SBP, DBP) *MECOM* (SBP, DBP) *FGF5* (SBP, DBP) *ARHGAP24* (SBP) *SLC39A8* (SBP, DBP) *GUCY1A3–GUCY1B3* (SBP) *NPR3–C5orf23* (SBP, DBP) *EBF1* (SBP, DBP) *HFE* (SBP, DBP) *BAT2–BAT5* (DBP) *ZNF318–ABCC10* (SBP) *RSPO3* (SBP, DBP) *PLEKHG1* (DBP)		

Table 1.2 (continued)

Population [Cohort/Consortium]	Type of study	Replication/follow-up performed?	Genes/Variants (Association) *(At genome-wide, array-wide, replication or study significance level)*	Year	Reference
			HOTTIP–EVX (SBP, DBP) *PIK3CG* (SBP) *BLK–GATA4* (SBP) *CACNB2* (SBP, DBP) *C10orf107* (SBP, DBP) *SYNPO2L* (SBP) *PLCE1* (SBP, DBP) *CYP17A1–NT5C2* (SBP, DBP) *ADRB1* (SBP, DBP) *LSP1–TNNT3* (SBP, DBP) *ADM* (SBP, DBP) *PLEKHA7* (SBP, DBP) *SIPA1* (SBP) *FLJ32810–TMEM133* (SBP, DBP) *PDE3A* (DBP) *ATP2B1* (SBP, DBP) *SH2B3* (SBP, DBP) *TBX5–TBX3* (DBP) *CYP1A1–ULK3* (SBP, DBP) *FURIN–FES* (SBP, DBP) *PLCD3* (SBP) *GOSR2* (SBP) *ZNF652* (DBP) *JAG1* (SBP, DBP) *GNAS–EDN3* (SBP, DBP)		

Table 1.2 (continued)

Population [Cohort/Consortium]	Type of study	Replication/follow-up performed?	Genes/Variants (Association) *(At genome-wide, array-wide, replication or study significance level)*	Year	Reference
Hispanic and trans-ethnic (Hispanics and African Americans) [Population Architecture using Genomics and Epidemiology (PAGE) consortium]	Meta-analysis using the Metabochip	Yes	Hispanics – known loci: *KCNK3* (DBP) *FGF5* (SBP, DBP) *ATXN2-SH2B3* (DBP) Trans-ethnic – novel variants in known loci: *ULK4* (DBP) *FGF5* (DBP) *HOXA-EVX1* (SBP)	2016	(Franceschini et al., 2016)

BP genetics in African individuals

The first BP GWAS to be carried out in individuals of African ancestry focused on an African-American population (Adeyemo et al., 2009). SNPs in or near six loci (*PMS1*, *SLC24A4*, *YWHA7*, *IPO7*, *CACNA1H* and *AL365365.23* (a pseudogene)) associated with SBP at a genome-wide significance level, but no SNPs met genome-wide significance for association with DBP or hypertension. Of the 17 SNPs taken through for replication in a West African population, three were associated with either SBP, DBP or hypertension. A combined meta-analysis of the African American and West African individuals revealed a significant association with five of the SNPs, including rs11160059 in *SLC24A4* (Adeyemo et al., 2009). A later evaluation of 16 top associated SNPs from this study in a different sample of African Americans showed no replication (Kidambi et al., 2012).

In a second GWAS in African-Americans, and the largest to date, two novel loci (rs2258119 in *C21orf91* - SBP and rs10474346 intergenic to *GPR98* and *ARRDC3* - DBP) reached genome-wide significance, but could not be replicated in independent African American cohorts (Fox et al., 2011). In addition, three loci previously reported in Americans of European ancestry (*SH2B3*, *TBX3-TBX5* and *CSK-ULK3*) were replicated in the African-Americans in this study.

A GWAS meta-analysis of 19 African American and one Yoruban sample from the Continental Origins and Genetic Epidemiology Network (COGENT) revealed an association between a SNP in *CYB5R2* (rs11041530) and SBP. A trans-ethnic meta-analysis of the discovery samples and additional African, European and East Asian samples failed to replicate this association, but identified three novel loci (*EVX1-HOXA*, *RSPO3*, *PLEKHG1)* and one novel association in a known locus *(SOX6)* and managed to fine map four previously identified loci (*EBF1, ATP2B1, NT5C2, ULK4*) (Franceschini et al., 2013). A further study in the same set of individuals integrated association evidence from summary statistics of multiple

traits and found genome-wide significant associations with *CHIC2*, *HOXA-EVX1*, *IGFBP1/IGFBP3* and *CDH17* (Zhu et al., 2015).

Several smaller scale studies have also identified BP or hypertension associated loci in African-Americans. For example, a SNP in *VNN1* (rs2272996) is associated with increased risk for hypertension (Zhu & Cooper, 2007); rs7726475, a SNP located between *SUB1* and *NPR3*, is associated with SBP and DBP (Zhu et al., 2011) and rs437470 in *CXADR* is associated with SBP and DBP (Shetty et al., 2012). In another study in African-Americans, suggestive associations were found between polymorphisms in *RGS5* and hypertension, SBP and DBP; polymorphisms in the *SELE* gene and SBP and BDP; and polymorphisms in *ATP1B1* and SBP (Faruque et al., 2011). These three genes had all previously been shown to be associated with hypertension in Americans of European ancestry and also replicated some of the findings from previous investigations in African-Americans (Chang et al., 2007).

Genetic association studies conducted in individuals of African ancestry other than African-Americans have been limited. An association was found between a SNP in *SLC4A5* (rs8179526) and SBP in West African women (Taylor et al., 2011). Some very small scale studies have been conducted in South African individuals. In addition to the aforementioned associations reported in ENaC subunit genes, *AGT* and *CYP11B2*, studies conducted in individuals from Johannesburg have shown associations between the 460-Trp variant in the α-adducin gene and hypertension (Barlassina et al., 2000) and an intron 2 polymorphism in the atrial natriuretic peptide (*ANP*) gene and the absence of hypertension (Nkeh et al., 2002). On the other hand, associations could not be found between a β_2 adrenoreceptor polymorphism (Candy et al., 2000) and hypertension.

1.5 Genotype imputation

GWAS have become a powerful alternative to family-based linkage studies (Marchini et al., 2007) and are a useful tool in genetics to discover variants associated with specific traits of interest. The chips or arrays used in GWAS, however, only contain a fraction of the SNPs that exist (Pei et al., 2008) and the findings from GWAS only explain a small proportion of the heritability of many phenotypes leading to what has become known as "missing heritability" (Manolio et al., 2009). In the case of BP/hypertension, the identified loci and variants to date explain less than 2.5% of the phenotypic variance for SBP and DBP (Ehret et al., 2011). A potential source of much of this "missing heritability" are rare variants (minor allele frequency (MAF) < 1%) (Dickson et al., 2010; Frazer et al., 2009), which are not generally included on the genotyping arrays or are removed during QC. Missing genotype information also results from removal of SNPs due to low call rates and deviations from Hardy-Weinberg equilibrium (HWE) (Pei et al., 2008) or variations in the depth and scope of assessment across different genotyping platforms (Zhang et al., 2011). Having more complete information can significantly enhance the power to detect causal variant(s) for a trait of interest (Zhang et al., 2011).

Genotype imputation, which involves the inference of genotype information at SNPs not initially genotyped, has become a useful addition to many GWAS to recover some of the missing or ungenotyped data and has become possible due to an increased understanding of the genome structure, substructure and population admixture (Zhang et al., 2011). The aim of imputation is to predict or impute the missing or ungenotyped data into the dataset based on the observed or genotyped data and information about the LD within the region from a chosen reference panel (Marchini et al., 2007; Biernacka et al., 2009). Imputation also relies on the fact that individuals with a common ancestor share extended haplotypes over short regions (Scheet & Stephens, 2006; Browning & Browning, 2007).

1.5.1 Benefits of imputation

The result of imputation is a larger and denser set of SNPs to test for association with the trait/phenotype under investigation and a much more detailed view of the associated region (Marchini et al., 2007), thus increasing the power of GWAS. Imputation can allow researchers to use genotyping arrays with slightly lower coverage and genotype more individuals rather than additional SNPs (Anderson et al., 2008). It has also increased the potential to identify increasingly subtle signals from large and complex datasets (Howie et al., 2009).

Imputation can help in the fine-mapping of a region of association through testing of a larger number of SNPs, with some imputed SNPs being more strongly associated with the trait of interest than genotyped SNPs (Li et al., 2009). This helps researchers to potentially pinpoint the associated regions more precisely (Nho et al., 2011) and has huge implications for follow-up studies where these imputed SNP signals can be studied in more detail (Ellinghaus et al., 2009). In addition, SNPs identified in pathway analyses, but that are not genotyped, can be imputed into the dataset for further analysis (Nho et al., 2011).

Many disease variants have small effects and are often undetected in individual studies. Meta-analyses, conducted by combining data across several studies, can facilitate detection of these variants (Marchini et al., 2007; Li et al., 2012) and identification of novel variants/loci (de Bakker et al., 2008). Different studies, however, often use different genotyping platforms or analyse a different set of SNPs. Imputation is useful in this case as it allows for a successful combination of different datasets and imputation of missing SNPs across studies to produce one common set of SNPs for analysis. Despite the possible benefit, however, Li et al (2012) suggested that the power of analysis of individual studies may still be higher than that of a meta-analysis with imputation, due to the genotype uncertainty that may be introduced with imputation being higher than the gained power from increasing the sample size. They suggested that the results

from the largest study be studied first as it may provide more power (Li et al., 2012). In addition, it has been found that imputation-induced bias can be introduced when combining all genotyped SNPs across arrays, even if very similar (Johnson et al., 2013). A more favourable option is to only combine those SNPs available on all arrays and then impute up to a common set of reference panel SNPs (Johnson et al., 2013).

In addition to increasing the power of GWAS and its usefulness in fine-mapping and meta-analysis, imputation can also help in QC steps by highlighting possible genotype errors and may help in the reconstruction of missing genotype data in ungenotyped family members in the context of pedigree data (Ellinghaus et al., 2009).

Central to imputation is the use of an appropriate software package with underlying imputation algorithm and a reference panel, which provides the necessary information to infer ungenotyped SNPs. Several imputation tools have been developed in recent years as the need for efficient imputation has increased (Zhang et al., 2011) and IMPUTE (initially version 1 and later version 2) is one such commonly used tool (Marchini et al., 2007; Howie et al., 2009).

1.5.2 IMPUTE and IMPUTE2

To infer missing genotypes, IMPUTE uses a set of known haplotypes from publically available data (reference panel) and the information from all markers in LD with the SNPs to be imputed (Zhao et al., 2008). The algorithm underlying the tool is based on a hidden Markov model (HMM). The set of haplotypes from the reference panel make up the pool of "hidden states" of the Markov chain and haplotypes and missing genotypes are inferred in the study samples according to these "hidden states". It compares the potential haplotype for each individual with all other observed haplotypes, rather than using a representative set of haplotypes (Li et al., 2009). The computation time of the program

increases linearly with the number of markers and quadratically (or less) with the number of "states" at each marker (Browning, 2008) making the process computationally quite intensive at times. IMPUTE imputes the missing genotypes without any reference to the phenotype data (Marchini & Howie, 2008) and generates an 'info' metric for each SNP which gives an indication of the certainty with which the SNP is imputed.

An update to the first version of IMPUTE (IMPUTE2) was developed to address the challenges faced by newer, more complex datasets and to improve imputation accuracy of ungenotyped SNPs by improving the accuracy of haplotype estimation (phasing) at genotyped SNPs (Howie et al., 2009). The newer generation of datasets may be larger, contain unphased and incomplete genotypes and offer additional reference data, even in the form of multiple reference panels with different SNP sets (Howie et al., 2009). IMPUTE2 uses a Markov chain Monte Carlo (MCMC) framework and separates the phasing and imputation stages into two steps by alternatively estimating haplotypes at SNPs genotyped in both the study sample and reference panel and imputing genotypes at SNPs genotyped in only the reference panel (Howie et al., 2009). In the phasing step information from both the reference panel and the study sample is used – more information than is used by most other imputation methods. This increases overall accuracy, but due to the multiple iterations, IMPUTE2 is generally slower than IMPUTE version 1 (Howie et al., 2009).

IMPUTE (1 and 2) shows generally favourable performance compared to other commonly used imputation tools. IMPUTE and MaCH (Li et al., 2010) outperformed fastPHASE (Scheet & Stephens, 2006), PLINK (Purcell et al., 2007; Purcell & Chang, 2014) and Beagle (Browning & Browning, 2007) in a comparison by Pei et al (2008) due to the fact that neither used haplotype clustering strategies (Pei et al., 2008). IMPUTE and MaCH again outperformed fastPHASE and PLINK in a comparison by Biernacka et al (2009), with both generating lower

imputation error rates and more reliable association test results. In another comparison with MaCH and Beagle, IMPUTE required less processing time than MaCH, but more than Beagle, and required less memory if imputation was carried out in smaller chromosomal subgroups (Ellinghaus et al., 2009). A pre-compiled version of IMPUTE is available for all major platforms and pre-prepared reference panels are also available. The produced output file is small in size and can be used immediately in association testing (Ellinghaus et al., 2009).

IMPUTE2 also outperforms other methods. It is able to use a larger reference panel and has higher specificity and sensitivity to detect copies of the minor allele at rare SNPs and is therefore efficient in imputation of rare variants (Howie et al., 2009). Compared to Beagle, one of its closest competitors, IMPUTE2 showed a haplotype best guess error rate 15-20% lower (Howie et al., 2009). Although IMPUTE2 is comparable to MaCH in terms of general performance, MaCH performed better than IMPUTE2 when using a HapMap reference panel and IMPUTE2 performed better than MaCH when using a 1000 Genomes or combined 1000 Genomes and HapMap reference panel (Nho et al., 2011). IMPUTE2 also required less time and memory than Beagle when considering the type of imputation datasets that are starting to emerge in the field (Howie et al., 2011). Another comparison of imputation performance in African Americans by Chanda et al (2012) showed that IMPUTE2 and MaCH outperformed Beagle and that IMPUTE2 was faster than MaCH. In yet another study comparing IMPUTE2, Beagle, MaCH and MaCH-Admix, MaCH and IMPUTE2 were computationally efficient and showed the highest accuracy (as measured by genotype concordance when masking and re-imputing genotyped SNPs). In addition, IMPUTE2 showed the highest accuracy (as measured by imputation quality score) when taking MAF into account and the highest quality (as measured by average r^2hat), regardless of the reference panel used (Hancock et al., 2012). One study has reported that MaCH performed slightly better than IMPUTE2 in most scenarios. This was, however, at the cost of increased computation time

(Roshyara et al., 2014). Another study, comparing IMPUTE2, minimac (Fuchsberger et al., 2015) and Beagle showed that IMPUTE2 and minimac outperformed Beagle (Liu et al., 2014).

1.5.3 Reference panel

One challenge of imputation is deciding on which reference panel to use. Early imputation methods required that a reference panel closely matching the ethnicity of the study sample was chosen (Hoffmann & Witte, 2015). Imputation uses the LD structure from the reference panel to infer the genotypes of ungenotyped SNPs (Pasaniuc et al., 2010). In theory, one would assume that an exact ethnic match would be best. Reference panels closely matching the study sample can indeed increase imputation accuracy but can also mean a decrease in diversity and therefore result in more missing genotype calls (Huang & Tseng, 2014). Choosing a reference panel is particularly challenging for populations with no available matching reference panel.

Mixed/"cosmopolitan" reference panels

More recent imputation methods are able to use a "cosmopolitan" reference panel which contains all available haplotypes from individuals of multiple ethnicities. Whether or not this is the most suitable reference panel to use in all scenarios is still debatable. It has given accurate results in a variety of populations (Huang et al., 2009; Howie et al., 2011) and could improve imputation in admixed populations (Chanda et al., 2012; Huang & Tseng, 2014), of rare variants (Howie et al., 2009; Liu et al., 2014; Howie et al., 2011) and in cases where no clear reference panel matches exist (Roshyara et al., 2016). It is reported to be advantageous to use "ancestrally inclusive" reference panels and that using a "cosmopolitan" reference panel might allow unexpected allele sharing between different populations to be captured (Howie et al., 2011).

Inclusion of more distantly related reference samples may or may not have an adverse effect on imputation. In a study in African Americans, inclusion of reference populations unrelated to African Americans didn't adversely affect imputation (Chanda et al., 2012). On the other hand, Hancock et al (2012) noted a reduction in imputation quality after addition of more distantly related reference panels, due to the introduction of low frequency SNPs which are monomorphic in some populations. They also noted that imputation of low frequency SNPs was of the highest quality when using a closely related reference panel and for other frequency SNPs was of the highest quality when using a more diverse reference panel (Hancock et al., 2012).

IMPUTE2 is able to handle these "cosmopolitan" reference panels. It uses local sequence similarity to select a custom reference panel from all available reference haplotypes for each study haplotype in each genomic region (Howie et al., 2011). In the process, it is able to ignore any unhelpful reference haplotypes (Howie et al., 2011) and is able to use the extra information more effectively than other imputation methods (Howie et al., 2009).

HapMap reference panels

In earlier studies, researchers used a HapMap (The International HapMap Consortium, 2003) reference panel for imputation. HapMap Phase 1 to 3 of the HapMap data consists of 4.1 million directly genotype SNPs in 1486 individuals from 11 locations and has facilitated imputation of about 2.5 million SNPs (Wood et al., 2013). Phase 2 data is limited in that it only covers a few ethnicities and the sample size within each ethnicity is low (Browning, 2008). Imputation with the Phase 2 reference panel produced the highest imputation accuracy in Europeans and the lowest in Africans when a single HapMap panel was used, with an improvement in accuracy when a mixed panel was used (Huang et al., 2009).

1000 Genomes reference panels

Reference panels from the 1000 Genomes Project (The 1000 Genomes Project Consortium, 2015) are being used more frequently than HapMap reference panels due to their denser SNP coverage and more diverse selection of reference populations (Hancock et al., 2012). Generation of Phase 1 1000 Genomes data involved whole-genome sequencing of 1094 individuals from a range of ethnic groups and covers more than 37 million variants, including low frequency and rare variants (Wood et al., 2013). Phase 3 1000 Genomes data was an improvement on Phase 1 in that it contained information for over 88 million variants (84.7 million SNPs, 3.6 million short insertions/deletions (indels) and 60 000 structural variants) from 2504 individuals from 26 populations (The 1000 Genomes Project Consortium, 2015). This has allowed for many more SNPs to be imputed. In a study in Europeans, although accuracy was the same for both HapMap and 1000 Genomes reference panels, 1000 Genomes Phase 1 imputation resulted in more than twice as many successfully imputed SNPs than HapMap imputation, with twice as many common SNPs, four times as many low frequency SNPs and eight times as many rare SNPs (Sung et al., 2012b). This shows that the 1000 Genomes reference panels can be useful across a range of MAFs, with particular interest in discovering rare variants (Sung et al., 2012b, 2012a). Another study using a 1000 Genomes reference panel allowed for stronger signals of association at known loci as well as the discovery of novel associations (Wood et al., 2013). Using a 1000 Genomes Phase 3 reference panel is thought to significantly increase the number of imputed variants even more compared to using HapMap or other 1000 Genomes reference panels (Zheng-Bradley & Flicek, 2016) and is therefore an appropriate choice in any current or future studies.

There is some concern that the quality of 1000 Genomes data is lower than that of HapMap data due to the depths of sequencing reads being low. Low quality SNPs are usually removed, but due to the significantly higher number of SNPs to

start with, the number remaining is still higher than if a HapMap reference panel is used (Sung et al., 2012b).

1.5.4 Factors affecting imputation performance and accuracy

In addition to the composition of the reference panel, several other factors influence imputation accuracy. These include the size of the reference panel, study sample size and SNP density of the region to be imputed, LD and MAF of the SNPs to be imputed.

Reference panel size

Although the running time and computational burden is increased, the larger the reference panel (in terms of number of samples and SNPs) or the more haplotypes in the reference panel, the higher the accuracy (Pei et al., 2008; Nho et al., 2011; Howie et al., 2011; Huang & Tseng, 2014). This is because a larger reference panel brings an increase in length of the haplotype stretches shared between study samples and reference panel samples and are easier to identify unambiguously with a larger reference panel (Li et al., 2009). Zhang et al (2011) recommended a reference sample size of more than 100 for homogenous populations and more than 200 for admixed populations. The improved accuracy with larger reference panels is specifically evident for less common variants (Liu et al., 2012; Roshyara & Scholz, 2015; Zheng et al., 2015a; Howie et al., 2011), although very rare variants remain difficult to impute. A study in African Americans showed that the addition of more distantly related populations to increase reference sample size improved imputation quality, but this was mostly for SNPs present in populations more closely related to African Americans (Hancock et al., 2012).

Study sample size and SNP density

The study sample size seems to have little effect on imputation accuracy (Zhao et al., 2008; Zhang et al., 2011). It does, however, influence the accuracy of haplotype estimation during phasing as, especially with IMPUTE2, phasing information is gained from all other individuals in the dataset, so having a larger sample increases haplotype estimation accuracy (Howie et al., 2009). SNP density in the study sample also influences imputation, with a higher SNP density in the region to be imputed leading to better imputation quality due to there being more neighbouring SNPs and therefore a higher LD (Zhang et al., 2011; Wang et al., 2012; Howie et al., 2012). This is particularly evident with low frequency and rare variants (Zheng et al., 2015a; Spencer et al., 2009). The quality of imputation of a region with less than 100 SNPs is likely not very stable and a window size of more than 500 SNPs or one SNP every 2 kilobases is recommended (Zhang et al., 2011). In addition, larger chromosomes are easier to impute due to lower recombination rates and subsequent higher levels of LD (The International HapMap Consortium, 2005) and an ungenotyped rate of less than 50% is favourable for good imputation (Zhang et al., 2011). It is also thought that the imputation quality depends more on the number of SNPs in the study sample rather than their quality (Roshyara et al., 2014).

LD and MAF

LD between genotyped and ungenotyped SNPs and the MAF of ungenotyped SNPs also influence imputation accuracy, with MAF having a weaker effect on accuracy than LD (Nho et al., 2011). Several studies have shown that stronger LD and lower MAF of ungenotyped SNPs together result in improved imputation accuracy (Pei et al., 2008; Biernacka et al., 2009; Huang et al., 2009; Nho et al., 2011). The influence of MAF seems to be lower for high LD regions and higher for low LD regions (Pei et al., 2008). One study found an increase in power with higher MAF under medium to high LD levels (Pei et al., 2010). The difference in

imputation accuracy between lower and higher MAF markers was shown to be highly variable in African populations (Huang et al., 2009).

Lower frequency or rare SNPs are, in general, more difficult to impute due to lower coverage in study samples, lower degrees of LD and more challenging haplotype reconstruction (Liu et al., 2012; Sung et al., 2012a; Chanda et al., 2012; Band et al., 2013; Zheng et al., 2015a). Inclusion of rare SNPs may affect the phasing step and lead to a lower imputation quality. On the other hand, including rare SNPs in the reference panel may also improve imputation quality of rarer SNPs (Liu et al., 2012). Rare indels can be more accurately imputed than rare SNPs (and *vice versa* for common variants) (Liu et al., 2014).

Other factors

Another factor which may influence imputation is the filtering or removal of SNPs during QC. Little or no SNP filtering seems to be favourable for imputing small to moderately sized datasets to keep the LD structure between SNPs intact (Roshyara et al., 2014). Other factors causing low imputation accuracy include low variant heterozygosity, high sequence similarity to other genomic regions, high GC content and segmental duplication (Liu et al., 2014). In addition, some regions such as GWAS loci for haematological measurements and immune system diseases are enriched with low imputability regions (Liu et al., 2014).

1.5.5 Imputation in African populations

Imputation accuracy across different populations varies, with the highest accuracy generally seen in European populations and the lowest seen in Africans (Zhao et al., 2008; Huang et al., 2009). Most of the studies involving imputation in populations of African ancestry have been carried out in African Americans, which are an admixed population and therefore don't give a true indication of imputation performance in all Africans.

Imputation in Africans, in general, is more challenging due to their high genetic diversity and lower levels of LD, which leads to more difficult haplotype estimation and a reduced imputation accuracy (Howie et al., 2011; Huang et al., 2011). This increases the required sample size to maintain power in imputation based GWAS in Africans (Huang et al., 2011). Early reference panels added to the challenge in that African populations were poorly captured (Huang et al., 2011), but imputation was improved with availability of more populations of African ancestry (Huang et al., 2011). In a recent study in three African populations (Kenya, Malawi and Gambia) imputation was able to capture most common variants in the three populations, but the accuracy and calibration of confidence (measured by the info score) was still lower than in European populations (Band et al., 2013).

One advantage of the greater genetic diversity in Africans is the improved accuracy of imputing rare variants in African ancestry populations compared to other populations. This is due to the larger number of haplotypes resulting from the diversity and the subsequent improved chance that a rare variant is tagged by one of the haplotypes (The 1000 Genomes Project Consortium, 2015).

In addition to the shorter LD blocks seen with all individuals of African ancestry, African Americans have proven challenging to use in imputation due to their high levels of admixture. There is no ideal single population reference panel to use for admixed populations (Sung et al., 2012a). Ideally for admixed populations, the reference panel should include samples from all the ancestral populations making up the admixed population under investigation (Zhang et al., 2011). In this respect, mixed or “cosmopolitan” reference panels have improved imputation performance in African Americans (Hao et al., 2009; Hancock et al., 2012).

1.5.6 Pre-phasing

Pre-phasing has become a popular option as it ultimately speeds up imputation as haplotypes can first be statistically estimated and missing genotypes then imputed directly into the inferred haplotypes (Howie et al., 2011, 2012). Pre-phasing significantly reduces the computational burden of imputation as a study sample only needs to be phased once and can therefore undergo multiple runs of imputation quickly with, for example, each new reference panel update, without re-phasing. In addition, it is faster to match phased study haplotypes to a single reference panel haplotype than to match unphased study haplotypes to more than one reference haplotype (Howie et al., 2012). Despite imputation of low frequency SNPs being generally less accurate, pre-phasing tends to achieve competitive accuracy at these variants (Howie et al., 2012). A tool called SHAPEIT has been developed to carry out the haplotype estimation/phasing step and can be used successfully in conjunction with IMPUTE2.

Chapter 2: STUDY PARTICIPANTS AND DATA QUALITY CONTROL

2.1 Study participants

This study used DNA samples and data from a mixed-sex subset of Bt20 participants and their female caregivers, with phenotype data taken from the year 17/18 (i.e. the year that the participants turned 17/18) and year 13 (i.e. the year that the participants turned 13) collection time points, respectively. The female caregivers are the female relatives who care for the Bt20 participants and include mothers, grandmothers, aunts and sisters.

Written informed assent was obtained from the participants in conjunction with written informed consent from the caregivers prior to each blood sample collection. Individuals also re-consented to the collection of data at each data collection session. Ethical clearance was previously obtained from the University of the Witwatersrand Human Research Ethics Committee (Medical) for collection of DNA samples and phenotype data from this cohort (M010556). Further ethical clearance was obtained to use the samples to identify genetic risks associated with obesity (M120647) and blood pressure (M1411116) in a black South African population. Ethics certificates, other relevant agreements/letters and the most recent consent forms are in **Appendix A**.

2.2 Phenotyping

Weight was measured to the nearest 0.1 kg using a digital scale with participants wearing light clothes and no shoes. Standing height was measured to the nearest 0.1 cm using a wall-mounted stadiometer (Holtain, Crosswell, UK). BMI was calculated as weight (kg) divided by height squared (m^2) (Kagura et al., 2015).

BP readings were taken using an Omron 6 automated machine (Kyoto, Japan). Measurements were taken with participants in a seated position. After five minutes of sitting in a resting position, three measurements were taken at intervals of two minutes. The first reading was discarded, in case of possible "white coat syndrome", and an average of the second and third measurements was calculated and used in all analyses (Kagura et al., 2015).

WC was measured at the level of the widest girth above the navel and hip circumference (HC) was measured as the widest part of the buttocks. Both were measured using a soft measuring tape to the nearest 0.5 cm with the participants in a standing position. WHR was calculated as WC divided by HC.

Body fat (g) and lean mass (g) were measured using dual-energy X-ray absorptiometry (DXA) (Hologic, Malborough, MA, USA) as per the guidelines recommended by the International Society of Clinical Densitometry (Gordon et al., 2008). Measurements were taken with participants wearing light clothes and with all jewellery or other metal objects removed. Percentage body fat was calculated as fat mass divided by total body mass (g).

2.3 DNA sample preparation and genotyping

DNA sample normalization and genotyping was carried out as part of another PhD study (Sahibdeen, 2016). Samples were extracted using the salting out method (Miller et al., 1988). They were then normalized to a concentration of 50ng.µl^{-1} following quantification using either a Tecan Infinite® 200 PRO NanoQuant or PicoGreen® dsDNA Quantitation Reagent. DNA is currently stored in the Division of Human Genetics at the National Health Laboratory Service (NHLS), Braamfontein, South Africa. Genotyping was conducted at the UC Davis Genome Center (California, USA) using the Metabochip (Illumina, San Diego, CA, USA) (Voight et al., 2012) *[See Preface]*. Samples were sent for genotyping in two

phases (the female caregiver samples were genotyped in May 2013 and the Bt20 participants were genotyped in November 2013). A few duplicate samples from both datasets were sent with the unique samples to make sure that there weren't "chip effects" from one chip to another and also to ensure that samples were genotyped consistently. Genotypes were called using GenomeStudio Software for Illumina (v2011.1) and a custom DNAtech cluster file and output was provided as final reports in the forward strand orientation.

2.4 Data quality control

2.4.1 Genotype data quality control

Prior to analysis of the genotype data, a QC process was carried out separately for the two datasets to convert the data into a usable format and to remove any SNPs or samples that may have affected downstream use of the data. The steps involved in the QC process and the number of SNPs and samples removed in each step are outlined in **Figure 2.1** and detailed below. Most of the steps were performed using PLINK (v1.9) (Purcell et al., 2007; Purcell & Chang, 2014), unless otherwise stated.

Before QC, the Bt20 participant dataset comprised data from 1248 samples (1240 unique samples and 8 duplicates) and the female caregiver dataset consisted of data from 1034 samples (1033 unique samples and 1 duplicate). Of the 2273 unique samples, there were 975 caregiver-participant pairs. Both datasets initially contained genotype data for 196725 SNPs.

Initial QC steps

Raw genotype data from GenomeStudio forward report files was converted into transposed PLINK files (.tped/.tfam) using a Python script written by Dr Tom Gaunt (University of Bristol, UK) (**Appendix B**). These files were then converted

into binary PLINK files (.bed/.bim/.fam) prior to initial removal of SNPs with 100% missing genotype data and samples with more than 20% missing genotype data.

SNP QC

SNPs with high **missingness**, low **MAF** and those failing **HWE** were then removed. Based on plots of *maximum missing rate* versus *number of SNPs remaining in the study*, SNPs with more than 2% missing data were removed. Default values of 0.01 for MAF (i.e. SNPs with MAF less than 0.01 were removed) and 0.00001 ($1x10^{-5}$) for HWE (i.e. SNPs with a HWE p-value of less than $1x10^{-5}$ were removed) were selected for both datasets (Anderson et al., 2010). In both datasets, the high number of SNPs removed due to low MAF was due to many of the SNPs being monomorphic in our black South African population. In both cases, some SNPs failed more than one of the SNP QC criteria.

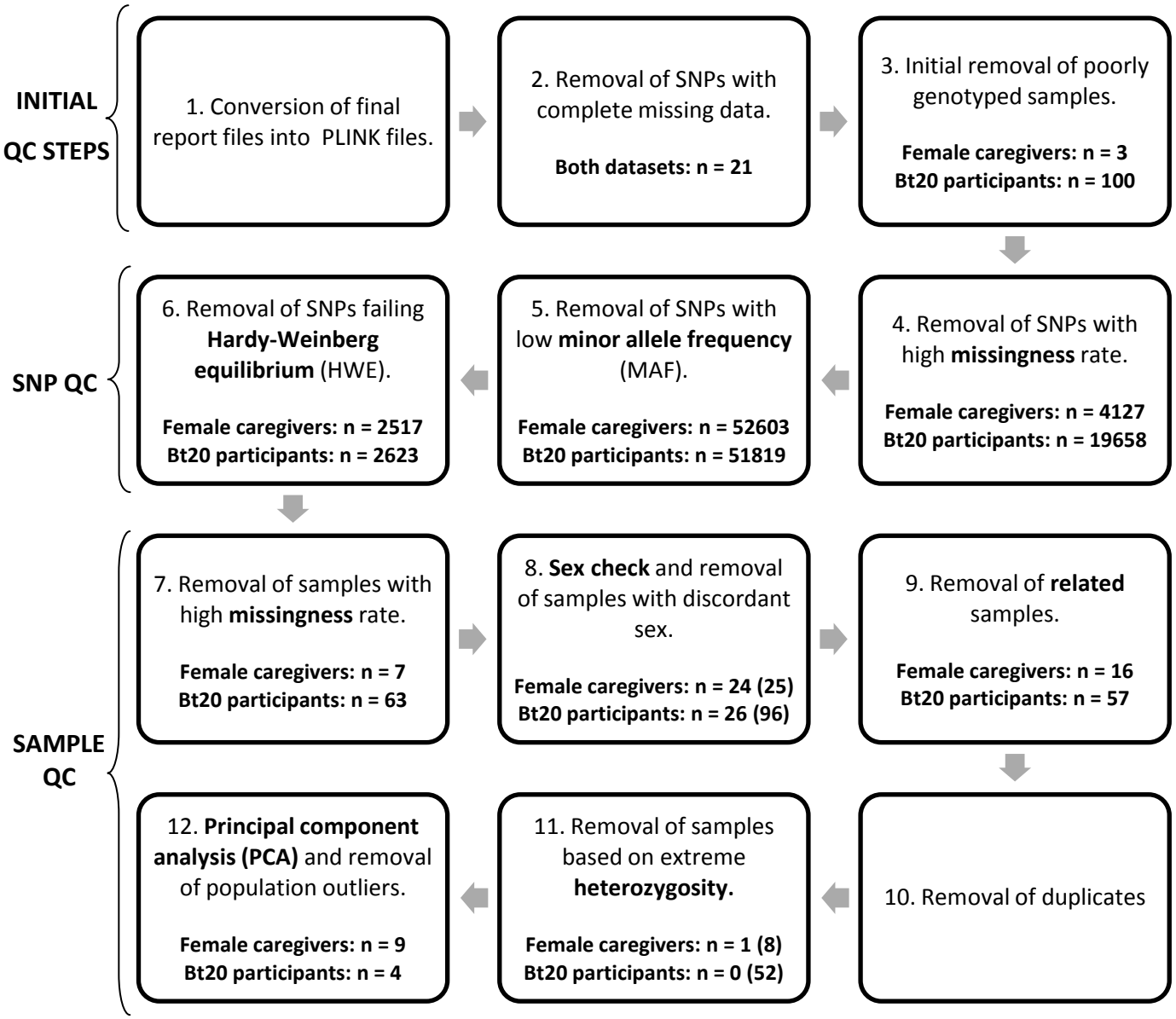


Figure 2.1 The steps involved in the genotype data QC process. The process was divided into initial QC steps, SNP QC and sample QC. In some sample QC steps, samples had already been removed in a previous step. If this was the case the number of samples actually removed in the step is given with the total number of "problem" samples identified in the steps shown in brackets.

Sample QC

A second step to remove individuals with a high **missingness** rate was performed. Based on plots of *maximum missing rate* versus *number of individuals removed*, individuals with more than 2% missing data in the female caregiver dataset and more than 3% missing data in the Bt20 participant dataset were removed.

A **sex check** was carried out on the original files to check for discrepancies between the genotype data and the recorded sex in the phenotype data, possibly arising from mislabelling of DNA samples or incorrect reporting of sex during recruitment. The sex check is done by calculating the homozygosity rate across all SNPs in the X chromosome for each individual and comparing it to the expected rate (males: homozygosity rate >0.8; females: homozygosity rate<0.2) (Anderson et al., 2010). Individuals are identified as being problems if there are discrepancies between the genotype and phenotype data or if the sexes are inconclusive/unspecified according to the genotype data. Problem individuals not already removed in a previous step were removed from the latest QC file for each dataset.

To avoid an introduction of bias into the analysis, any duplicate individuals or cryptically **related** individuals (individuals that are unknown second-degree relatives or higher) were removed from the dataset. These individuals were identified by calculating an identity by state (IBS) metric for each pair of individuals. This metric is based on the average proportion of alleles in common at the genotyped autosomal chromosome SNPs. Identity by decent (IBD) scores were then estimated from the IBS data. IBD = 1 for duplicates or monozygotic twins, IBD = 0.5 for first-degree relatives, IBD = 0.25 for second-degree relatives and IBD = 0.125 for third-degree relatives. The initial step involved LD pruning of the datasets so that only independent SNPs were included in the calculations. This was followed by generation of IBD scores (PI_HAT column in the .genome

file) and then removal of one of each of the pair of individuals with IBD score (PI_HAT) > 0.1875. This is a common value to use and is halfway between the expected IBD for third- and-second-degree relatives. The number of individuals removed due to cryptic relatedness is quite high. This may be due to the fact that all samples were taken from the same very integrated community and the chances of second- or third-degree relatives may be high. (Anderson et al., 2010)

All **duplicate** individuals should have been identified in the previous step by an IBD=1 between individuals in a duplicate pair. This was, however, not the case for all duplicates and, as we were no longer sure of their identities, all suspected duplicates were removed at this point.

Individuals with **extreme mean heterozygosity** were also removed, if not already removed in previous steps. All individuals are expected to have a certain proportion of heterozygous genotypes. Mean heterozygosity/observed heterozygosity rate can be calculated across all individuals as (number of non-missing genotypes (N) – observed number of homozygous genotypes (O))/N) and individuals are removed in cases of extreme heterozygosity. Excessive mean heterozygosity could indicate DNA sample contamination while reduced mean heterozygosity could indicate inbreeding. To decide on a reasonable threshold at which to exclude individuals based on extreme heterozygosity, a graph of *observed heterozygosity rate per individual* (x-axis) versus *proportion of missing SNPs per individual* (y-axis) was plotted. A .het file was generated, using genotype files produced from the initial removal of SNPs with complete missing data, to give the observed number of homozygous genotypes (third column) and the number of non-missing genotypes (fifth column) per individual. The proportion of missing SNPs per individual was obtained following a calculation of missingness (the sixth column in the generated .imiss file is the proportion of missing SNPs per individual) and all individuals with a heterozygosity rate ± 3 standard deviations from the mean were excluded (Anderson et al., 2010).

Population stratification, where genotype differences may be due to different population origins rather than actual disease risk differences (Cardon & Palmer, 2003), is a source of bias that can be introduced into analysis. This was dealt with in two ways: removal of extreme **population outliers** during this QC step (using visual cut-offs from the principal component analysis (PCA) plots) and further control for population stratification during analysis, when deemed necessary through visualisation of quantile–quantile (Q-Q) plots, by including principal components (PCs) as covariates *[See Chapter 4]*. The initial step involved LD pruning of the datasets and PCA was then run using SMARTPCA (Patterson et al., 2006), an EIGENSTRAT program. PCA plots were drawn in Genesis (http://www.bioinf.wits.ac.za/software/genesis) to first identify and remove outliers and then to visualise the genetic structure of the datsets in the context of other African populations following all QC steps [*See 2.6.2 below*].

2.4.2 Phenotype data quality control

The female caregiver phenotype data also went through a QC process and corrections were made where inconsistencies were found between the original questionnaires and captured data. As the year 17/18 Bt20 participant data were captured using electronic questionnaires, no QC could be carried out on this data.

2.5 Merging of datasets and update to Build 37

The female caregiver and Bt20 participant datasets were merged for association analysis in PLINK and then pruned to remove any non-overlapping SNPs between the datasets. During the merging step, nine SNPs were found to have the same base pair positions as other SNPs in the dataset and were removed at this point.

SNP names and coordinates were also updated, manually and using PLINK `--update-name` and `--update-map` commands, from Build 36 (Metabochip SNPs

are mapped to Build 36) to Build 37, as is required by the imputation programs [*See Chapter 5*].

2.6 Post-quality control data characteristics

Following all QC steps, there were 971 individuals and 140649 SNPs remaining in the female caregiver dataset and 976 individuals and 127764 SNPs remaining in the Bt20 participant dataset, with 763 caregiver-participant pairs. The merged and pruned dataset contained 1947 individuals and 125906 SNPs.

2.6.1 Descriptive statistics

Descriptive statistics of the individuals remaining after QC in the individual datasets are shown in **Table 2.1**.

2.6.2 Population structure

PCA plots drawn using each of the datasets in combination with other African populations (**Figures 2.2 and 2.3**) showed good clustering of the Bt20 samples with southeastern Bantu-speakers (SEB) and other black Sowetans (BSO), who are also southeastern Bantu-speakers. Also evident is the very clear separation of Bt20 individuals and individuals from the Yoruba in Ibadan, Nigeria (YRI); the Luhya in Webuye, Kenya (LWK) and the Maasai in Kinyawa, Kenya (MKK), indicating differences in genetic structure across African populations.

Table 2.1 Descriptive statistics of the individuals remaining after QC in the individual datasets.

	Bt20 Participants (n=976)						Female Caregivers (n=971)		
	Females (n=456)			Males (n=518)					
	Mean (SD)	Median	Range	Mean (SD)	Median	Range	Mean (SD)	Median	Range
Age (years)	17.9 (0.4)	17.9	17.3-18.9	17.9 (0.4)	17.9	17.3-18.9	41.9 (8.6)	41.0	18.0-84.0
Weight (kg)	59.5 (12.9)	56.9	36.1-136.6	59.2 (9.8)	57.8	38.5-128.4	76.1 (17.0)	74.7	36.3-135.7
Height (m)	1.6 (0.1)	1.6	1.4-1.8	1.7 (0.1)	1.7	1.5-1.9	1.6 (0.1)	1.6	1.2-1.8
BMI ($kg.m^{-2}$)	23.3 (4.8)	22.3	15.5-53.1	20.3 (3.1)	19.8	14.9-48.3	30.4 (6.6)	30.0	16.6-58.8
SBP (mmHg)	115.7 (10.5)	115.5	87.0-172.0	120.5 (11.4)	119.8	81.5-170.0	117.4 (20.9)	112.5	77.0-206.5
DBP (mmHg)	72.4 (9.1)	71.5	44.5-129.5	70.7 (8.3)	70.0	50.0-102.0	76.6 (12.5)	75.0	46.5-126.5

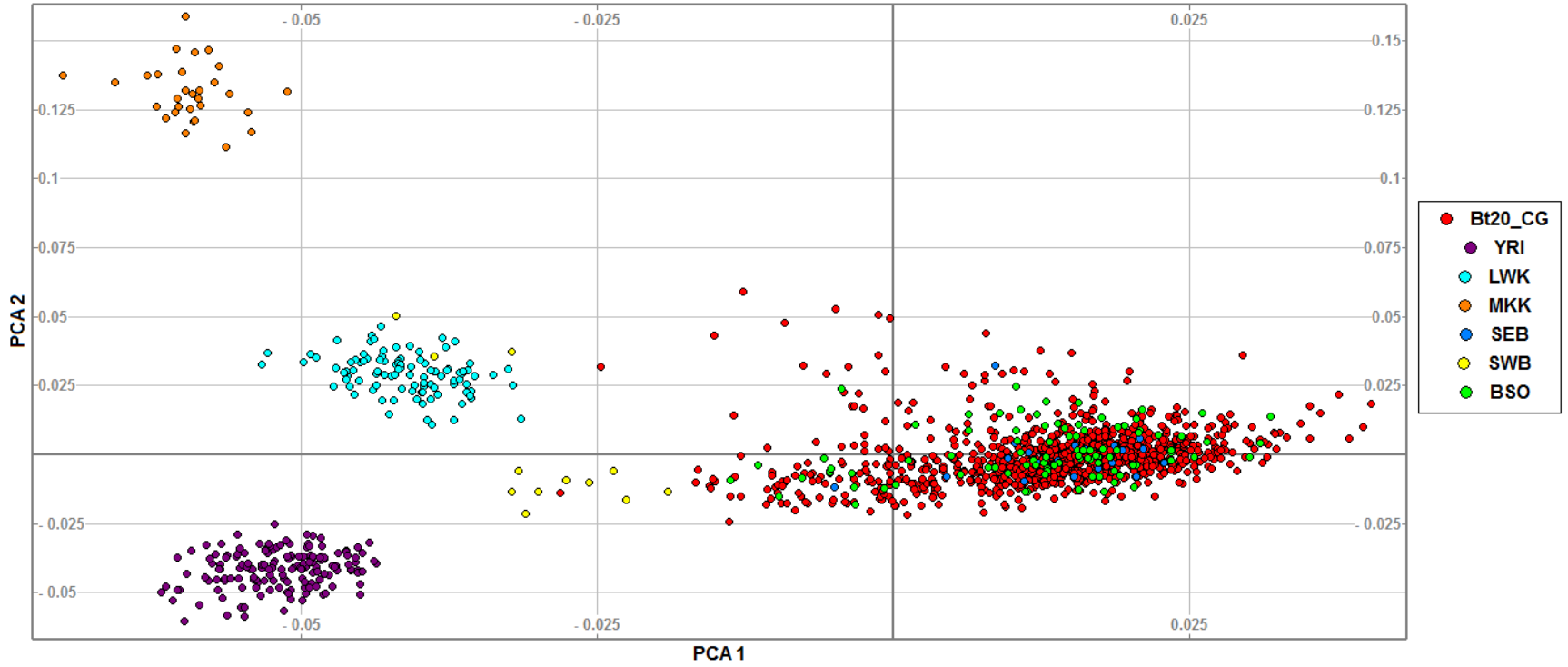


Figure 2.2 PCA plot of population structure in the female caregivers and other African populations. PC1 and PC2 are shown. Bt20_CG = female caregivers; YRI = Yoruba in Ibadan, Nigeria; LWK = Luhya in Webuye, Kenya; MKK = Maasai in Kinyawa, Kenya; SEB = southeastern Bantu-speakers; SWB = southwestern Bantu-speakers and BSO = black Sowetans.

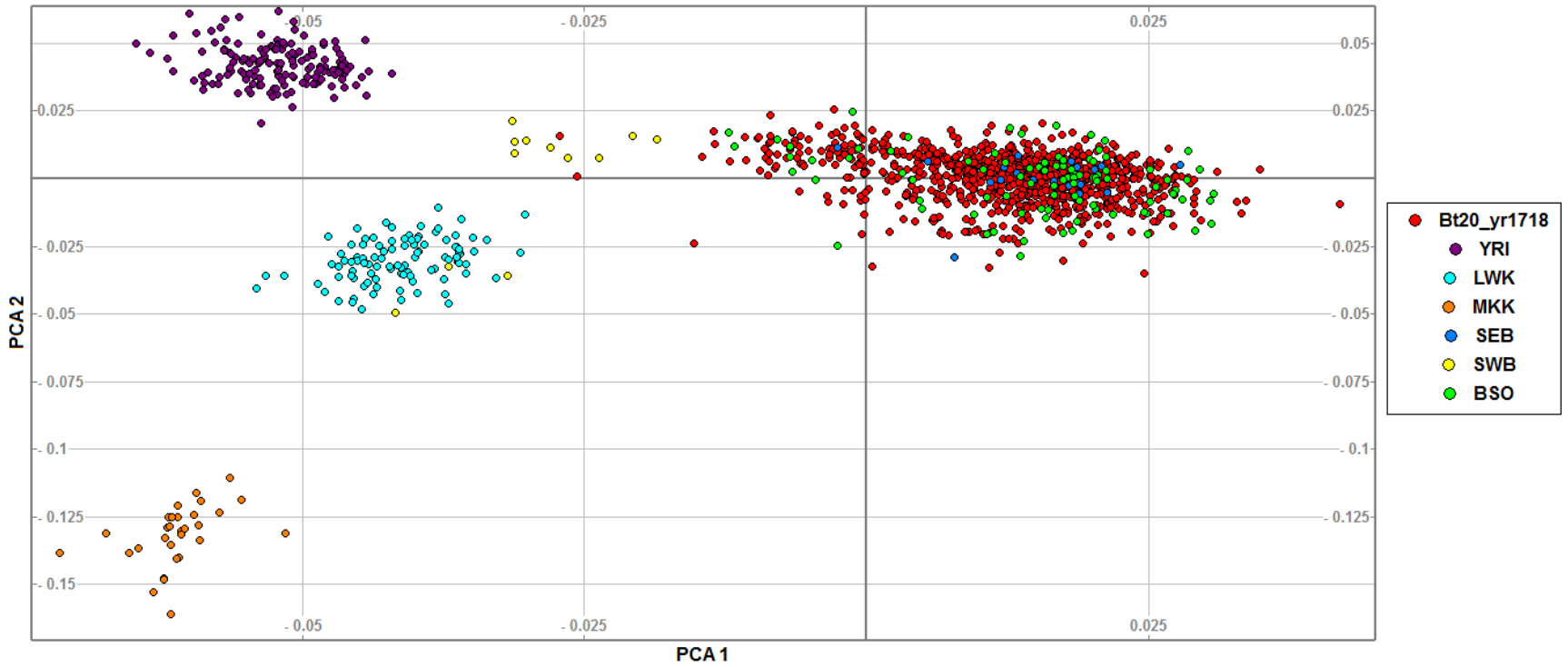


Figure 2.3 PCA plot of population structure in the Bt20 participants and other African populations. PC1 and PC2 are shown. Bt20_yr1718 = Bt20 participants; YRI = Yoruba in Ibadan, Nigeria; LWK = Luhya in Webuye, Kenya; MKK = Maasai in Kinyawa, Kenya; SEB = southeastern Bantu-speakers; SWB = southwestern Bantu-speakers and BSO = black Sowetans.

Chapter 3: DESIGN OF A QUERYABLE CARDIOMETABOLIC DATABASE

Advances in technology, high-throughput experiments and multi-disciplinary research have resulted in a huge growth in the amount and complexity of biological data being generated in recent years. There is an increasing need for the development of databases to efficiently store and manage the data and make it more accessible. Typically, publicly available datasets and the associated annotation data are accessed through internet-based browsers that offer a user-friendly interface to a database, which makes complex queries simple to execute. Research project-specific data are, however, not commonly stored in this way.

This chapter focuses on the development of MetaboBTT, a queryable cardiometabolic database, and accompanying user interface, which houses the Bt20 phenotype, Metabochip SNP annotation and association analysis data from a current project focused on identifying risk factors for cardiometabolic disease in South Africans. Although the current study involved a cross-sectional analysis, the database has been designed for longitudinal data (i.e. data from multiple data collection time points within the cohort can be incorporated and queried). A secondary output is a model/structure for research groups with similar data (phenotype, genotype, annotation and association analysis data) to use for implementation of their own databases. The aim is for the data to be more easily accessible and queryable for useful information by all members of a research group to ultimately accelerate biological knowledge discovery.

3.1 Database construction

The database was developed in MySQL (v 5.7). Tables were constructed using standard SQL code (**Appendix D**) and data was entered into the database tables

using a MySQL database connector (MySQLdb) in Python (v 2.7.12) (**Appendix E**). The database is a relational, multiuser (supports multiple users at the same time), centralised (data located at a single site) database.

3.1.1 Database content

Each individual recorded in the database has a unique Individual ID (with a suffix 'C' or 'CG' for the Bt20 participants and female caregivers, respectively). This ID is the primary key for all tables containing phenotype data.

As previously mentioned, the Bt20 cohort is a longitudinal cohort consisting of data collected at multiple time points since its inception. **Phenotype data** currently present in the database is that which is appropriate to the current study (i.e. year 17/18 data for the Bt20 participants and the year 13 data for the female caregivers). Tables have, however, been constructed for other data collection time points for the Bt20 participants (year 5, year 7, year 9/10, year 11/12, year 13, year 14, year 15, year 16, year 19, year 20) and can be populated with the data when obtained. Phenotype data only exists in the database for individuals with available genotype data.

SNP annotation/Metabochip data exists for all 196725 SNPs on the Metabochip and includes Build 36 and 37 SNP information and locations of the SNPs within the genome. Information is also given about whether or not the SNP remained in the dataset for either or both of the datasets after QC.

All **association analysis data** for the available phenotypes under investigation are recorded and includes corrected and uncorrected p-values and beta values/odds ratios (OR) and a list of covariates (up to three can be listed in the current design) that were corrected for in each analysis. The datasets available can be analysed as individual or merged datasets – results for all possible scenarios are recorded.

The phenotype, Metabochip and association analysis data are arranged in tables in the database as indicated in the relational model in **Figure 3.1**. The QC'ed **genotype data** exists as separate binary PLINK format files *(.bed/.bim/.fam*) for both the Bt20 participants and the female caregivers and can be accessed on request.

3.1.2 Database design evaluation

The database has been normalised to a certain extent as there are no repeating groups in individual tables, there exist separate tables for each set of related data and each set of data is identified by a primary key. There are some dependencies in the SNP tables (for example the Build 37 base pair position is dependent on the Build 37 SNP ID), meaning that the current database is in first normal form. The phenotype data was split into data which doesn't change over time and data which does change over time, rather than all being included in one table. Separate tables then exist for the changeable phenotype data at each data collection time point. This allows for easy extension of the database to include additional time points, without possible corruption of already existing data. The phenotype data which exists in the tables is data which is currently relevant to the larger project. SQL allows for the addition of extra columns to the tables, making it easy for additional phenotypes to be included. The different association analysis scenarios are also split across multiple tables for easier initial input of the data into the tables and more efficient querying. The database is largely free of redundant data with the only information repeated across tables being the Individual ID (for the phenotype tables) and the Build 36 SNP IDs (SNPID36) (for the SNP tables).

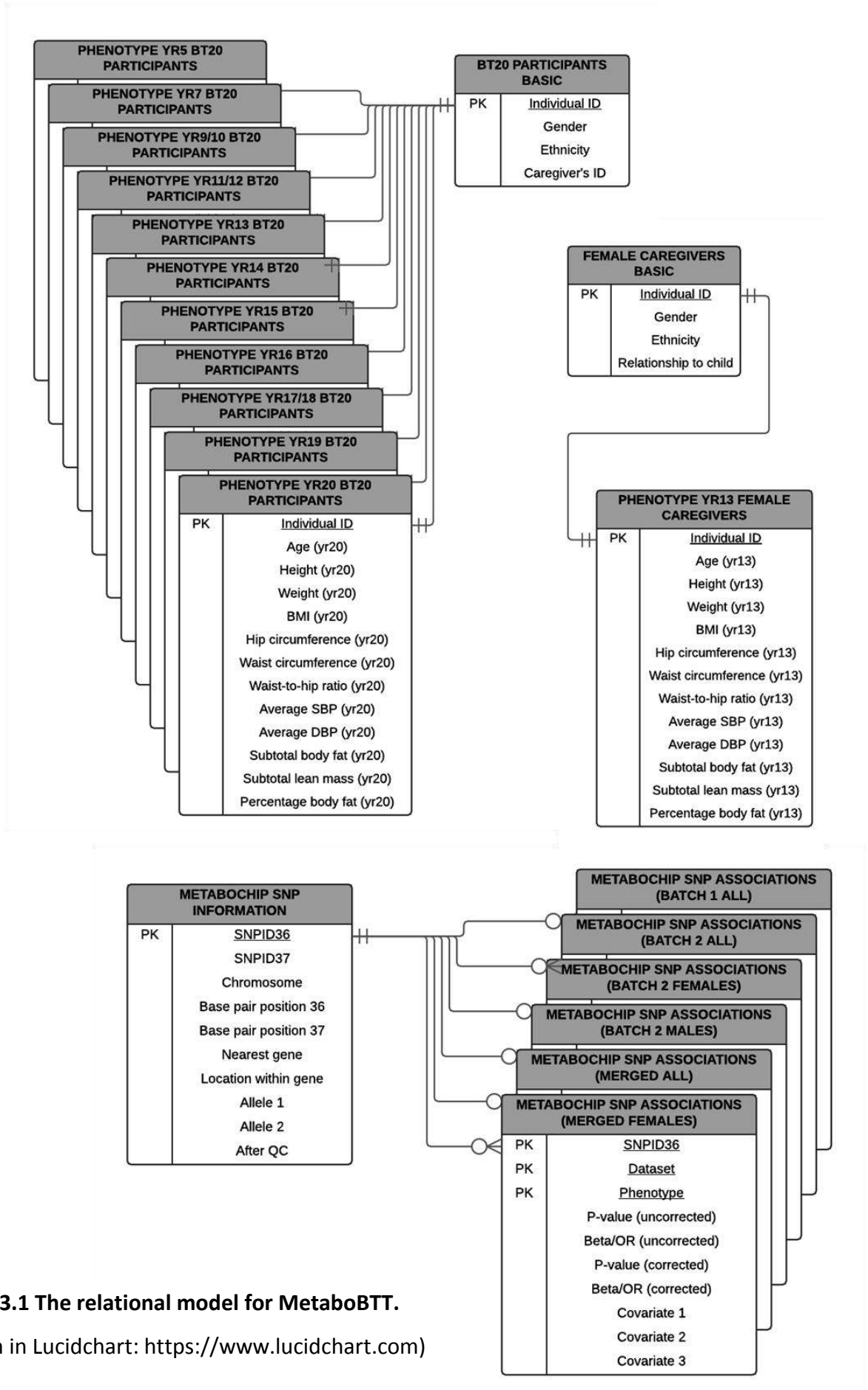


Figure 3.1 The relational model for MetaboBTT.

(Drawn in Lucidchart: https://www.lucidchart.com)

3.2 Web interface

The data within the MySQL database can be accessed and queried via a web interface (which will be made available at http://www.bioinf.wits.ac.za/software/metabobtt to members of the research group with permitted access) constructed using CSS, HTML and PHP code (**See Web_interface.pdf available at https://github.com/ LieslH/Liesl-Hendry-PhD-Code**).

3.2.1 Access and security

The interface is username and password protected, therefore allowing only those individuals allocated access to view and query the data. Individuals who request permission to access the database will be assigned a username and will then be able to register their own passwords via a link available on the login page. The passwords are stored as hashed passwords along with the usernames in a separate MySQL table within the database and users can be added or removed when necessary. Once logged in, the home page (**Figure 3.2**) allows for easy navigation to other pages. The interface also has a 'session timeout' feature which ensures that any user inactive for longer than eight minutes will be automatically logged out of the interface and will be required to login again to query the database further. Alert/error messages have also been included to alert users when an inappropriate selection or input is made or if users forget to select required options, therefore further preventing some misuse of the database.

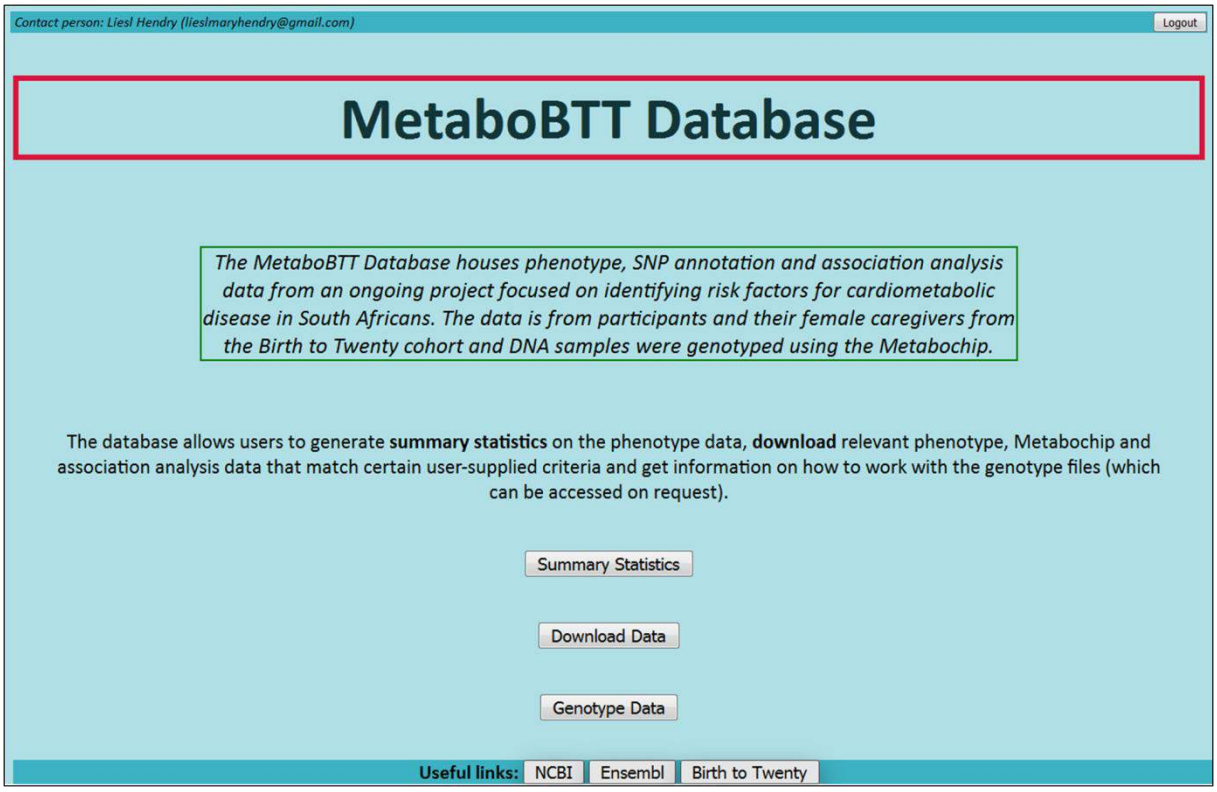


Figure 3.2 Home page of the MetaboBTT Database web interface.

3.2.2 Function

Users can access the database from the web interface to generate **summary statistics** (basic and complex counts and average/minimum/maximum) on the phenotype data, **download** relevant phenotype, Metabochip and association analysis data that match certain user-supplied criteria and get information on how to work with the **genotype** files in PLINK (**Figure 3.3**).

Summary statistics

Users are able to perform basic counts (count of number of individuals, males or females in a particular dataset) (**Figures 3.4 and 3.5**), complex counts (count of number of individuals, males or females in a particular dataset and data collection time point matching up to three criteria) (**Figures 3.6 and 3.7**) and calculate or retrieve the average, minimum or maximum of a particular

phenotype for all, male, female or specific individuals in a particular dataset and time point (**Figures 3.8 and 3.9**). The summary statistics are printed to the screen.

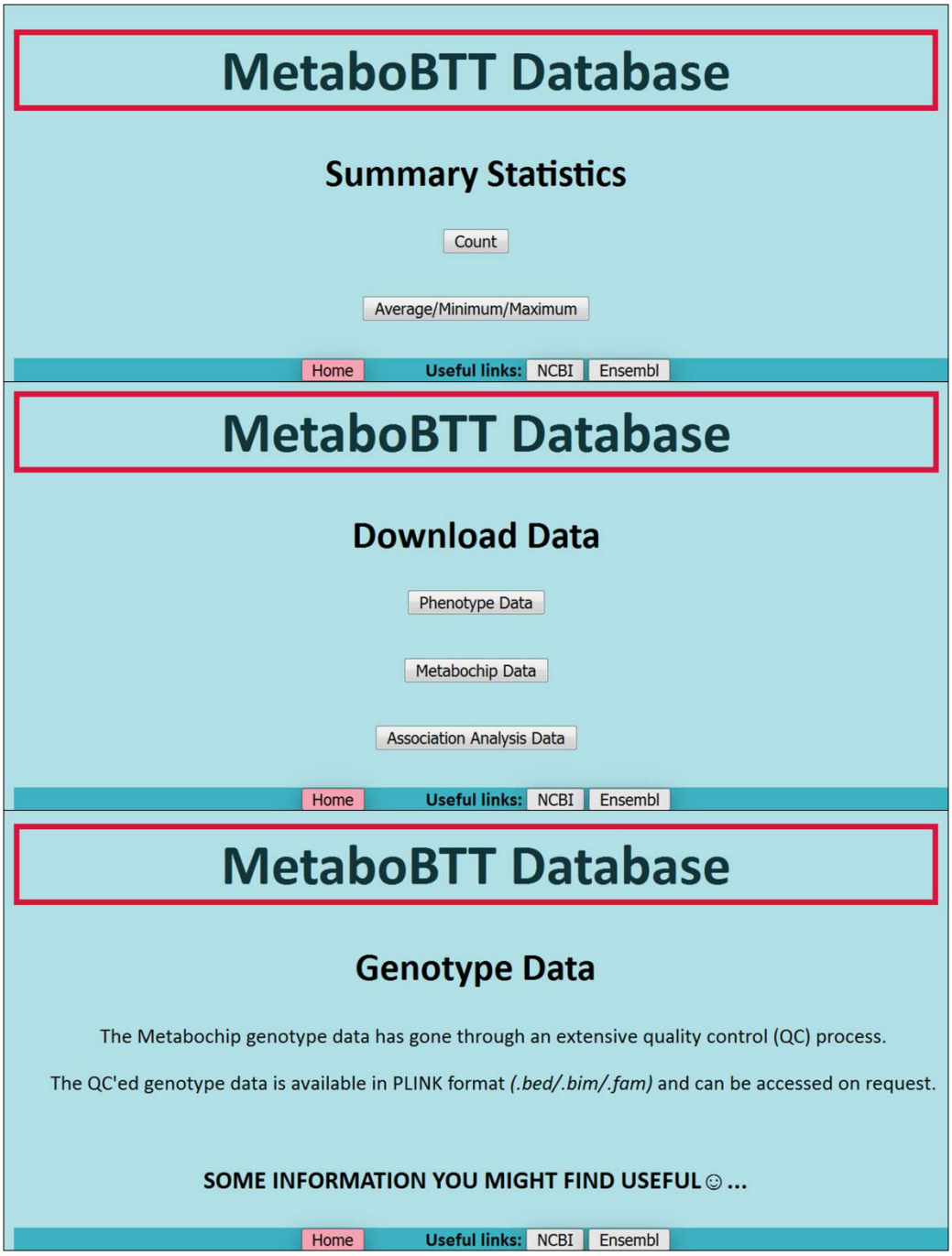


Figure 3.3 Summary statistics, data download and genotype data landing pages.

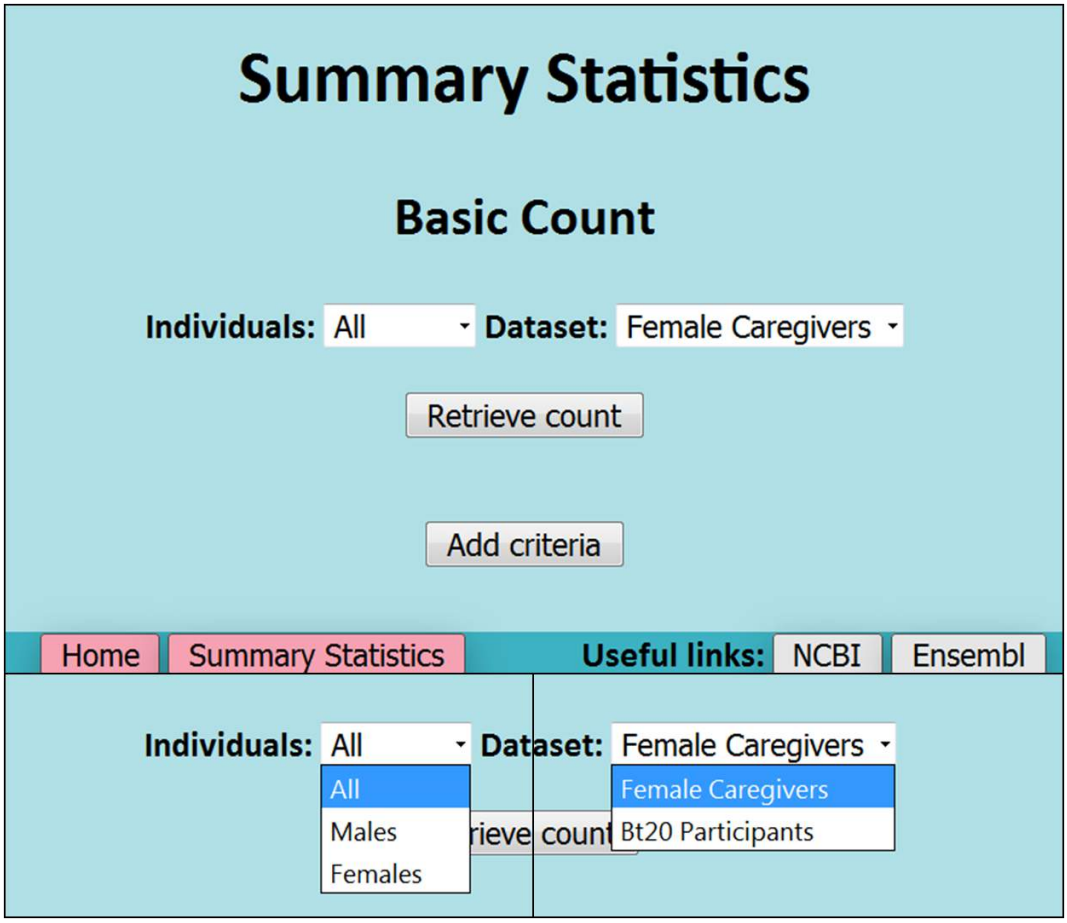


Figure 3.4 Layout of the basic count page.

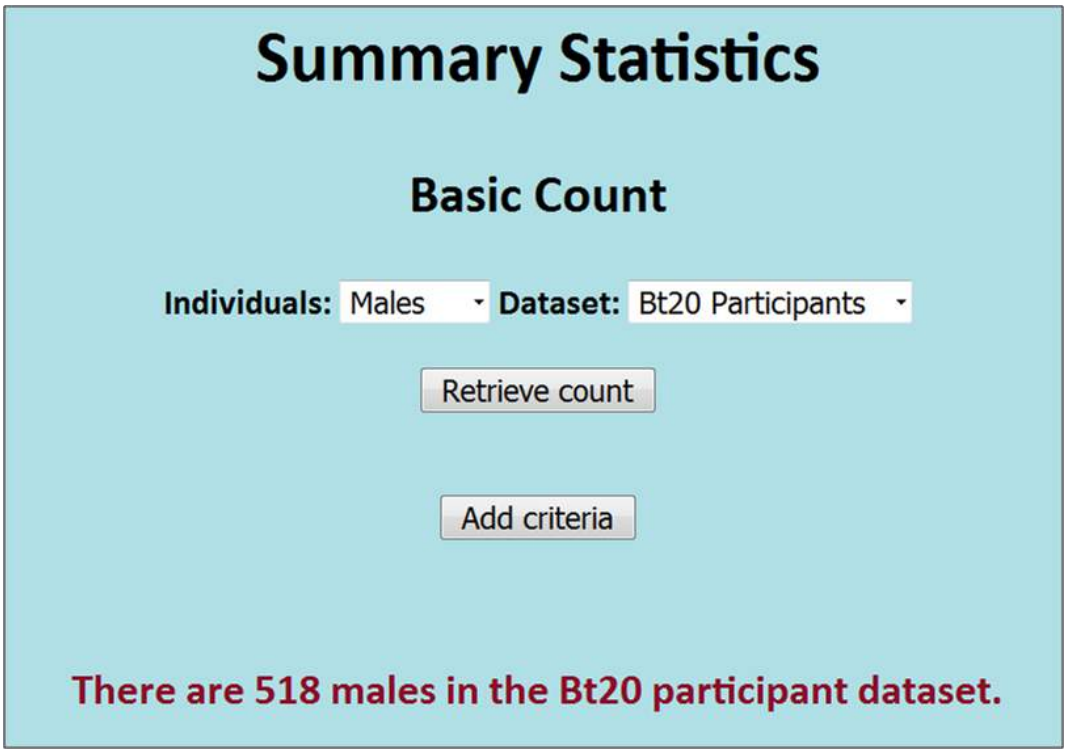


Figure 3.5 An example of the input and output of a basic count query.

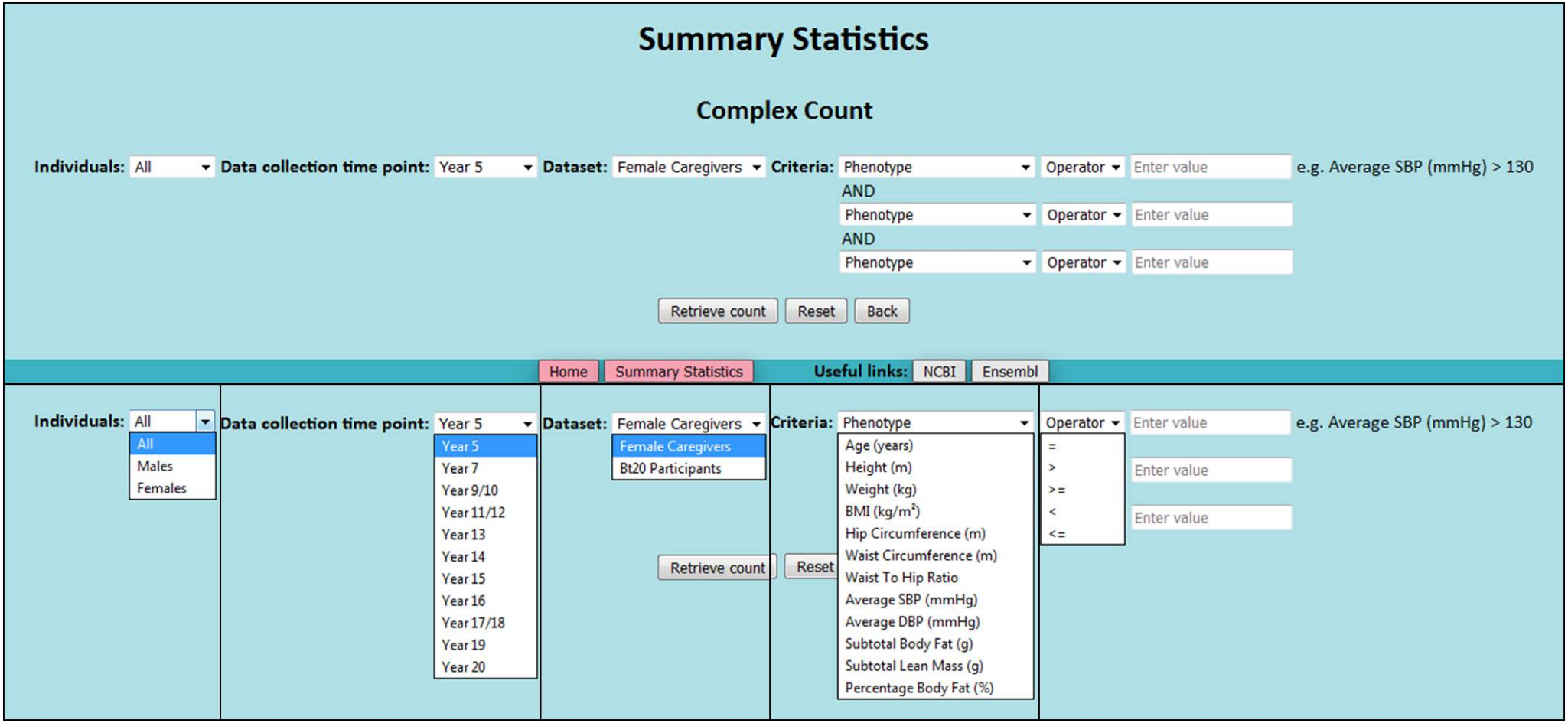


Figure 3.6 Layout of the complex count page.

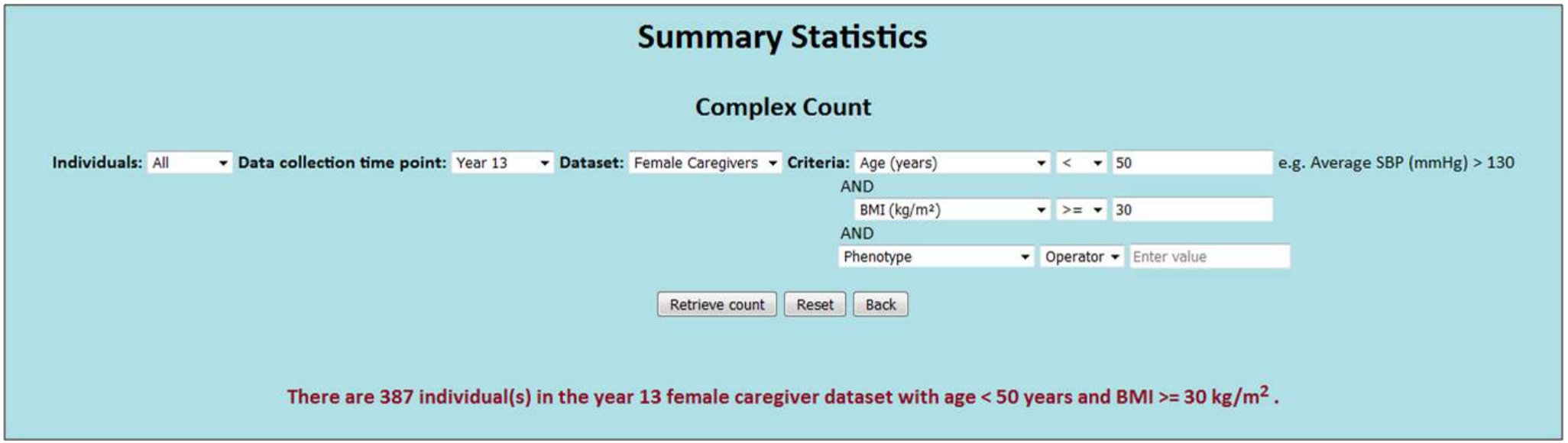


Figure 3.7 An example of the input and output of a complex count query.

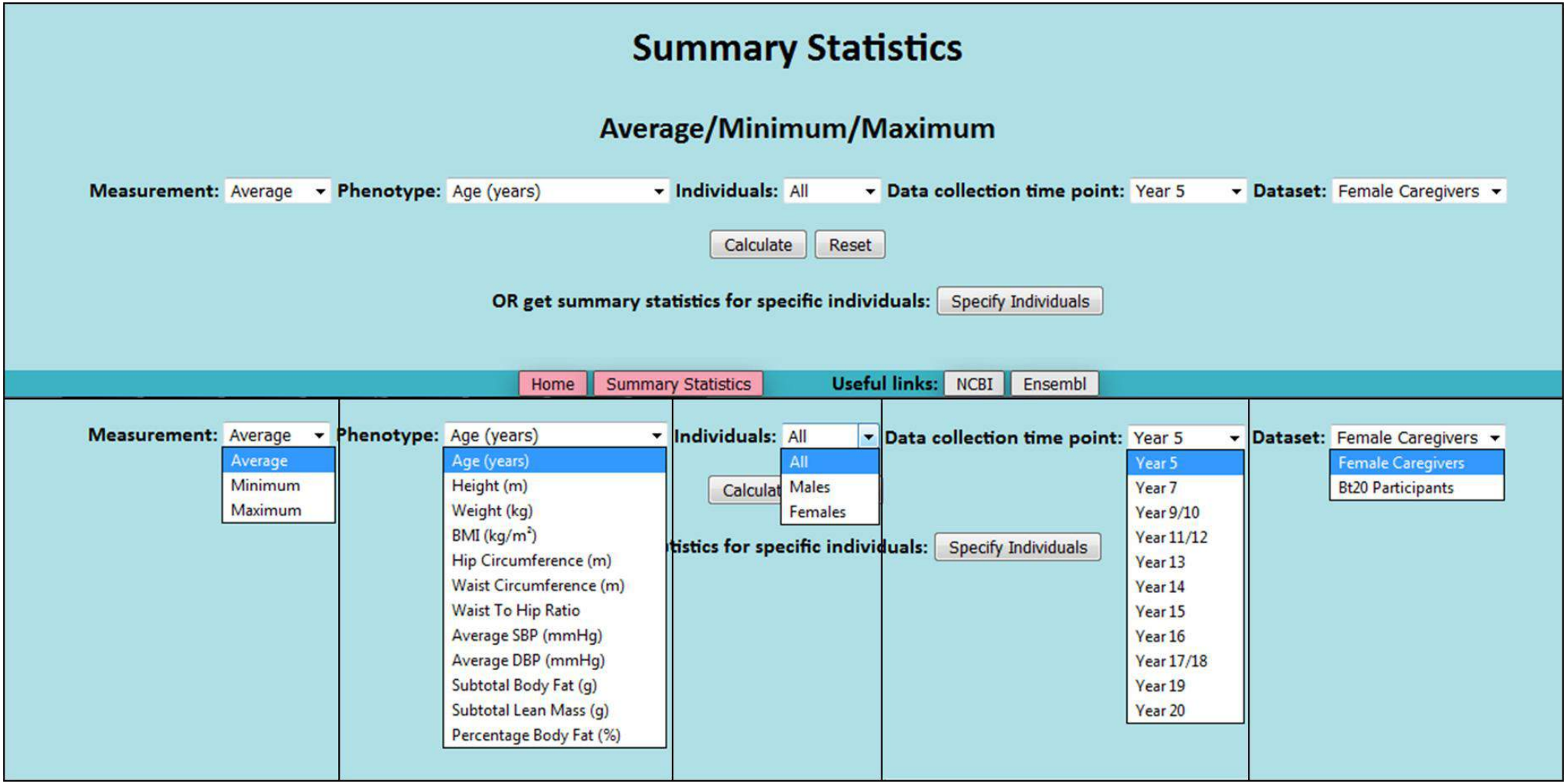


Figure 3.8a Layout of the average/minimum/maximum page where the user can specify all/male/female individuals.

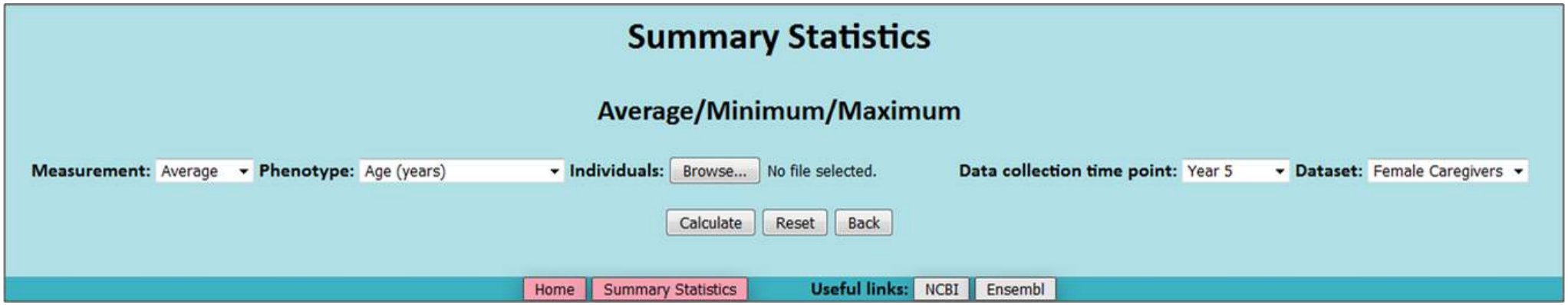


Figure 3.8b Layout of the average/minimum/maximum page where the user can specify specific individuals by uploading a file of Individual IDs.

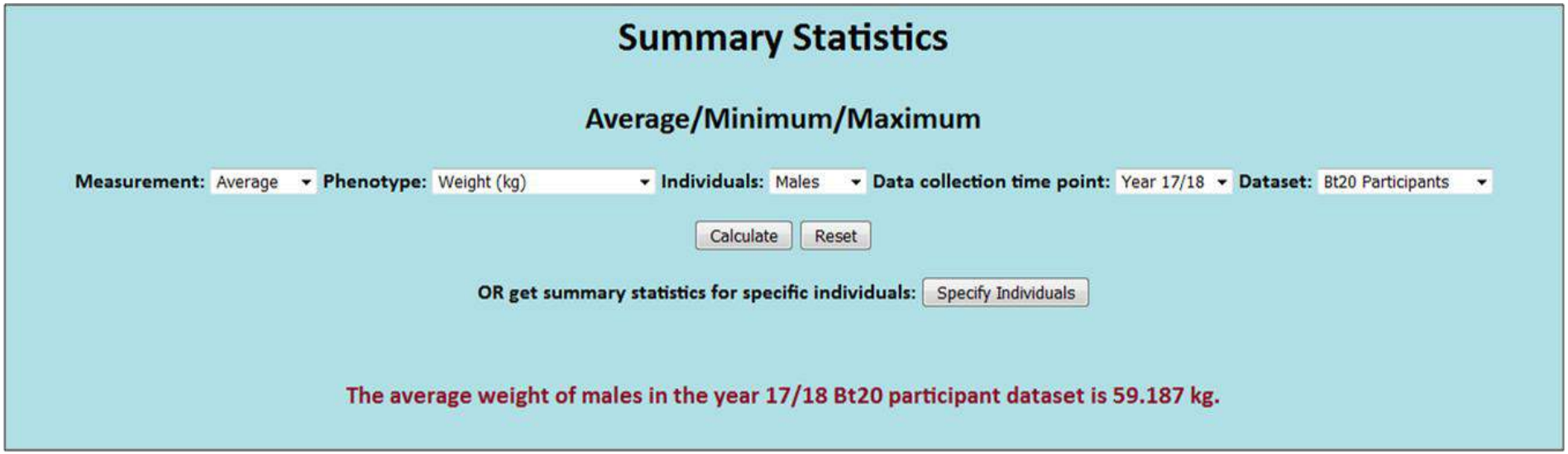


Figure 3.9 An example of the input and output of an average/minimum/maximum query.

Data download

Users are able to download:

- Phenotype data for specific or all/male/female individuals from a specific dataset and data collection time point matching certain criteria (**Figures 3.10, 3.11 and 3.12**). Separate pages exist for the female caregiver and Bt20 participant datasets.
- SNP-related data filtering by SNPID (Build 36 or 37), base pair position (Build 36 or 37), gene, chromosome or location within the gene (**Figures 3.13 and 3.14**).
- Association analysis-related data for specific phenotypes in all or a specific subset of the dataset, filtering by SNPID (Build 36 or 37) or gene and/or p-value threshold (**Figures 3.15 and 3.16**).

The data can be printed to the screen (by selecting 'Print to screen') or downloaded as a comma separated values (CSV) file (by selecting 'Save to file (.csv)').

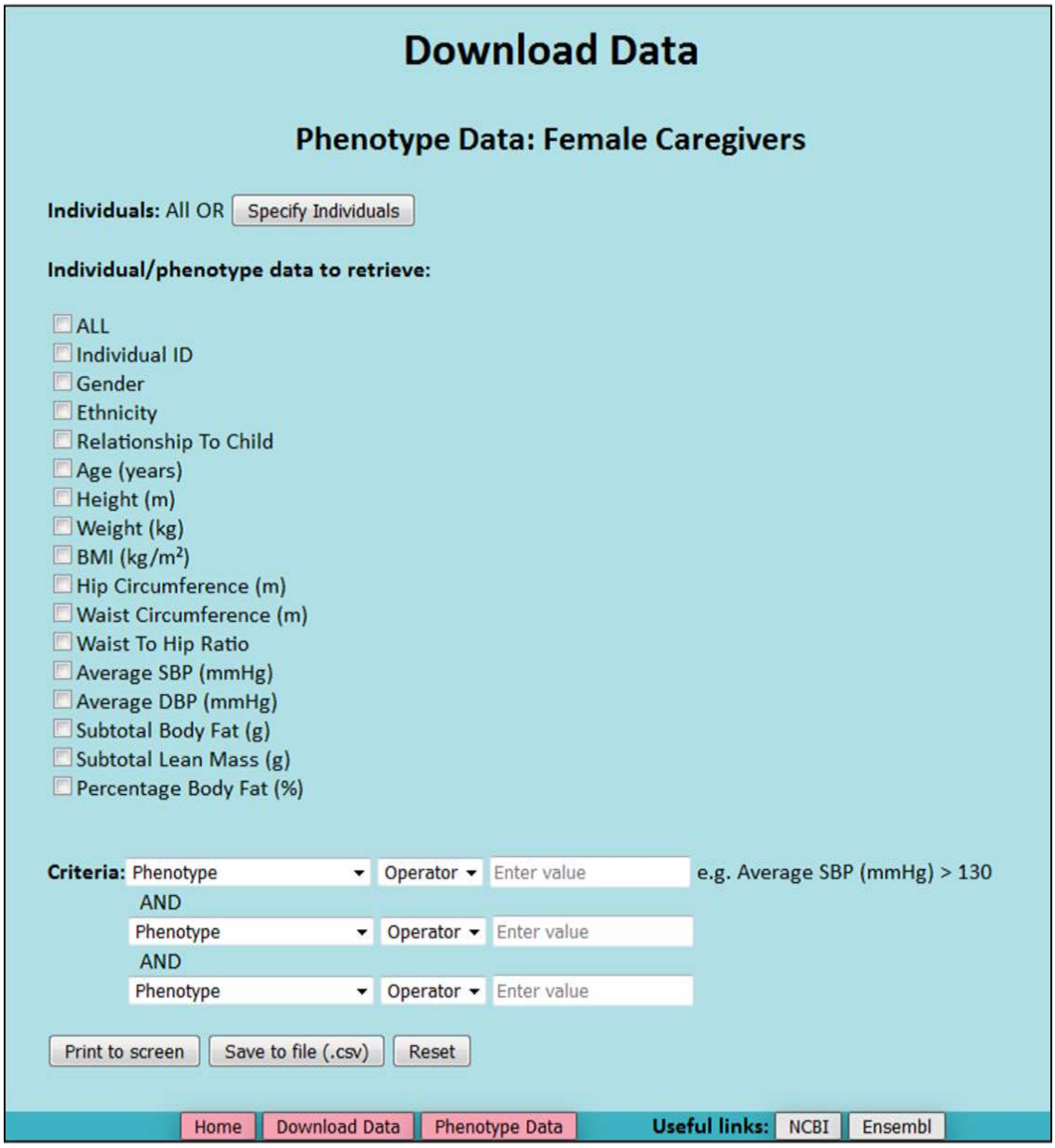


Figure 3.10a Layout of the female caregiver phenotype data download page where the user can specify all individuals. The phenotype and operator options for the criteria are the same as in the summary statistics.



Figure 3.10b Layout of the female caregiver phenotype data download page where the user can specify specific individuals by uploading a file of Individual IDs.

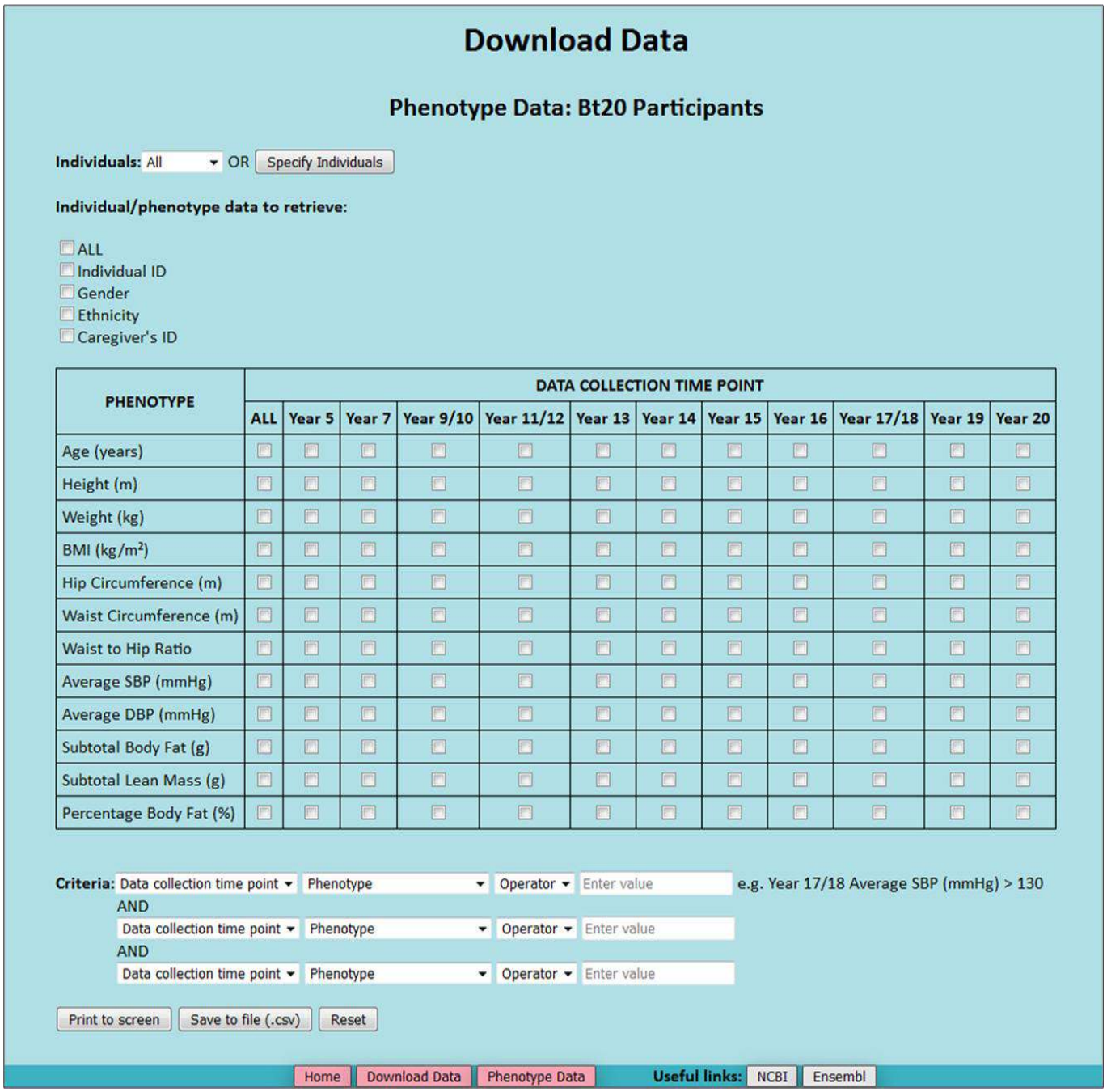


Figure 3.11a Layout of the Bt20 participant phenotype data download page where the user can specify all/male/female individuals. The phenotype and operator options for the criteria are the same as in the summary statistics and the data collection time point options for the criteria are the same as in the larger table.

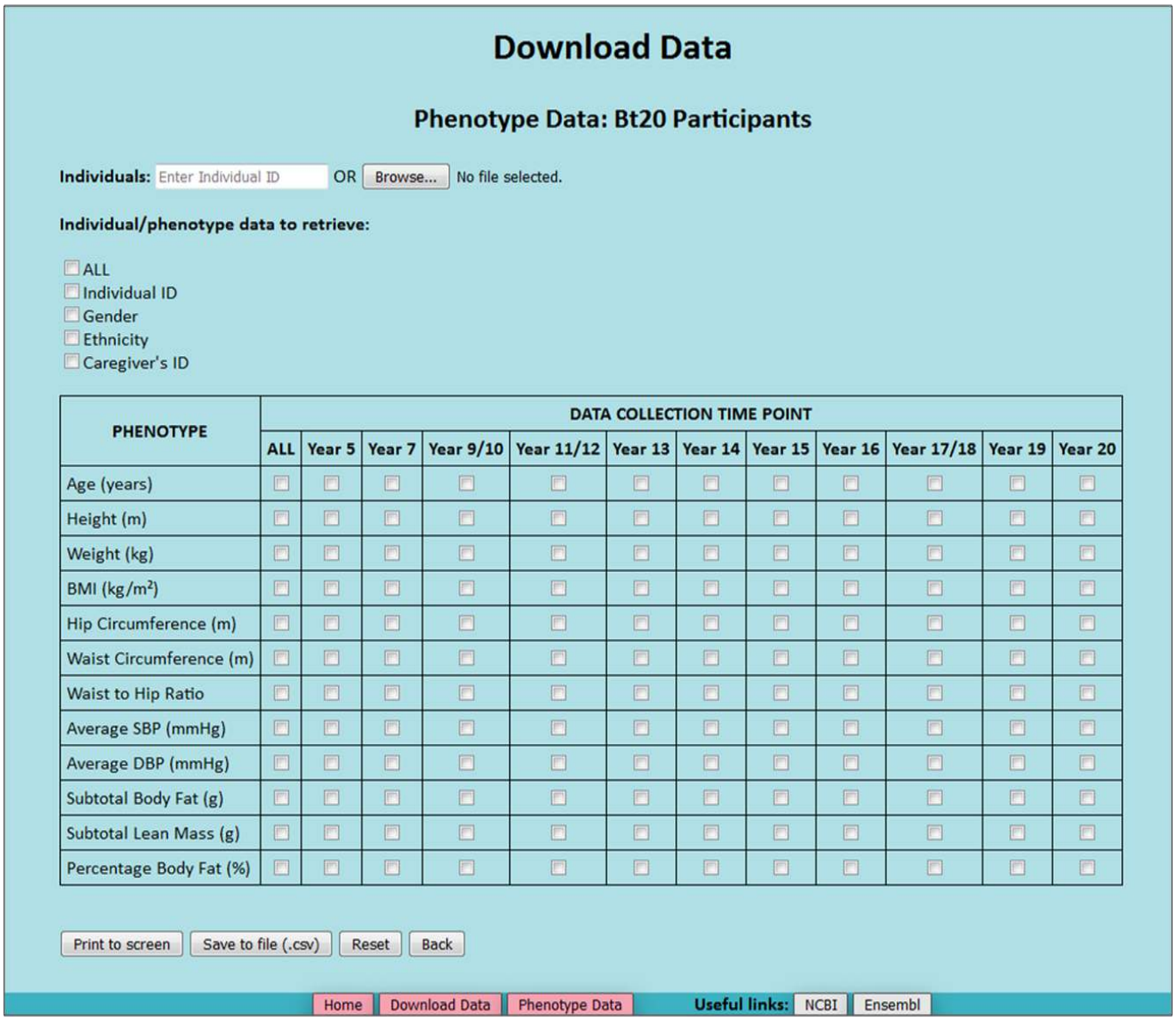
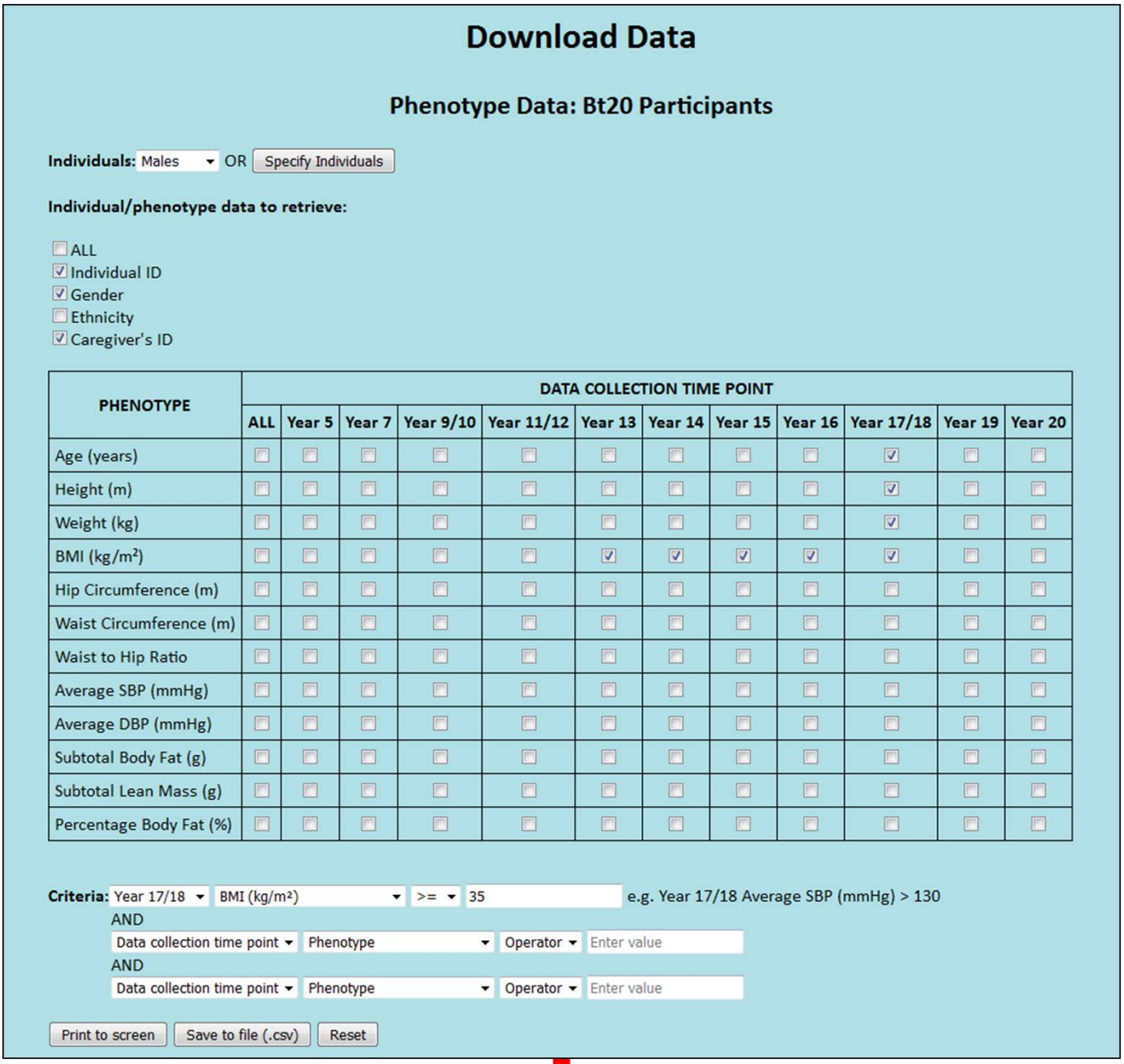


Figure 3.11b Layout of the Bt20 participant phenotype data download page where the user can specify specific individuals by uploading a file of Individual IDs.



2 result(s) matched the selected criteria

IndividualID	Gender	CaregiversID	Age_yr1718	Height_yr1718	Weight_yr1718	BMI_yr13	BMI_yr14	BMI_yr15	BMI_yr16	BMI_yr1718
667C	1		17.591	1.825	128.4					38.551
673C	1	667CG	17.706	1.608	125					48.343

Figure 3.12 An example of the input and output of a phenotype data download query. The empty BMI columns are due to the fact that only yr17/18 data currently exists in the database.

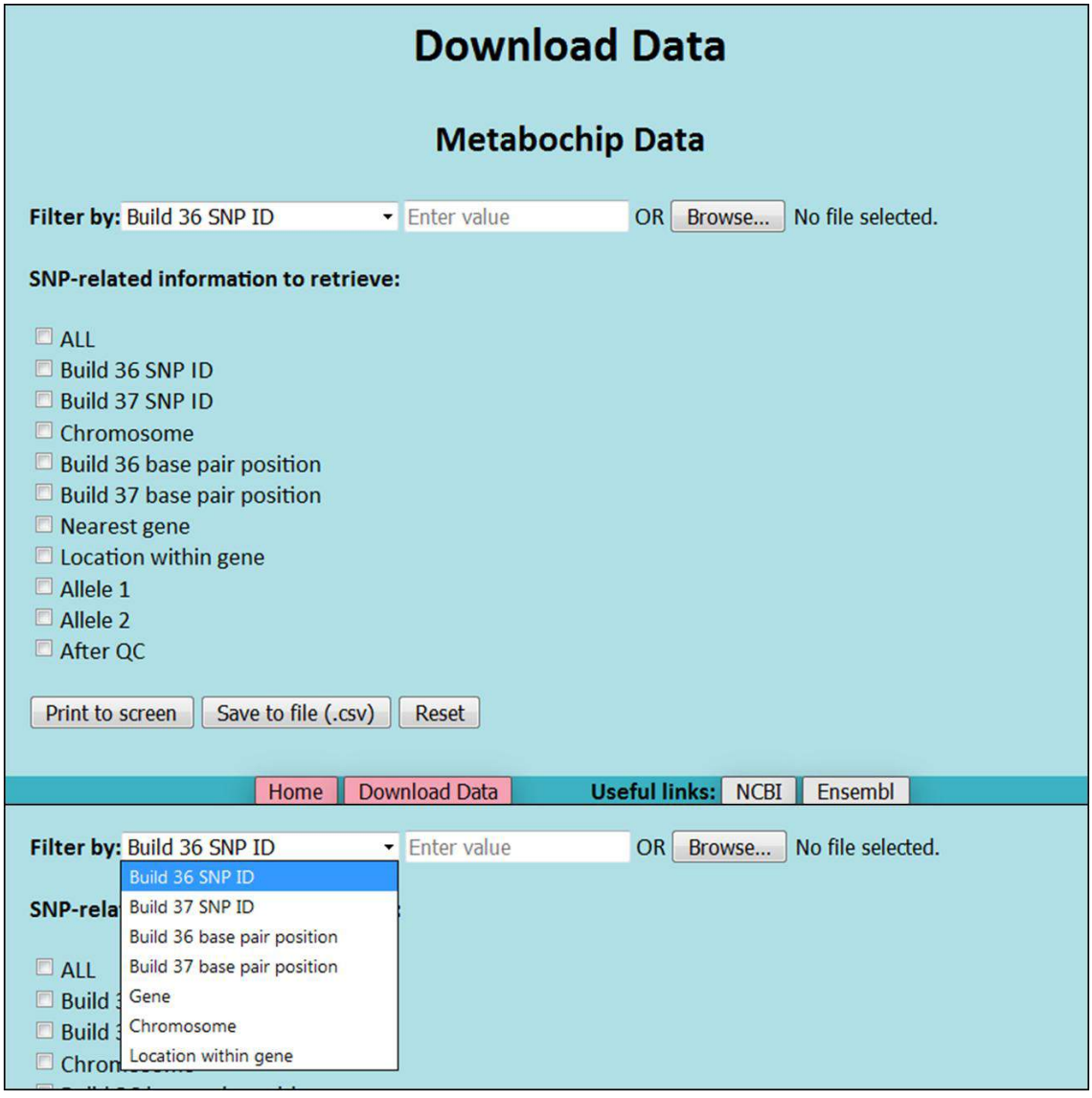


Figure 3.13 Layout of the Metabochip data download page.

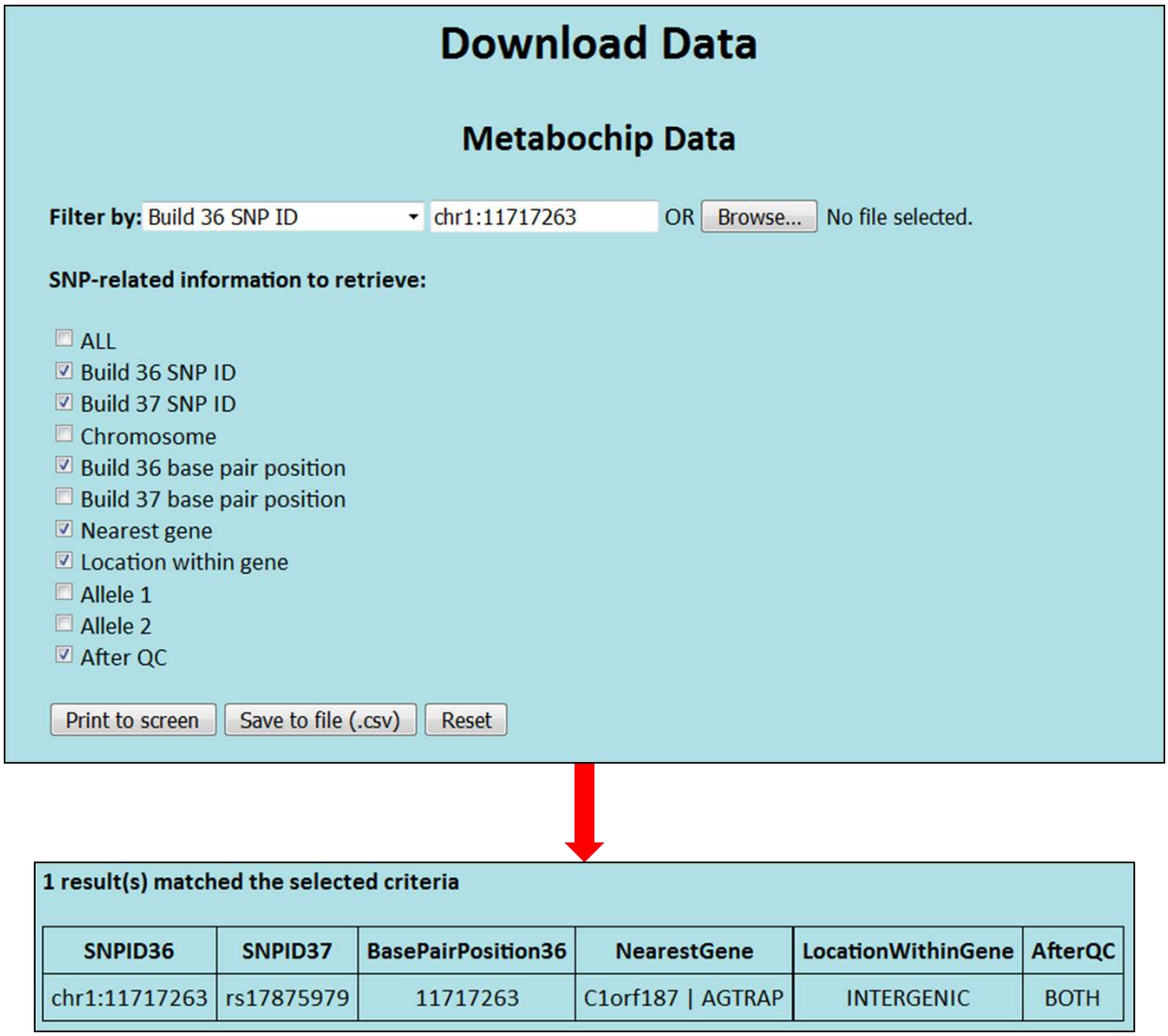
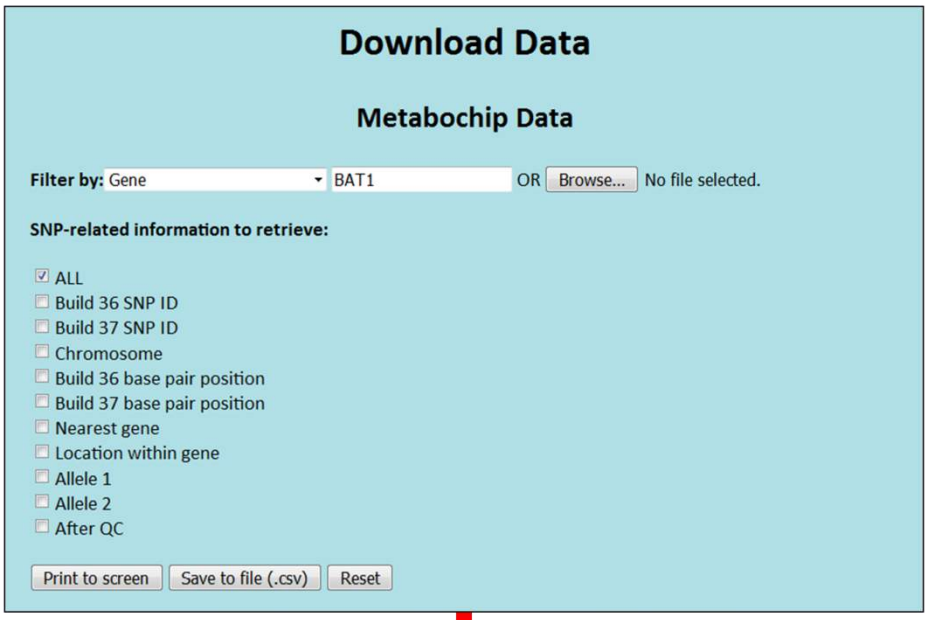


Figure 3.14a Example one of the input and output of a Metabochip data download query.



12 result(s) matched the selected criteria

SNPID36	SNPID37	Chromosome	BasePairPosition36	BasePairPosition37	NearestGene	LocationWithinGene	Allele1	Allele2	AfterQC
rs12665501	rs12665501	6	31616048	31508069	BAT1	INTRON	0	T	
rs2239709	rs2239709	6	31615426	31507447	BAT1	INTRON	A	G	BOTH
rs2516393	rs2516393	6	31614723	31506744	BAT1	INTRON	T	G	BOTH
rs2516478	rs2516478	6	31606716	31498737	BAT1	INTRON	T	C	BOTH
rs2523506	rs2523506	6	31617946	31509967	BAT1	UTR	A	C	BOTH
rs2734583	rs2734583	6	31613459	31505480	BAT1	INTRON	C	T	BOTH
rs2844509	rs2844509	6	31618903	31510924	BAT1 | ATP6V1G2	INTERGENIC	C	T	BOTH
rs3130058	rs3130058	6	31613866	31505887	BAT1	INTRON	T	C	BOTH
rs3131628	rs3131628	6	31610746	31502767	BAT1	INTRON	G	A	BOTH
rs3219190	rs3219190	6	31605954	31497975	MCCD1 / BAT1	COMPLEX	G	A	
rs7738430	rs7738430	6	31616815	31508836	BAT1 | SNORD84	INTERGENIC	C	T	
rs9267487	rs9267487	6	31619329	31511350	BAT1 | ATP6V1G2	INTERGENIC	C	T	BOTH

Figure 3.14b Example two of the input and output of a Metabochip data download query.

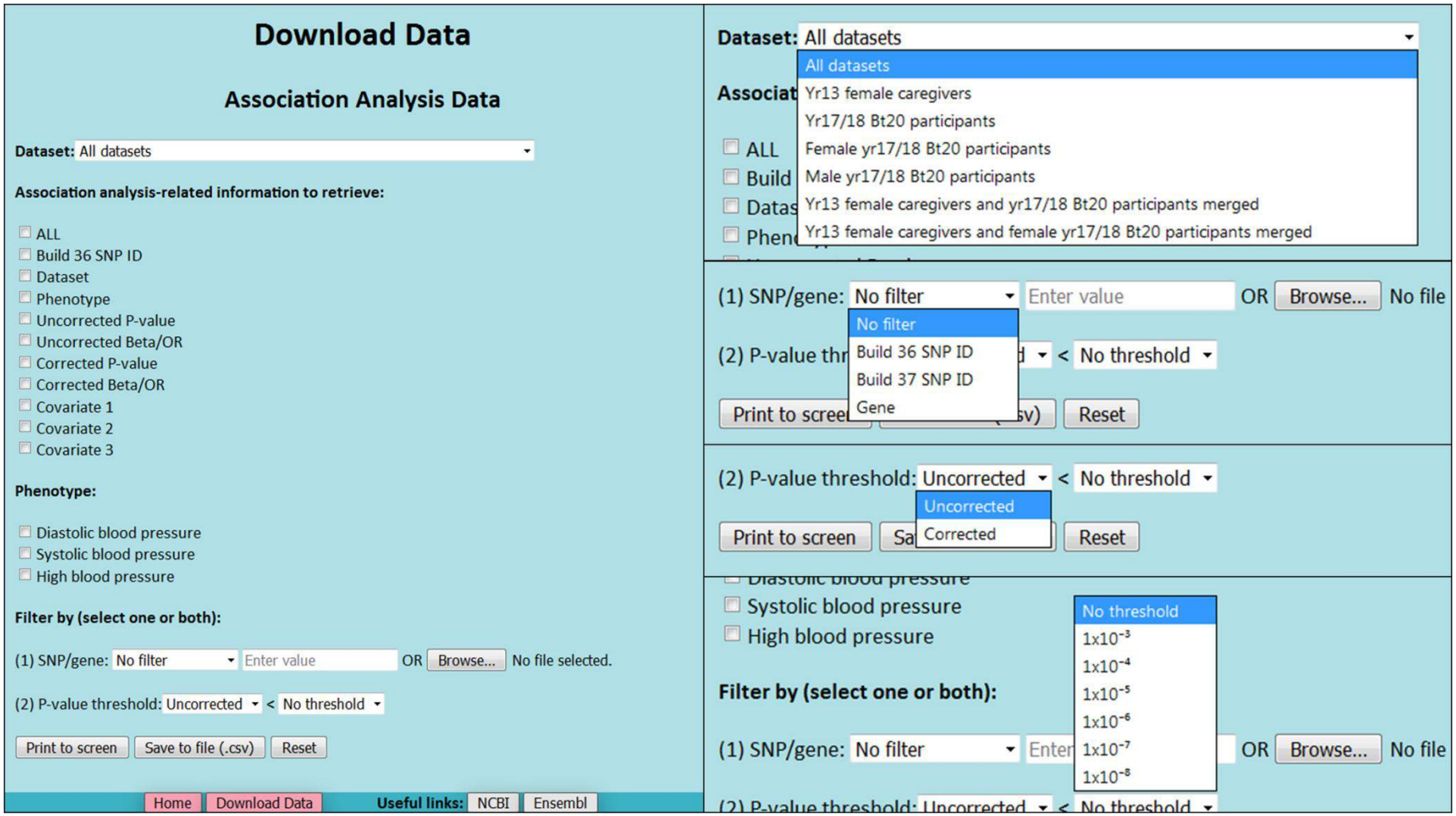
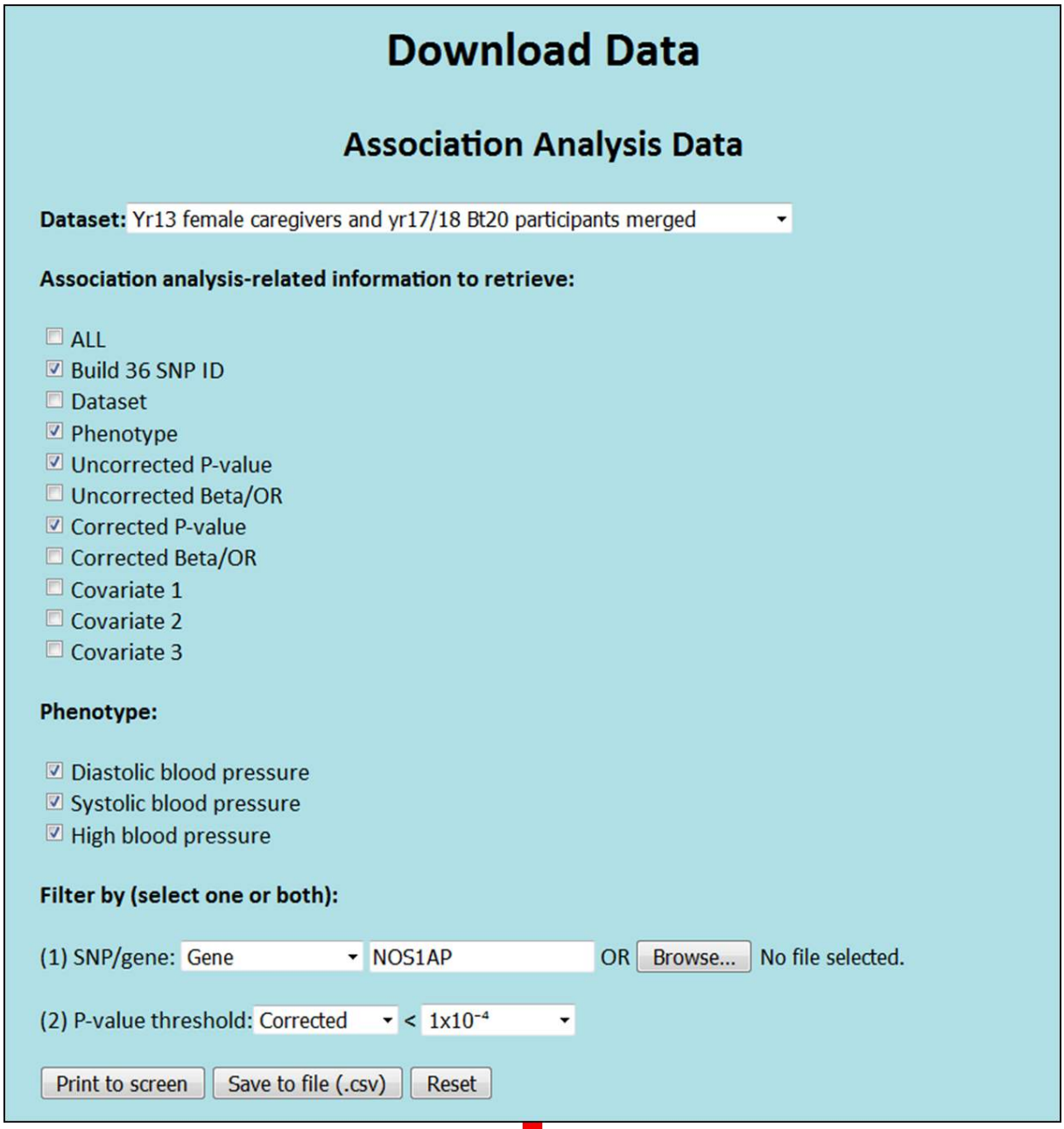


Figure 3.15 Layout of the association analysis data download page.



2 result(s) matched the selected criteria

SNPID36	Phenotype	P_valueUncorrected	P_valueCorrected
chr1:160462273	DBP	0.00004866439	0.00007179409
chr1:160522009	SBP	0.00007254314	0.00004043616

Figure 3.16 An example of the input and output of an association analysis data download query.

Genotype data

Genotype data, although not directly downloadable from the web interface, is available in PLINK format on request. For convenience, useful information is provided on the genotype data page for manipulating the genotype files and generating basic genotype data summary statistics (missingness, HWE, MAF) in PLINK as follows:

(a) Manipulating genotype files in PLINK:

To **extract/keep** genotype data for specific **individuals** (*saved as a list of Family ID/Individual ID pairs in a file called myIDs.txt*):

```
plink --bfile {INPUT GENOTYPE FILENAME} --keep myIDs.txt --make-bed --out {OUTPUT GENOTYPE FILENAME}
```

To **exclude/remove** genotype data for specific **individuals** (*saved as a list of Family ID/Individual ID pairs in a file called myIDs.txt*):

```
plink --bfile {INPUT GENOTYPE FILENAME} --remove myIDs.txt --make-bed --out {OUTPUT GENOTYPE FILENAME}
```

To **extract/keep** genotype data for a specific **SNP**, multiple specific SNPs (*saved as a list of SNP IDs in a file called mySNPs.txt*) OR a specific chromosome:

```
plink --bfile {INPUT GENOTYPE FILENAME} --snp {SNP ID} --make-bed --out {OUTPUT GENOTYPE FILENAME}
```

```
plink --bfile {INPUT GENOTYPE FILENAME} --extract mySNPs.txt --make-bed --out {OUTPUT GENOTYPE FILENAME}
```

```
plink --bfile {INPUT GENOTYPE FILENAME} --chr {CHROMSOME NUMBER} --make-bed --out {OUTPUT GENOTYPE FILENAME}
```

To **exclude/remove** genotype data for specific **SNPs** (*saved as a list of SNP IDs in a file called mySNPs.txt*):

```
plink --bfile {INPUT GENOTYPE FILENAME} --exclude mySNPs.txt --make-bed --out {OUTPUT GENOTYPE FILENAME}
```

(b) Generating basic summary statistics on the genotype data

Missingness:

```
plink --bfile {INPUT GENOTYPE FILENAME} --missing --out {OUTPUT FILENAME}
```

Hardy-Weinberg Equilibrium:

```
plink --bfile {INPUT GENOTYPE FILENAME} --hardy --out {OUTPUT FILENAME}
```

Minor allele frequencies:

```
plink --bfile {INPUT GENOTYPE FILENAME} --freq --out {OUTPUT FILENAME}
```

Additional features

Each page has a footer with links to useful websites (NCBI (*https://www.ncbi.nlm.nih.gov/)* and Ensembl (*www.ensembl.org)*) and the home and main landing pages. Pages that allow for the selection of various criteria/options have a reset button to easily clear all previously selected fields before selecting new options. Some queries allow for users to upload a file containing a list of Individuals, SNPs etc. The uploaded file should be a text file containing a list of Individual IDs/SNPIDs etc. each on a separate line.

A user manual/README has been set up to allow users to easily navigate themselves around the web interface (**Appendix F**). It contains information about the data stored in the database, including the unit of measurement of each variable.

3.3 Expandability of the database

The database has been designed in such a way that all fixed/unchangeable data (IDs, gender, ethnicity, relationship to child, caregiver's ID) for a particular dataset are in separate tables to the phenotype data that changes (age, height, weight, etc.) at each data collection time point. Therefore, data from future time points can easily be added to the database, with a simple edit made to the user interface to include the additional time points. For example, if new data for the Bt20 participants from year X becomes available, a new phenotype table will be created using the 'CREATE TABLE phenotype_yrN_bt20_participants' SQL code in Appendix D. Each phenotype column name will be appended with 'yrX' to allow for easy queries. Data will then be added into the table using the Python code shown in Appendix E. A "Year X" data collection time point will be added as an option to the drop down menu of the complex count and average/minimum/maximum pages and the year selection checkboxes and criteria of the phenotype data download pages of the web interface. Appropriate changes to the actual queries will also be made.

The Metabochip and Bt20 phenotype data also has the potential to be used further in genetic association analyses. As association analysis data involving other traits becomes available, this can also be added to the database with another simple edit made to the user interface to allow selection of data for these additional traits.

3.4 Use by other research groups

The MetaboBTT database and user interface provide a useful model for setting up databases for the storage and querying of similar data from other genetic association studies. The MetaboBTT code has been generalised and some notes added to it to guide users to modify it to fit to their own phenotype, SNP annotation and association analysis data (**See General_database_model.pdf available at https://github.com/ LieslH/Liesl-Hendry-PhD-Code**).

3.5 Discussion

Here presented is the first version of MetaboBTT, a research group-specific relational database for the storage and querying of data relating to a specific project in the field of genetics investigating risk factors for cardiometabolic-related diseases. Users are able to query the database for summary statistics relating to the phenotype data and download phenotype, SNP annotation and association analysis data matching certain user-supplied criteria. Also presented is the general code for the implementation of this database model in other research groups.

The growth in and increasing complexity of data across many biological fields poses an ongoing challenge to store, manage and analyse the data efficiently. At the most basic level is data from projects within individual research groups. As is the case in many research groups, including our own, data are often stored in spreadsheets due to their simplicity and ease of use. Unless individuals know some basic command-line scripting or a programming language, the files require manual manipulation to sort through the data and obtain useful information from it. This can often be slow and error prone. In addition, datasets are reaching sizes that are too big to handle within spreadsheets, with manipulation of the data again being slow or in some cases not possible.

Although sometimes challenging to design and develop, databases, and in particular relational databases, are a lot more flexible and have several advantages over storing data in simple spreadsheets. With a database, all the data is stored in one central location making access to the data quick and easy. There is also increased consistency and reduced updating errors as there aren't multiple copies of the data which could be edited or updated. If changes or updates to the data are required, this only needs to be done once. In a relational database, data are split across multiple tables which are linked by a primary key or column of unique identifiers. This makes individual records easy to locate. In addition, complex queries can be made that extract data from more than one table at a time using a simple join function. Queries like this would usually be more complicated to perform manually. Relational databases also have increased security. As the data exists in multiple tables, users can have restricted access to any tables that contain sensitive data. This may be particularly useful if the Bt20 data were to become publically available in the future. The current database could also become a publically available tool for anyone to access the phenotype and SNP annotation data, but the association analysis results could be hidden. Lastly, when manipulating data in spreadsheets, many intermediate spreadsheets are often generated which creates disorder and confusion as to which version of the data is correct. Databases eliminate this problem and help to ensure that there is minimal corruption of the data and that the integrity of the data is maintained.

A study showed that two thirds of users accessing online biological databases have limited programming experience (Schultheiss, 2011) and therefore require a web interface to aid data access and retrieval (Helmy et al., 2016). The interface should ideally be easy to navigate by non-experts and should require minimal user actions to retrieve the desired information quickly (Helmy et al., 2016). A simple, user-friendly web interface has therefore been designed to accompany this database to allow users in a biological setting with little or no

experience in MySQL and various programming languages to easily query and access the data contained within the database. This is possible with the click of a few buttons. The ease of use is further ensured by the inclusion of messages that alert users when an inappropriate selection or input is made or if users forget to select required options. The interface has also been tested by a few potential users or those with some knowledge of web design to ensure that it works and is easy to use.

MetaboBTT, with its current functionality, is a useful tool to provide to all members of the research group involved in investigations into cardiometabolic diseases and their genetic or non-genetic risk factors. The calculation of summary statistics allows users to get a feel for the data contained within the database. The basic count option is particularly useful for first time users of the data to obtain information about the number of individuals and the composition of males and females within each dataset. The complex count and average/minimum/maximum options provide further information about the phenotypic and physiological characteristics of the individuals making up the datasets. The data download functionality allows users to download only the data that match very specific criteria and that is relevant to them for a particular investigation. In the case of the investigation into the genetics of BP/hypertension, the SBP and DBP data along with appropriate covariate data (age, sex and BMI) is easily downloaded for the Bt20 participants and their female caregivers for year 17/18 and year 13, respectively, and used in conjunction with the available genotype data. In addition, association analysis results arising from the investigation were added to the database.

An important feature of MetaboBTT is its ability to store data of a longitudinal nature (from past and future data collection time points). Users can obtain useful summary statistics and download data from multiple years, which is useful for conducting longitudinal studies and investigating changes over time.

The performance of the database is encouraging. The interface performs well in Firefox, Chrome and Internet Explorer web browsers. The query return time for most queries is within seconds, with some more complex queries (mostly those involving the association analysis data) taking slightly longer. The MySQL tables and data only take up about 355 megabytes, which, by database standards, is relatively small. Overall, the current database is therefore inexpensive on time and space. MySQL also has support for larger databases. The database currently runs on an InnoDB engine. For a default page size of 16 kilobytes (i.e. 16 kilobytes of data are transferred between the disk and memory at any one time), the maximum size for a table is 64 terabytes. With this table size limit and a row size limit of about 65.5 kilobytes, over 1 billion rows can exist in a table. The database is, therefore, highly scalable for larger studies to include many more participants and SNPs, with possible limitations being the efficiency and speed of the queries and the available storage space and memory.

The database also has room for growth, not only through the addition of data from future data collection time points, but also through the addition of further association analysis results obtained using the Bt20 and Metabochip data. Having a comprehensive record of SNPs or genes that are associated with a particular phenotype(s) and having easy access to this information is useful to inform future replication or functional studies within the research group. In addition, the database is such that it can apply to any set of phenotype, SNP annotation and association analysis data. Although some table creation in MySQL, python data loading and manipulation of PHP/HTML/CSS scripts would be required, a general model with editing instructions has been set up to make it easy for researchers to adjust for their own data. The tools required to implement the database and interface are also freely available, making this a cost-free option for data storage and management.

As databases are used more readily and data becomes more accessible and well organised, researchers will be able to do a lot more with the data they have and will ultimately be able to accelerate biological knowledge discovery. This in turn will lead to great advancements in the scientific and medical fields in the coming years.

Chapter 4: INSIGHTS INTO THE GENETICS OF BLOOD PRESSURE IN BLACK SOUTH AFRICANS

As described previously [*See Chapter 1*], CVDs are the leading causes of non-communicable NCD deaths globally, with a major risk factor of CVDs being hypertension or raised BP. This makes studying their aetiology very important. Hypertension and BP are multifactorial in nature, with many genetic factors contributing to the disease phenotype or trait in addition to various non-genetic risk factors. Numerous studies have already reported on genes and variants that have associations with hypertension, SBP and DBP. To date, however, most of the large scale studies that have been published were carried out in individuals of non-African ancestry.

This chapter focuses on findings from the association analysis carried out in our black South African sample with the aim of providing some insight into the genetics of blood pressure in individuals of African ancestry. The work is in the form of a manuscript submitted to the Journal of Hypertension. Data were analysed mainly as a merged dataset (all participants and caregivers together). This was considered the best option to make use of the largest possible sample. Some methodology details not provided in detail in the manuscript are also presented.

The individual contributions of each of the authors to the manuscript are as follows:

- Liesl Hendry: QC, analysis and interpretation of the data, writing of the manuscript.

- Venesa Sahibdeen: Acquisition of genetic data, QC of the data, critically assessed the manuscript for intellectual content and approved final version to be published.
- Ananyo Choudhury: Assistance with generation of Evoker plots, advice/help on the QC and power calculation, critically assessed the manuscript for intellectual content and approved final version to be published.
- Shane Norris: Involvement in the larger Bt20 cohort study, critically assessed the manuscript for intellectual content and approved final version to be published.
- Michèle Ramsay: Involvement in the larger AWI-Gen study, critically assessed the manuscript for intellectual content and approved final version to be published.
- Zané Lombard: Conception and design of the study, critically assessed the manuscript for intellectual content and approved final version to be published.

4.1 Post-analysis quality control

A post-analysis QC process was carried out following initial association analysis to check the association signals and to confirm whether or not they were real. Q-Q plots were drawn and genomic inflation factors were calculated in R (v3.0.3) (R Development Core Team, 2009) (**Appendix G**). If there was any deviation of points from the normal line in the Q-Q plots or if the genomic inflation factors were significantly greater than one, association analysis was re-done with correction for principal components.

This was followed by a prioritisation step. Any SNPs with $p \leq 1x10^{-4}$ were examined further and compared to previously reported top hits (known hits) for DBP, SBP and hypertension. SNPs of interest were run through an annotation

pipeline provided by Dr Daniel Suveges from the Wellcome Trust Sanger Institute. SNPs were disregarded due to poor separation or clustering examined in the cluster plots generated in Evoker (Morris et al., 2010).

4.2 Manuscript

(as it appeared in the Journal of Hypertension submission)

Insights into the genetics of blood pressure in black South African individuals: the Birth to Twenty cohort

Blood pressure genetics in black Africans

Liesl M HENDRY[a,b], Venesa SAHIBDEEN[b,c], Ananyo CHOUDHURY[b], Shane A NORRIS[d], Michele RAMSAY[b,c] & Zané LOMBARD[a,b,c] of the AWI-Gen study and as members of the H3Africa Consortium

[a] School of Molecular & Cell Biology, Faculty of Science, University of the Witwatersrand, Johannesburg, South Africa

[b] Sydney Brenner Institute for Molecular Bioscience, University of the Witwatersrand, Johannesburg, South Africa

[c] Division of Human Genetics, School of Pathology, Faculty of Health Sciences, National Health Laboratory Service & University of the Witwatersrand, Johannesburg, South Africa

[d] MRC/Wits Developmental Pathways for Health Research Unit, Department of Paediatrics, School of Clinical Medicine, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa

Previous oral/poster presentations of part of the work

- 16th Biennial Congress of the Southern African Society for Human Genetics (SASHG) (August 2015, Centurion South Africa)
- 7th H3Africa Consortium Meeting (October 2015, Washington DC and Bethesda Maryland, USA)
- Molecular Biosciences Research Thrust Research Day 2015 (December 2015, Johannesburg, South Africa)
- 7th Cross-Faculty Symposium (March 2016, Johannesburg, South Africa)

Sources of support for the authors

This study was made possible by funding from the AWI-Gen Collaborative Centre, which is funded by the National Institutes of Health (1U54HG006938) as part of the H3Africa Consortium, and the Thuthuka Programme of the South African National Research Foundation (NRF) of South Africa for the grant, Unique Grant No. 94007 & 80702. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the NRF. LH is supported by an NRF Innovation Doctoral Scholarship for her PhD studies. Birth to Twenty is supported by funding from the University of the Witwatersrand, South African Medical Research Council and the DST-NRF Centre of Excellence in Human Development.

Conflicts of interest

The authors do not have any conflicts of interest to declare.

Correspondence

Liesl Hendry; Sydney Brenner Institute for Molecular Bioscience, The Mount, 9 Jubilee Rd, Johannesburg, 2193, South Africa; Tel: 0117176632; E-mail: lieslmaryhendry@gmail.com

Word count: 6496

Number of tables: 2

Number of figures: 1

Number of supplementary digital content files: 9 (4 tables and 5 figures)

Abstract

Objectives: Cardiovascular diseases (CVDs) are the leading cause of non-communicable disease deaths globally, with hypertension being a major risk factor contributing to CVDs. Hypertension itself has numerous risk factors and is known to be approximately 30-50% heritable, but very few studies have been performed to investigate the role of genetics in African populations. This study aimed to identify single nucleotide polymorphisms and genes associated with systolic (SBP) and diastolic (DBP) blood pressure in a black South African population.

Methods: Analysis was carried out in a merged sample of adult female caregivers (median age 41.0) and mixed sex participants (median age 17.9) (n=1947) from the Birth to Twenty cohort genotyped using the Metabochip. The dataset was further stratified to identify possible sex-and age-related associations.

Results: Association analysis identified regions of interest in the *NOS1AP* (DBP: rs112468105 – $p=7.18x10^{-5}$ and SBP: rs4657181 – $p=4.04x10^{-5}$), *MYRF* (SBP: rs11230796 – $p=2.16x10^{-7}$, rs400075 – $p=2.88x10^{-7}$) and *POC1B* (SBP: rs770373 – $p=7.05x10^{-5}$, rs770374 – $p=9.05x10^{-5}$) genes as well as some intergenic regions (*DACH1*|*LOC440145* (DBP: rs17240498 – $p=4.91x10^{-6}$ and SBP: rs17240498 – $p=2.10x10^{-5}$) and *INTS10*|*LPL* (SBP: rs55830938 – $p=1.30x10^{-5}$, rs73599609 – $p=5.78x10^{-5}$, rs73667448 – $p=6.86x10^{-5}$)). All of these were novel findings.

Conclusions: The study provided insight into the genetics of blood pressure in black South Africans, with several novel blood pressure variants identified. Further functional and replication studies in larger samples are required to confirm the role of the genes identified in blood pressure regulation and whether or not the genetic links are African-specific.

Keywords

Genetics, blood pressure, black South Africans, Metabochip, Birth to Twenty

Introduction

Research on communicable or infectious diseases have been the focus until now, particularly in Africa where they are a main cause of morbidity and mortality. Non-communicable diseases (NCDs) are, however, gaining increasing interest and becoming as significant due to their increasing burden to the continent. In the latest global status report on NCDs, it was reported that in 2012 approximately 38 million deaths (63% of total deaths) were due to NCDs [1] with this figure expected to increase to 52 million by 2030 [2]. Low-and middle-income countries (LMIC) are the most affected, with about 28 million of the NCD deaths occurring in these countries and the NCD death rate being 625 and 673 per 100 000 in low-income and lower-middle-income countries, respectively,

compared to 397 per 100 000 in high-income countries [3]. In addition, 82% of premature deaths (deaths before the age of 70) occur in LMIC.

Cardiovascular diseases (CVDs) are the leading cause of NCD deaths globally ahead of cancers, respiratory diseases and diabetes [3]. CVDs were responsible for approximately 17.5 million (46%) NCD deaths in 2012, with more than 80% occurring in LMIC, and this figure is expected to increase to about 22.2 million in 2030, with approximately 85% occurring in LMIC [2,3].

A major risk factor contributing to CVDs is hypertension or raised blood pressure (BP). In 2014 the global prevalence of raised BP was approximately 22% in adults aged 18 years and older, with the highest prevalence reported in Africa at 30% for all adults combined [3]. A recent review by Rayner and Spence, looking specifically at South Africans, highlighted the differences in patho-physiology of hypertension between black and white individuals and stated that black individuals generally warrant a different approach to causation, outcome and treatment of hypertension [4].

Risk factors associated with BP and hypertension include obesity or increased body mass index (BMI), increased salt intake, age, sex, insulin resistance, physical inactivity, alcohol intake, psychosocial stress and consumption of high-fat foods [5–11]. BP and hypertension are multifactorial traits, with a significant genetic contribution (approximately 30-50% heritable [12]) in addition to the various non-genetic risk factors [13]. The genetic contribution itself is polygenic, with small contributions from risk alleles in multiple genes playing a role in the aetiology of the trait or disorder [14]. Numerous studies have already reported on genes and variants that have associations with hypertension, systolic blood pressure (SBP) and diastolic blood pressure (DBP). To date, however, most of the large scale studies have been in individuals of European Ancestry [15], with fewer conducted in individuals of Asian or African ancestry [16]. African-related

studies have also largely been carried out in African-Americans with studies conducted in native Africans being limited.

The current study aims to contribute to what is known about the genetics of blood pressure, looking specifically at African individuals. The specific aim is to identify single nucleotide polymorphisms (SNPs) and genes associated with SBP and DBP in a black South African population. The study uses participants and their female caregivers from the Birth to Twenty (Bt20) cohort [17] using the Metabochip [18] as a genotyping tool.

Methods

Study Participants

The study participants were taken from the Bt20 cohort, which is the most extensive longitudinal study on child and adolescent health and development in Africa. The cohort initially enrolled 3273 individuals born as single births during a 7-week period in early 1990 to women living in Soweto [17]. The cohort consists of African (78%), White (6%), Coloured (12%) and Indian (4%) individuals, which is an approximate representation of the race groups in the South African population [17], although only African individuals were used in this particular study. Data, including body composition and cardiometabolic data, have been collected at regular time points since inception of the cohort, with some individuals dropping out of the study along the way due to migration and loss of follow up [17].

This study made use of DNA samples from a mixed sex subset of Bt20 participants (n=1240), with phenotype data collected in year 17/18 of the study, and their female caregivers (n=1033), with phenotype data collected at year 13 of the study. Of the 2273 samples in total, there were 975 caregiver-participant

pairs and of those with a known relationship between caregiver and participant, 864 were mother-participant pairs.

Informed consent was obtained for the collection of data and DNA samples. Ethical clearance was obtained from the University of the Witwatersrand Human Research Ethics Committee (Medical) for collection of DNA samples and phenotype data from this cohort (M010556). Further clearance was obtained for use of these samples to identify genetic risks associated with obesity (M120647) and blood pressure (M1411116) in a black South African population. DNA is currently stored in the Division of Human Genetics at the National Health Laboratory Service (NHLS), Braamfontein, South Africa.

Phenotype measurements

Blood pressure (BP) readings were taken using an Omron 6 automated machine (Kyoto, Japan). Measurements were taken with participants in a seated position. After five minutes of sitting in a resting position, three measurements were taken at intervals of two minutes. The first reading was discarded, in case of possible "white coat syndrome", and an average of the second and third measurements was calculated and used in all analyses [19].

Weight was measured to the nearest 0.1 kg using a digital scale with participants wearing light clothes and no shoes. Standing height was measured to the nearest 0.1 cm using a wall-mounted stadiometer (Holtain, UK). Body mass index (BMI) was calculated as weight (kg) divided by height squared (m^2) [19].

Genotyping

DNA, extracted using the salting out method [20], was normalized to 50ng.ul^{-1} prior to genotyping. The DNA samples were genotyped in two separate batches

(participants and caregivers) at the UC Davis Genome Centre (California, USA) for almost 200,000 SNPs known to influence cardiometabolic traits using the Metabochip (Illumina, San Diego, CA, USA). A few duplicate samples from each batch were sent with the unique samples to assess genotyping consistency. Genotypes were called using GenomeStudio Software for Illumina (v2011.1) and a custom DNAtech cluster file and final output was provided as final reports in the forward strand orientation.

Data Quality Control

Pre-analysis quality control (QC) of the data was carried out separately for the two datasets using PLINK (v1.9) [21,22]. Additional tools used included SMARTPCA (to run the principal component analysis (PCA) for identification of population outliers) [23] and Genesis (to visualize the PCA) [24]. The final report files were converted into binary PLINK format files. An initial SNP and sample removal step involved removing SNPs with complete missing genotype data and poorly genotyped samples (more than 20% missing genotype data). Further SNP QC involved removal of SNPs with high missingness rate (> 2%), low minor allele frequency (MAF) (< 0.01) and those failing Hardy-Weinberg equilibrium (HWE) ($p<1x10^{-5}$). Additional sample QC involved removal of samples with high missingness rate (> 2% for caregivers and > 3% for participants), those with discordant sex, related samples (PI_HAT > 0.1875), duplicates, samples with extreme heterozygosity (heterozygosity rate ± 3 standard deviations from the mean [25]) and population outliers. A few SNPs were also disregarded due to poor clustering examined in the cluster plots generated in Evoker [26].

Checks were also carried out on the phenotype data and corrections were made where inconsistencies were found between the original questionnaires and captured data.

Association analysis

Data were initially analysed as a merged dataset (all participants and caregivers together). Possible sex-related associations were further examined by stratifying the sample into female (female caregivers and female Bt20 participants merged) and male (male Bt20 participants) individuals, while age-related associations were examined by stratifying the sample into old (female caregivers) and young (Bt20 participants) individuals.

Association analysis of the merged datasets was performed in GEMMA (v0.94.1) [27] using univariate linear mixed models and incorporating a centered relatedness matrix to account for the relatedness between individuals from the different datasets. Analysis of the individual datasets was done by linear regression under an additive model using PLINK (v1.9). All analyses included adjustments for age, sex (where appropriate) and BMI, as well as principal components if deemed necessary after examination of quantile-quauntile (Q-Q) plots, constructed in R (v3.0.3) [28].

The standard Bonferroni significance threshold for correction for multiple testing used in genome-wide association studies ($p<5x10^{-8}$) was considered too strict in this study as this is a replication study and the Metabochip is a fine-mapping array with variants in high linkage disequilibrium (LD). The threshold to measure "array-wide" significance was thus calculated as 0.05/number of unlinked markers: $p<6.7x10^{-7}$ (0.05/74475), $p<6.1x10^{-7}$ (0.05/82239) and $p<6.6x10^{-7}$ (0.05/75834) for the merged, caregiver and Bt20 participant datasets, respectively. To address the possible introduction of Type II errors through the

application of this rigorous correction, we chose to also present results where a cut-off of $p \leq 1 \times 10^{-4}$ was met.

Association analysis results were visualized using Manhattan plots (for genome-wide visualisation) drawn in R(v3.0.3) [28] and LocusZoom plots (for regional visualisation) [29].

Power calculation

The power to detect associations was assessed using Quanto (v 1.2.4) [30]. Power was assessed for a range of beta and allele frequency (AF) values for reported associations obtained from the genome-wide association study (GWAS) catalogue [31] for both DBP and SBP. We were more than 80% powered to detect associations with an effect size of approximately 1.02 and higher for DBP and 1.53 and higher for SBP. In both cases the smaller effect sizes were only detectable at higher MAFs, with the power to detect associations at low MAFs increasing as the effect size increased.

Results

Pre-analysis QC of the data resulted in 976 participants (median age = 17.9) and 971 female caregivers (median age = 41.0), with 127,764 and 140,649 SNPs, respectively remaining. Upon merging of the datasets, 1947 individuals and 125,906 SNPs remained for analysis. Descriptive statistics of the individuals remaining after QC in the merged dataset are shown in **Table 1**.

Table 1 Descriptive statistics of the individuals remaining after QC in the merged dataset.

		Mean (SD)	Median	Range
Merged (n=1947)				
Females: 73.3%	Age (years)	30.1 (13.5)	30.0	17.3-84.0
Males: 26.6%	Weight (kg)	68.0 (16.7)	64.4	36.1-136.6
	Height (m)	1.6 (0.1)	1.6	1.2-1.9
	BMI ($kg.m^{-2}$)	26.2 (7.1)	24.6	14.9-58.8
	SBP (mmHg)	117.8 (16.8)	116.0	77.0-206.5
	DBP (mmHg)	74.1 (11.1)	72.5	44.5-129.5

Analysis of the merged dataset across the genome can be visualised in the Manhattan plots in **Figure 1**. All SNPs associated with DBP or SBP at $p \leq 1 \times 10^{-4}$ in the merged dataset are shown (**see Table, Supplemental Digital Content 1, associations with DBP and SBP at $p \leq 1 \times 10^{-4}$**). Regions that contained two or more SNPs associated with one of the two traits under investigation or that were associated with both traits were examined further (**Table 2**). SNPs in introns in *NOS1AP* and intergenic to *DACH1*|*LOC440145* associated with both DBP and SBP. SNPs intronic to *MYRF* and *POC1B* and intergenic to *INTS10*|*LPL* associated with SBP only. The only SNPs to reach "array-wide" significance are two intronic SNPs in *MYRF* (rs11230796-G and rs400075-T) which are associated with SBP. Regional plots centered around the lead SNPs of *NOS1AP* (DBP and SBP), *MYRF* (SBP), *POC1B* (SBP) and the intergenic region of *INTS10*|*LPL* are shown (**see Figures, Supplemental Digital Content 2-6, LocusZoom plots**).

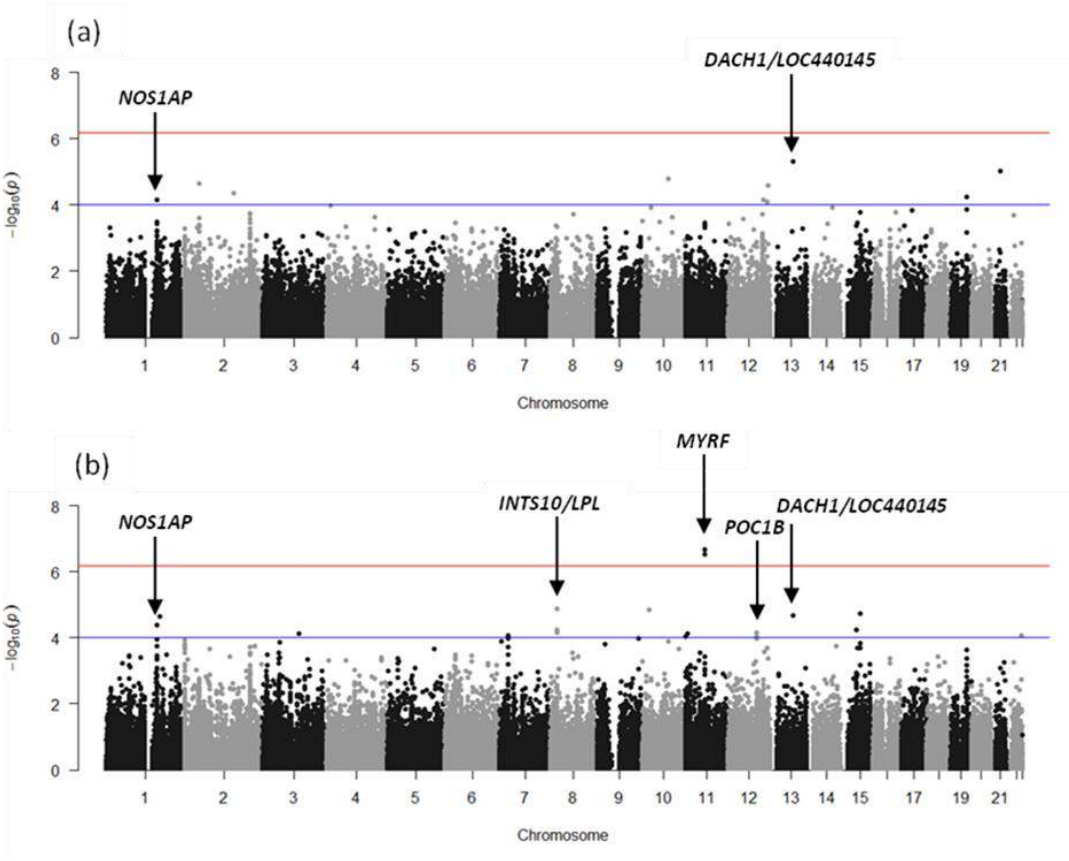


Figure 1. Manhattan plots drawn from the association results (with correction for covariates and PCs where necessary) of the all merged dataset. Plots are shown for association with (a) DBP and (b) SBP. The upper horizontal line indicates the calculated "array-wide" significance cutoff ($p=6.7x10^{-7}$) while the lower horizontal line shows the cutoff of $p=1x10^{-4}$. Identified regions of interest for further investigation are indicated by arrows. *DACH1* – Dachshund Family Transcription Factor 1; *INTS10* – integrator complex subunit 10; *LPL* – lipoprotein lipase; *MYRF* – myelin regulatory factor; *NOS1AP* – nitric oxide synthase 1 (neuronal) adaptor protein; *POC1B* – POC1 centriolar protein.

Table 2 Identified regions of interest associated with DBP or SBP in the merged dataset.

CHROMOSOME	GENE/REGION	SNP ID [a]	BASE PAIR POSITION [a]	ALLELES		A1 FREQUENCY [c]			DBP		SBP	
				A1 [b]	A2	Bt20	YRI	CEU	P-value [d]	Beta [e]	P-value [d]	Beta [e]
1	*NOS1AP*	rs112468105	162195649	G	C	0.011	0.023	0.000	$7.18x10^{-5}$	6.69	-	-
		rs4657181	162255385	T	A	0.046	0.023	0.556	-	-	$4.04x10^{-5}$	-5.62
13	*DACH1 | LOC440145*	rs17240498	72965307	C	T	0.011	0.000	0.182	$4.91x10^{-6}$	7.93	$2.10x10^{-5}$	11.68
11	*MYRF*	rs11230796	61529267	G	T	0.058	0.056	0.222	-	-	**$2.16x10^{-7}$**	6.12
		rs400075	61528814	T	C	0.058	0.056	0.217	-	-	**$2.88x10^{-7}$**	6.02
12	*POC1B*	rs770373	89818289	T	C	0.176	0.329	0.449	-	-	$7.05x10^{-5}$	-2.95
		rs770374	89818022	T	G	0.231	0.366	0.561	-	-	$9.05x10^{-5}$	-2.62
8	*INTS10 | LPL*	rs55830938	19735188	G	T	0.026	0.028	0.000	-	-	$1.30x10^{-5}$	7.49
		rs73599609	19756974	C	G	0.050	0.051	0.000	-	-	$5.78x10^{-5}$	5.14
		rs73667448	19747475	C	A	0.028	0.042	0.000	-	-	$6.86x10^{-5}$	6.72

[a] All SNP IDs and base pair positions are reported using *GRCh37* (Genome Reference Consortium human genome Build *37).*

[b] A1 corresponds to the minor allele in the dataset.

[c] Frequencies of allele 1 are recorded for the merged dataset used in this study (Bt20) and for an African and European 1000 Genomes population - the Yoruba in Ibadan, Nigeria (YRI) and the Utah Residents (CEPH) with Northern and Western Ancestry (CEU).

[d] p-value adjusted for age, sex, BMI and principal components. P-values that pass the "array-wide" significance threshold ($p<6.7x10^{-7}$) are shown in bold.

[e] Beta values are with respect to the minor allele in the sample. A positive beta indicates that the minor allele is associated with an increased blood pressure relative to the major allele, and *vice versa*

DACH1 – Dachshund Family Transcription Factor 1; *INTS10* – integrator complex subunit 10; *LPL* – lipoprotein lipase; *MYRF* – myelin regulatory factor; *NOS1AP* – nitric oxide synthase 1 (neuronal) adaptor protein; *POC1B* – POC1 centriolar protein

Stratification of the sample into males and females and older and younger individuals revealed possible sex- and age-related associations (**see Table, Supplemental Digital Content 7, sex- and age-related associations**). Associations specific to females and older individuals included SNPs in *NOS1AP* (DBP and SBP). Associations of SNPs in *MYRF* (SBP) and intergenic to *DACH1* and *LOC440145* (DBP and SBP) were specific to females only. SNPs in *POC1B* or intergenic to *DUSP6 and POC1B* were associated with SBP in both older and younger individuals, but the SNPs were different in each case.

Discussion

This study revealed several associations with DBP and SBP in black South Africans. The analysis pointed to regions of interest in the *NOS1AP* (DBP and SBP), *MYRF* (SBP) and *POC1B* (SBP) genes as well as two intergenic regions (*DACH1*|*LOC440145* and *INTS10*|*LPL*). Two SNPs in the *MYRF* gene met the calculated "array-wide" significance threshold ($p<6.7x10^{-7}$ for the merged dataset) for multiple testing. All SNPs identified were novel associations with BP.

Of the non-intergenic regions identified, only *POC1B (*POC1 centriolar protein B) has previously shown any kind of link to blood pressure. In our study, two intronic SNP alleles (rs770373-T and rs770374-T) associated with decreased SBP. rs770373 and rs770374 are shown to be in high LD in both the YRI and CEU populations and could also be in high LD in the South African population used in this study. In a previous study testing the SNP main effects and SNP-age interactions, a SNP (rs4842666) intergenic to *POC1B* (encoding a POC1 protein with a role in basal body and cilia formation) and *ATP2B1* (ATPase plasma membrane Ca^{2+} transporting 1) was found to be associated with SBP ($7.91x10^{-13}$), DBP ($2.13x10^{-10}$) and mean arterial pressure (MAP) ($7.24x10^{-13}$). This study was conducted among individuals of age 40 to 49 years in a meta-analysis to determine age-specific genetic effects [32]. Polymorphisms in *ATP2B1* are known

to be associated with BP/hypertension [33] so it may be that rs4842666 is in LD with other variants within *ATP2B1*, thus presenting as an association with BP. Apart from a link to BP, a missense mutation in *POC1B* is known to cause an autosomal recessive form of cone-rod dystrophy [34].

In our study, SNPs in the *NOS1AP (*nitric oxide synthase 1 (neuronal) adaptor protein) gene were found to be associated with increased DBP (rs112468105-G) and decreased SBP (rs4657181-T). Polymorphisms in this gene have not previously been associated with blood pressure, but have been associated with other cardiovascular phenotypes in various populations, most notably QT interval length [35–44]. Other associations, with significance at a genome-wide level, have been found with sudden cardiac death [45] and change in body mass index over time [46]. Polymorphisms in *NOS1AP* have also shown a suggestive association with hip circumference [47] and word reading [48]. One SNP identified in this study (rs4657181) showed a previous association with schizophrenia in a South American population [49]. Interestingly, several genes in the chromosome 1q linkage region, in which *NOS1AP* falls, have previously been reported to be associated with hypertension [13] which motivates for this gene and regions on chromosome 1 to be investigated further for their role in blood pressure/hypertension. Of all the regions of interest identified, *NOS1AP* could have the most plausible functional link to blood pressure regulation or hypertension risk. The gene encodes an adaptor protein for neuronal nitric oxide synthase (nNOS), an enzyme involved in nitric oxide (NO) synthesis in the nervous tissue in the central and peripheral nervous systems. NO has several roles in the body, one of which is as a vasodilator and regulator of vascular tone and blood flow [50]. Most research until now has implicated NO synthesised in the blood vessels by endothelial NOS (eNOS) in regulation of vascular tone and blood flow, but more recent animal and human studies have suggested the involvement of the neuronal-derived NO in this process too [50]. NO has also been implicated in BP regulation and impaired NO bioactivity has shown to be

associated with hypertension, although the mechanism is unclear [51]. Research has also shown that NO synthesised in the central nervous system by nNOS is involved in the central regulation of blood pressure and inhibition of nNOS activity in the medulla and hypothalamus has been linked to systemic hypertension [52]. Studies have debated the possible association of mutations in eNOS with hypertension, with some studies showing an association and others not [51], but no studies have identified associations between mutations in the nNOS gene and hypertension. The exact role of the NOS1AP protein in the regulation of nNOS function in human disease is not clear, but a recent study showed that over-expression of NOS1AP increased nNOS activity [53].

Despite associations of the two SNPs in *MYRF (*myelin regulatory factor) (rs11230796-G and rs400075-T) with increased SBP at "array-wide" significance ($p<6.7x10^{-7}$ for the merged dataset), there is no clear functional link to blood pressure. *MYRF* encodes a transcription factor involved in myelination in the central nervous system. Polymorphisms in this gene have not shown any previous associations with BP or hypertension, but have been associated with fatty acid, phospholipid and blood metabolite levels [54–57] and colorectal cancer risk in East Asians [58], with most associations being at a genome-wide significance level.

Of the intergenic regions of interest identified through analysis of the merged dataset, *INTS10* (Integrator complex subunit10)/*LPL* (lipoprotein lipase) is the most interesting finding. The *LPL* gene encodes a lipase expressed in the heart, muscle and adipose tissue. Linkage studies have linked SBP to a region at or near the *LPL* gene in Taiwanese individuals [59] and in a more recent study an association was suggested between the *LPL* gene and hypertension following haplotype analysis [60]. The findings are, however, inconclusive and the same has not been seen in Caucasian individuals [61]. Associations have also been reported between SNPs in *LPL* and coronary artery disease (rs264-A) [62] and

triglycerides-BP (rs15285-A) [63]. Interestingly, the SNPs identified in this study in this region are monomorphic in European, East Asian and South Asian populations.

An accurate investigation into the genetics of hypertension as a binary disease trait was not possible in this study due to the number of cases and controls being insufficient to observe any meaningful effect. A simple analysis was nevertheless carried out after classifying individuals into having high blood pressure (SBP ≥140mmHg and/or DBP ≥90mmHg) versus normal or low blood pressure and these findings (**see Table, Supplemental Digital Content 8, associations with high BP**) along with possible age- and sex-related findings are shown (**see Table, Supplemental Digital Content 9, age- and sex-related associations with high BP**). One interesting observation is the association of three intronic SNPs (rs1557647-A, rs179434-A and rs179431-G) in *KCNQ1 (*channel, voltage gated KQT-like subfamily Q, member 1) with high blood pressure. *KCNQ1* encodes a voltage-gated potassium channel which is required in the repolarisation phase of the cardiac action potential. A study conducted in African Americans identified a variant near KCNQ1 (rs4930130) with a suggestive association with DBP [64]. In another study in the Han Chinese, a suggestive association was found between rs10832417 and MAP responses to high-sodium intervention [65]. Polymorphisms in *KCNQ1* have in the past been most commonly linked to type 2 diabetes risk, in a range of populations including Europeans [66–68], African Americans [69], Asians [68,70–75], American Indians [76] and Hispanics [77–79], and QT interval [35,38–40,42,43]. Polymorphisms in this gene were also chosen to be part of a locus fine mapping region on the Metabochip based on previous association with the QT trait and type 2 diabetes. Other associated phenotypes have included height [80–82] and BMI [83] (at genome-wide significance level) and plasma amyloid beta peptide concentrations [84], platelet aggregation [85] and lactic dehydrogenase levels [86] (suggestive associations).

The three SNPs intergenic to *INTS10* and *LPL* that associated with SBP also associated with high blood pressure in the binary analysis, with one SNP (rs55830938-G) reaching "array-wide" significance. This association, however, could have been driven by the high SBP.

The Population Architecture using Genomics and Epidemiology (PAGE) study, whose main goal is to assess the generalizability of GWAS-identified variants across different populations, assessed the fine mapping capability of the Metabochip in African-Americans [87] and found it to be successful. Although it is known that African-Americans are genetically different to our African population, the Metabochip served as a good starting point in the investigation of blood pressure genetics in black South Africans, despite the tool being developed from data on European populations. The Metabochip was developed in 2009 and several new BP/hypertension associated variants and regions have been identified since then, therefore possibly limiting the capacity to replicate in our population what has previously been found. In addition, as the Metabochip only contains variants known to be previously associated with cardiometabolic traits, the chances of identifying novel associations in Africans is reduced.

This study has provided some insight into the genetics of blood pressure in black South Africans. Studies in larger samples could enable us to identify more associated variants that have modest to small effects. The functional significance of the associations identified is unclear, though some have plausible biological explanations for their role in regulating blood pressure. Functional and replication studies in larger African studies, as are proposed within the H3Africa Consortium [88] and more specifically the AWI-Gen study [89], will no doubt provide more insight into the genetics in African populations.

Acknowledgements

The authors wish to thank Richard Munthali for assistance with data quality control. This study was made possible by funding from the AWI-Gen Collaborative Centre, which is funded by the National Institutes of Health (1U54HG006938) as part of the H3Africa Consortium, and the Thuthuka Programme of the South African National Research Foundation (NRF) of South Africa for the grant, Unique Grant No. 94007 & 80702. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the NRF. LH is supported by an NRF Innovation Doctoral Scholarship for her PhD studies. The authors are grateful to the participants of the Bt20 study. Bt20 is supported by funding from the University of the Witwatersrand, South African Medical Research Council and the DST-NRF Centre of Excellence in Human Development.

References

1 World Health Organization. Global Health Estimates : Death by cause, age, sex and country, 2000 -2012. WHO. 2014.

2 Mathers CD, Loncar D. Projections of global mortality and burden of disease from 2002 to 2030. *PLoS Med* 2006; 3:2011–2030.

3 World Health Organization. Global status report on noncommunicable diseases 2014. 2014; :176.

4 Rayner BL, Spence JD. Hypertension in blacks: insights from Africa. *J Hypertens* 2017; 35:234–239.

5 Forrester T. Historic and early life origins of hypertension in Africans. *J Nutr* 2004; 134:211–216.

6 Banerjee S. Hypertension in children. *Clin Queries Nephrol* 2013; 2:78–83.

7 Boutin-Foster C, Ogedegbe G, Ravenell JE, Robbins L, Charlson ME.

Ascribing meaning to hypertension: a qualitative study among African Americans with uncontrolled hypertension. *Ethn Dis* 2007; 17:29–34.

8 Douglas JG, Bakris GL, Epstein M, Ferdinand KC, Ferrario C, Flack JM, *et al.* Management of high blood pressure in African Americans: consensus statement of the Hypertension in African Americans Working Group of the International Society on Hypertension in Blacks. *Arch Intern Med* 2003; 163:525–541.

9 Roger VL, Go AS, Lloyd-Jones DM, Benjamin EJ, Berry JD, Borden WB, *et al.* Heart disease and stroke statistics--2012 update: a report from the American Heart Association. *Circulation* 2012; 125:e2–e220.

10 Reckelhoff JF. Gender differences in the regulation of blood pressure. *Hypertension* 2001; 37:1199–1208.

11 Seedat YK. Race, environment and blood pressure: the South African experience. *J Hypertens* 1983; 1:7–12.

12 Munroe PB, Barnes MR, Caulfield MJ. Advances in blood pressure genomics. *Circ Res* 2013; 112:1365–1379.

13 Faruque MU, Chen G, Doumatey A, Huang H, Zhou J, Dunston GM, *et al.* Association of ATP1B1, RGS5 and SELE polymorphisms with hypertension and blood pressure in African-Americans. *J Hypertens* 2011; 29:1906–1912.

14 Doris PA. Hypertension genetics, single nucleotide polymorphisms, and the common disease:common variant hypothesis. *Hypertension* 2002; 39:323–331.

15 Franceschini N, Le T. Genetics of Hypertension: discoveries from the bench to human populations. *Am J Physiol - Ren Physiol* 2013; 306:F1–F11.

16 Popejoy A, Fullerton S. Genomics is failing on diversity. *Nature* 2016; 538:161–164.

17 Richter L, Norris S, Pettifor J, Yach D, Cameron N. Cohort Profile:

Mandela's children: the 1990 Birth to Twenty study in South Africa. *Int J Epidemiol* 2007; 36:504–511.

18 Voight BF, Kang HM, Ding J, Palmer CD, Sidore C, Chines PS, *et al.* The Metabochip, a Custom Genotyping Array for Genetic Studies of Metabolic, Cardiovascular, and Anthropometric Traits. *PLoS Genet* 2012; 8:e1002793.

19 Kagura J, Adair LS, Musa MG, Pettifor JM, Norris SA. Blood pressure tracking in urban black South African children: birth to twenty cohort. *BMC Pediatr* 2015; 15:1–7.

20 Miller SA, Dykes DD, Polesky HF. A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Res* 1988; 16:1215.

21 Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007; 81:559–575.

22 Purcell S, Chang C. PLINK 1.9. 2014.https://www.cog-genomics.org/plink2

23 Patterson N, Price AL, Reich D. Population structure and eigenanalysis. *PLoS Genet* 2006; 2:e190.

24 Buchmann R, Hazelhurst S. Genesis Manual . 2014.http://www.bioinf.wits.ac.za/software/genesis/Genesis.pdf

25 Anderson CA, Pettersson FH, Clarke GM, Cardon LR, Morris AP, Zondervan KT. Data quality control in genetic case-control association studies. *Nat Protoc* 2010; 5:1564–1573.

26 Morris JA, Randall JC, Maller JB, Barrett JC. Evoker: A visualization tool for genotype intensity data. *Bioinformatics* 2010; 26:1786–1787.

27 Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. *Nat Genet* 2012; 44:821–824.

28 R Development Core Team. *R: A language and environment for statistical*

computing. Vienna, Austria: ; 2009. doi:10.1007/978-3-540-74686-7

29 Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, Gliedt TP, *et al.* LocusZoom: Regional visualization of genome-wide association scan results. *Bioinformatics* 2010; 26:2336–2337.

30 Gauderman W, Morrison J. QUANTO 1.1: A computer program for power and sample size calculations for genetic-epidemiology studies. *http://hydra.usc.edu/gxe* 2006; :35–50.

31 Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* 2014; 42:D1001–D1006.

32 Simino J, Shi G, Bis JC, Chasman DI, Ehret GB, Gu X, *et al.* Gene-age interactions in blood pressure regulation: A large-scale investigation with the CHARGE, Global BPgen, and ICBP consortia. *Am J Hum Genet* 2014; 95:24–38.

33 Levy D, Ehret GB, Rice K, Verwoert GC, Launer LJ, Dehghan A, *et al.* Genome-wide association study of blood pressure and hypertension. *Nat Genet* 2009; 41:677–687.

34 Durlu YK, Köroglu C, Tolun A. Novel recessive cone-rod dystrophy caused by POC1B mutation. *JAMA Ophthalmol* 2014; 132:1185–1191.

35 Arking DE, Pulit SL, Crotti L, van der Harst P, Munroe PB, Koopmann TT, *et al.* Genetic association study of QT interval highlights role for calcium signaling pathways in myocardial repolarization. *Nat Genet* 2014; 46:826–836.

36 Kim JW, Hong KW, Go MJ, Kim SS, Tabara Y, Kita Y, *et al.* A common variant in SLC8A1 is associated with the duration of the electrocardiographic QT interval. *Am J Hum Genet* 2012; 91:180–184.

37 Marroni F, Pfeufer A, Aulchenko YS, Franklin CS, Isaacs A, Pichler I, *et al.* A genome-wide association scan of RR and QT interval duration in 3

European genetically isolated populations: The EUROSPAN project. *Circ Cardiovasc Genet* 2009; 2:322–328.

38 Newton-Cheh C, Eijgelsheim M, Rice KM, de Bakker PIW, Yin X, Estrada K, *et al.* Common variants at ten loci influence QT interval duration in the QTGEN Study. *Nat Genet* 2009; 41:399–406.

39 Pfeufer A, Sanna S, Arking DE, Müller M, Gateva V, Fuchsberger C, *et al.* Common variants at ten loci modulate the QT interval duration in the QTSCD Study. *Nat Genet* 2009; 41:407–414.

40 Smith JG, Avery CL, Evans DS, Nalls MA, Meng YA, Smith EN, *et al.* Impact of ancestry and common genetic variants on QT interval in African Americans. *Circ Cardiovasc Genet* 2012; 5:647–655.

41 Nolte IM, Wallace C, Newhouse SJ, Waggott D, Fu J, Soranzo N, *et al.* Common genetic variation near the phospholamban gene is associated with cardiac repolarisation: meta-analysis of three genome-wide association studies. *PLoS ONE* 2009; 4:e6138.

42 Sano M, Kamitsuji S, Kamatani N, Hong KW, Han BG, Kim Y, *et al.* Genome-wide association study of electrocardiographic parameters identifies a new association for PR interval and confirms previously reported associations. *Hum Mol Genet* 2014; 23:6668–6676.

43 Holm H, Gudbjartsson DF, Arnar DO, Thorleifsson G, Thorgeirsson G, Stefansdottir H, *et al.* Several common variants modulate heart rate, PR interval and QRS duration. *Nat Genet* 2010; 42:117–122.

44 Chambers JC, Zhao J, Terracciano CMN, Bezzina CR, Zhang W, Kaba R, *et al.* Genetic variation in SCN10A influences cardiac conduction. *Nat Genet* 2010; 42:149–152.

45 Kao WH, Arking DE, Post W, Rea TD, Sotoodehnia N, Prineas RJ, *et al.* Genetic variations in nitric oxide synthase 1 adaptor protein are associated with sudden cardiac death in US white community-based populations.

Circulation 2009; 119:940–951.

46 Scannell Bryan M, Argos M, Pierce B, Tong L, Rakibuz-Zaman M, Ahmed A, *et al.* Genome-wide association studies and heritability estimates of body mass index related phenotypes in Bangladeshi adults. *PLoS One* 2014; 9:e105062.

47 Comuzzie AG, Cole SA, Laston SL, Voruganti VS, Haack K, Gibbs RA, *et al.* Novel genetic loci identified for the pathophysiology of childhood obesity in the hispanic population. *PLoS One* 2012; 7:e51954.

48 Luciano M, Evans DM, Hansell NK, Medland SE, Montgomery GW, Martin NG, *et al.* A genome-wide association study for reading and language abilities in two population cohorts. *Genes, Brain Behav* 2013; 12:645–652.

49 Kremeyer B, García J, Kymäläinen H, Wratten N, Restrepo G, Palacio C, *et al.* Evidence for a role of the NOS1AP (CAPON) gene in schizophrenia and its clinical dimensions: An association study in a South American population isolate. *Hum Hered* 2009; 67:163–173.

50 Melikian N, Seddon MD, Casadei B, Chowienczyk PJ, Shah AM. Neuronal nitric oxide synthase and human vascular regulation. Trends Cardiovasc. Med. 2009; 19:256–262.

51 Hermann M, Flammer A, Lüscher TF. Nitric oxide in hypertension. *J Clin Hypertens* 2006; 8:17–29.

52 Förstermann U, Sessa WC. Nitric oxide synthases: regulation and function. Eur. Heart J. 2012; 33:829–837.

53 Lu C-J, Hao G, Nikiforova N, Larsen HE, Liu K, Crabtree MJ, *et al.* CAPON modulates neuronal calcium handling and cardiac sympathetic neurotransmission during dysautonomia in hypertension. *Hypertension* 2015; 65:1288–1297.

54 Mozaffarian D, Kabagambe EK, Johnson CO, Lemaitre RN, Manichaikul A, Sun Q, *et al.* Genetic loci associated with circulating phospholipid trans

fatty acids: A meta-analysis of genome-wide association studies from the CHARGE consortium. *Am J Clin Nutr* 2015; 101:398–406.

55 Tintle NL, Pottala J V, Lacey S, Ramachandran V, Westra J, Rogers A, *et al.* A genome-wide association study of fourteen red blood cell fatty acids in the Framingham Heart Study. *Prostaglandins, Leukot Essent Fat Acids* 2015; 94:65–72.

56 Lemaitre RN, Tanaka T, Tang W, Manichaikul A, Foy M, Kabagambe EK, *et al.* Genetic loci associated with plasma phospholipid n-3 fatty acids: a meta-analysis of genome-wide association studies from the CHARGE Consortium. *PLoS Genet* 2011; 7:e1002193.

57 Shin S-Y, Fauman EB, Petersen A-K, Krumsiek J, Santos R, Huang J, *et al.* An atlas of genetic influences on human blood metabolites. *Nat Genet* 2014; 46:543–550.

58 Zhang B, Jia W-H, Matsuda K, Kweon S-S, Matsuo K, Xiang Y-B, *et al.* Large-scale genetic study in East Asians identifies six new loci associated with colorectal cancer risk. *Nat Genet* 2014; 46:533–542.

59 Wu DA, Bu X, Warden CH, Shen DD, Jeng CY, Sheu WH, *et al.* Quantitative trait locus mapping of human blood pressure to a genetic region at or near the lipoprotein lipase gene locus on chromosome 8p22. *J Clin Invest* 1996; 97:2111–2118.

60 Li B, Ge D, Wang Y, Zhao W, Zhou X, Gu D, *et al.* Lipoprotein lipase gene polymorphisms and blood pressure levels in the Northern Chinese Han population. *Hypertens Res* 2004; 27:373–378.

61 Hunt S, Province M, Atwood L, Sholinsky P, Lalouel J, Rao D, *et al.* No linkage of the lipoprotein lipase locus to hypertension in Caucasians. *J Hypertens* 1999; 17:39–43.

62 Dichgans M, Malik R, König IR, Rosand J, Clarke R, Gretarsdottir S, *et al.* Shared genetic susceptibility to ischemic stroke and coronary artery

disease: a genome-wide analysis of common variants. *Stroke* 2014; 45:24–36.

63 Kraja AT, Vaidya D, Pankow JS, Goodarzi MO, Assimes TL, Kullo IJ, *et al.* A bivariate genome-wide approach to metabolic syndrome: STAMPEED Consortium. *Diabetes* 2011; 60:1329–1339.

64 Fox ER, Young JH, Li Y, Dreisbach AW, Keating BJ, Musani SK, *et al.* Association of genetic variation with systolic and diastolic blood pressure among African Americans: the Candidate Gene Association Resource study. *Hum Mol Genet* 2011; 20:2273–2284.

65 He J, Kelly TN, Zhao Q, Li H, Huang J, Wang L, *et al.* Genome-wide association study identifies 8 novel loci associated with blood pressure responses to interventions in Han Chinese. *Circ Cardiovasc Genet* 2013; 6:598–607.

66 Mahajan A, Go MJ, Zhang W, Below JE, Gaulton KJ, Ferreira T, *et al.* Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nat Genet* 2014; 46:234–44.

67 Voight BF, Scott LJ, Steinthorsdottir V, Morris ADP, Dina C, Welch RP, *et al.* Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nat Genet* 2010; 42:579–589.

68 Unoki H, Takahashi A, Kawaguchi T, Hara K, Horikoshi M, Andersen G, *et al.* SNPs in KCNQ1 are associated with susceptibility to type 2 diabetes in East Asian and European populations. *Nat Genet* 2008; 40:1098–1102.

69 Ng MC, Shriner D, Chen BH, Li J, Chen WM, Guo X, *et al.* Meta-analysis of genome-wide association studies in African Americans provides insights into the genetic architecture of type 2 diabetes. *PLoS Genet* 2014; 10:e1004517.

70 Hara K, Fujita H, Johnson TA, Yamauchi T, Yasuda K, Horikoshi M, *et al.*

Genome-wide association study identifies three novel loci for type 2 diabetes. *Hum Mol Genet* 2014; 23:239–246.

71 Li H, Gan W, Lu L, Dong X, Han X, Hu C, *et al.* A genome-wide association study identifies GRK5 and RASGRP1 as type 2 diabetes loci in Chinese Hans. *Diabetes* 2013; 62:291–298.

72 Cui B, Zhu X, Xu M, Guo T, Zhu D, Chen G, *et al.* A genome-wide association study confirms previously reported loci for type 2 diabetes in Han Chinese. *PLoS One* 2011; 6:e22353.

73 Tsai FJ, Yang CF, Chen CC, Chuang LM, Lu CH, Chang CT, *et al.* A genome-wide association study identifies susceptibility variants for type 2 diabetes in Han Chinese. *PLoS Genet* 2010; 6:e1000847.

74 Takeuchi F, Serizawa M, Yamamoto K, Fujisawa T, Nakashima E, Ohnaka K, *et al.* Confirmation of multiple risk loci and genetic impacts by a genome-wide association study of type 2 diabetes in the Japanese population. *Diabetes* 2009; 58:1690–1699.

75 Yasuda K, Miyake K, Horikawa Y, Hara K, Osawa H, Furuta H, *et al.* Variants in KCNQ1 are associated with susceptibility to type 2 diabetes mellitus. *Nat Genet* 2008; 40:1092–1097.

76 Hanson RL, Muller YL, Kobes S, Guo T, Bian L, Ossowski V, *et al.* A genome-wide association study in American Indians implicates DNER as a susceptibility locus for type 2 diabetes. *Diabetes* 2014; 63:369–376.

77 Williams AL, Jacobs SBR, Moreno-Macias H, Huerta-Chagoya A, Churchhouse C, Marquez-Luna C, *et al.* Sequence variants in SLC16A11 are a common risk factor for type 2 diabetes in Mexico. *Nature* 2014; 506:97–101.

78 Parra EJ, Below JE, Krithika S, Valladares A, Barta JL, Cox NJ, *et al.* Genome-wide association study of type 2 diabetes in a sample from Mexico City and a meta-analysis of a Mexican-American sample from Starr County,

Texas. *Diabetologia* 2011; 54:2038–2046.

79 Palmer ND, Goodarzi MO, Langefeld CD, Wang N, Guo X, Taylor KD, *et al.* Genetic Variants Associated With Quantitative Glucose Homeostasis Traits Translate to Type 2 Diabetes in Mexican Americans: The GUARDIAN (Genetics Underlying Diabetes in Hispanics) Consortium. *Diabetes* 2015; 64:1853–1866.

80 He M, Xu M, Zhang B, Liang J, Chen P, Lee J-Y, *et al.* Meta-analysis of genome-wide association studies of adult height in East Asians identifies 17 novel loci. *Hum Mol Genet* 2015; 24:1791–1800.

81 Lango Allen H, Estrada K, Lettre G, Berndt SI, Weedon MN, Rivadeneira F, *et al.* Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* 2010; 467:832–838.

82 Wood AR, Esko T, Yang J, Vedantam S, Pers TH, Gustafsson S, Chu AY, Estrada K, Luan J, Kutalik Z, Amin N, Buchkovich ML, Croteau-Chonka DC, Day FR, Duan Y, Fall T, Fehrmann R, Ferreira T, Jackson AU, Karjalainen J, Lo KS, Locke AE, Mägi R, Mihailov E, Por FT. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat Genet* 2014; 46:1173–1186.

83 Wen W, Zheng W, Okada Y, Takeuchi F, Tabara Y, Hwang JY, *et al.* Meta-analysis of genome-wide association studies in East Asian-ancestry populations identifies four new loci for body mass index. *Hum Mol Genet* 2014; 23:5492–5504.

84 Chouraki V, De Bruijn RFAG, Chapuis J, Bis JC, Reitz C, Schraen S, *et al.* A genome-wide association meta-analysis of plasma Aβ peptides concentrations in the elderly. *Mol Psychiatry* 2014; 19:1326–1335.

85 Johnson AD, Yanek LR, Chen M-H, Faraday N, Larson MG, Tofler G, *et al.* Genome-wide meta-analyses identifies seven loci associated with platelet aggregation in response to agonists. *Nat Genet* 2010; 42:608–13.

86 Melzer D, Perry JR, Hernandez D, Corsi A-MM, Stevens K, Rafferty I, *et al.* A genome-wide association study identifies protein quantitative trait loci (pQTLs). *PLoS Genet* 2008; 4:e1000072.

87 Buyske S, Wu Y, Carty CL, Cheng I, Assimes TL, Dumitrescu L, *et al.* Evaluation of the metabochip genotyping array in African Americans and implications for fine mapping of GWAS-identified loci: the PAGE study. *PLoS One* 2012; 7:e35651.

88 H3Africa Consortium. Enabling the genomic revolution in Africa. *Science (80-)* 2014; 344:1346–1348.

89 Ramsay M, Crowther N, Tambo E, Agongo G, Baloyi V, Dikotope S, *et al.* H3Africa AWI-Gen Collaborative Centre: a resource to study the interplay between genomic and environmental risk factors for cardiometabolic diseases in four sub-Saharan African countries. *Glob Heal Epidemiol Genomics* 2016; 1:1–13.

Supplemental Digital Content

Supplemental Digital Content 1. All SNPs associated with DBP or SBP at $p \leq 1 \times 10^{-4}$ in the merged dataset

PHENOTYPE	CHROMOSOME	GENE/REGION	LOCATION	SNP ID [a]	BASE PAIR POSITION [a]	ALLELES		A1 FREQUENCY [c]			P-value [d]	Beta [e]
						A1 [b]	A2	Bt20	YRI	CEU		
DBP	13	DACH1 | LOC440145	INTERGENIC	rs17240498	72965307	C	T	0.011	0.000	0.182	4.91×10^{-6}	7.93
	21	ADAMTS5 | C21orf94	INTERGENIC	rs469709	28801007	A	G	0.051	0.019	0.131	9.21×10^{-6}	3.50
	10	LOC642666 | LOC727960	INTERGENIC	rs12761063	82533425	T	C	0.112	0.120	0.101	1.58×10^{-5}	2.47
	2	PLEKHH2 | LOC728819	INTERGENIC	rs13423605	43886460	C	T	0.127	0.176	0.040	2.25×10^{-5}	2.29
	12	SCARB1	INTRON	rs10846744	125312425	G	C	0.227	0.241	0.854	2.49×10^{-5}	-1.79
	2	ARL6IP6 | LOC391453	INTERGENIC	rs2114653	153648587	G	A	0.163	0.144	0.364	4.28×10^{-5}	-1.94
	19	EML2 | GIPR	INTERGENIC	rs4994276	46164172	T	C	0.150	0.199	0.187	5.87×10^{-5}	-1.96
	12	TRPV4	INTRON	rs16939725	110250587	G	A	0.137	0.139	0.015	7.08×10^{-5}	-2.04
	1	NOS1AP	INTRON	rs112468105	162195649	G	C	0.011	0.023	0.000	7.18×10^{-5}	6.69
	12	DNAH10	COMPLEX	rs6488908	124377814	G	A	0.127	0.144	0.040	7.94×10^{-5}	-2.02

Supplemental Digital Content 1 (continued)

PHENOTYPE	CHROMOSOME	GENE/REGION	LOCATION	SNP ID [a]	BASE PAIR POSITION [a]	ALLELES		A1 FREQUENCY [c]			P-value [d]	Beta [e]
						A1 [b]	A2	Bt20	YRI	CEU		
SBP	11	MYRF	INTRON	rs11230796	61529267	G	T	0.058	0.056	0.222	**2.16×10^{-7}**	6.12
	11	MYRF	INTRON	rs400075	61528814	T	C	0.058	0.056	0.217	**2.88×10^{-7}**	6.02
	8	INTS10 | LPL	INTERGENIC	rs55830938	19735188	G	T	0.026	0.028	0.000	1.30×10^{-5}	7.49
	10	LOC100128511 | C10orf114	INTERGENIC	rs6482175	21573536	C	T	0.081	0.111	0.192	1.42×10^{-5}	4.31
	13	DACH1 | LOC440145	INTERGENIC	rs17240498	72965307	C	T	0.011	0.000	0.182	2.10×10^{-5}	11.68
	1	FMO4 | BAT2D1	INTERGENIC	rs10798391	171389938	T	G	0.022	0.014	0.187	2.19×10^{-5}	8.16
	1	NOS1AP	INTRON	rs4657181	162255385	T	A	0.046	0.023	0.556	4.04×10^{-5}	-5.62
	15	CYP19A1	INTRON	rs10459592	51536141	G	T	0.196	0.273	0.581	5.75×10^{-5}	2.81
	8	INTS10 | LPL	INTERGENIC	rs73599609	19756974	C	G	0.050	0.051	0.000	5.78×10^{-5}	5.14
	8	INTS10 | LPL	INTERGENIC	rs73667448	19747475	C	A	0.028	0.042	0.000	6.86×10^{-5}	6.72

Supplemental Digital Content 1 (continued)

	CHROMOSOME	GENE/REGION	LOCATION	SNP ID [a]	BASE PAIR POSITION [a]	ALLELES		A1 FREQUENCY [c]			P-value [d]	Beta [e]
						A1 [b]	A2	Bt20	YRI	CEU		
PHENOTYPE	12	POC1B	INTRON	rs770373	89818289	T	C	0.176	0.329	0.449	$7.05x10^{-5}$	-2.95
	11	STK33	INTRON	rs1596888	8455344	T	G	0.432	0.481	0.722	$7.51x10^{-5}$	2.20
	3	C3orf17 | BOC	INTERGENIC	rs1881941	112852476	A	T	0.123	0.282	0.066	$7.66x10^{-5}$	3.39
	22	FLJ46257 | FAM19A5	INTERGENIC	rs6519991	48725535	A	G	0.173	0.157	0.101	$8.31x10^{-5}$	-2.89
	7	TAX1BP1	INTRON	rs6944913	27826523	G	A	0.019	0.005	0.242	$8.63x10^{-5}$	8.09
	11	KCNQ1	INTRON	rs1557647	2551363	A	G	0.255	0.315	0.682	$9.00x10^{-5}$	2.51
	12	POC1B	INTRON	rs770374	89818022	T	G	0.231	0.366	0.561	$9.05x10^{-5}$	-2.62

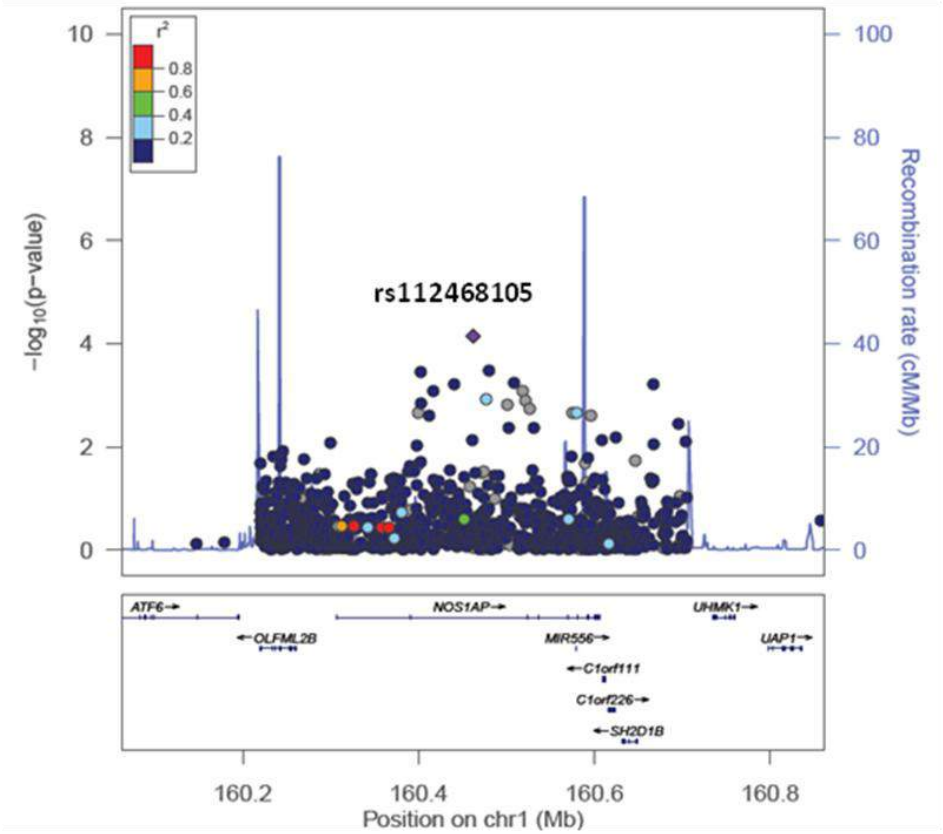
[a] All SNP IDs and base pair positions are reported using *GRCh37* (Genome Reference Consortium human genome Build *37).*

[b] A1 corresponds to the minor allele in the dataset.

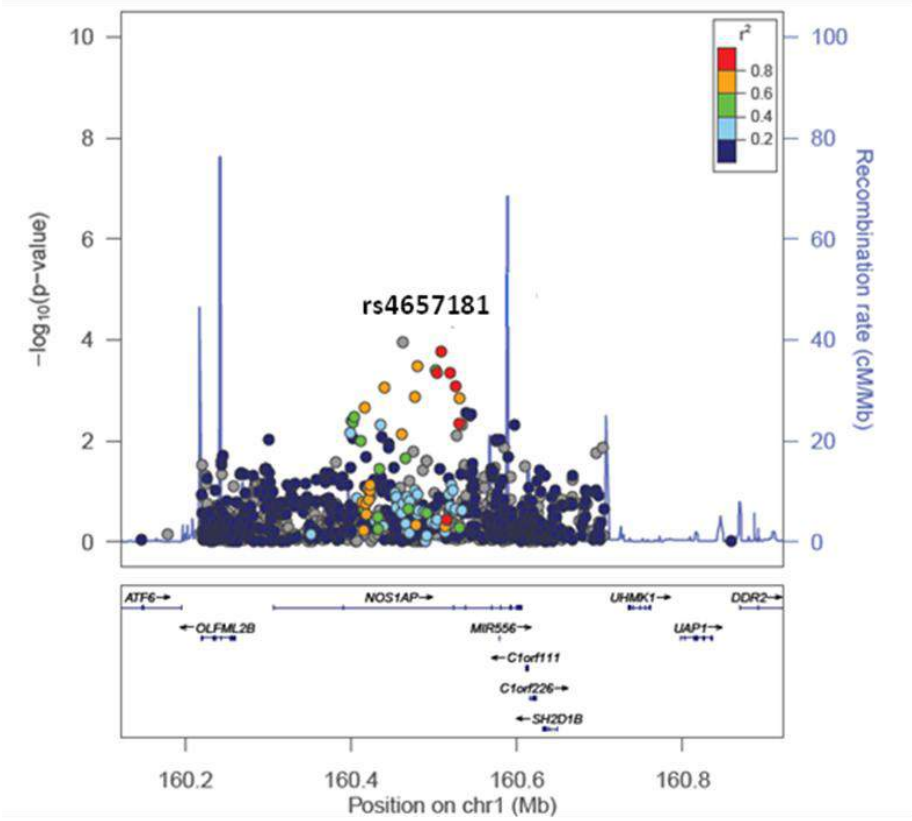
[c] Frequencies of allele 1 are recorded for the merged dataset used in this study (Bt20) and for an African and European 1000 Genomes population - the Yoruba in Ibadan, Nigeria (YRI) and the Utah Residents (CEPH) with Northern and Western Ancestry (CEU).

[d] p-value adjusted for age, sex, BMI and principal components. P-values that pass the "array-wide" significance threshold ($p<6.7x10^{-7}$) are shown in bold.

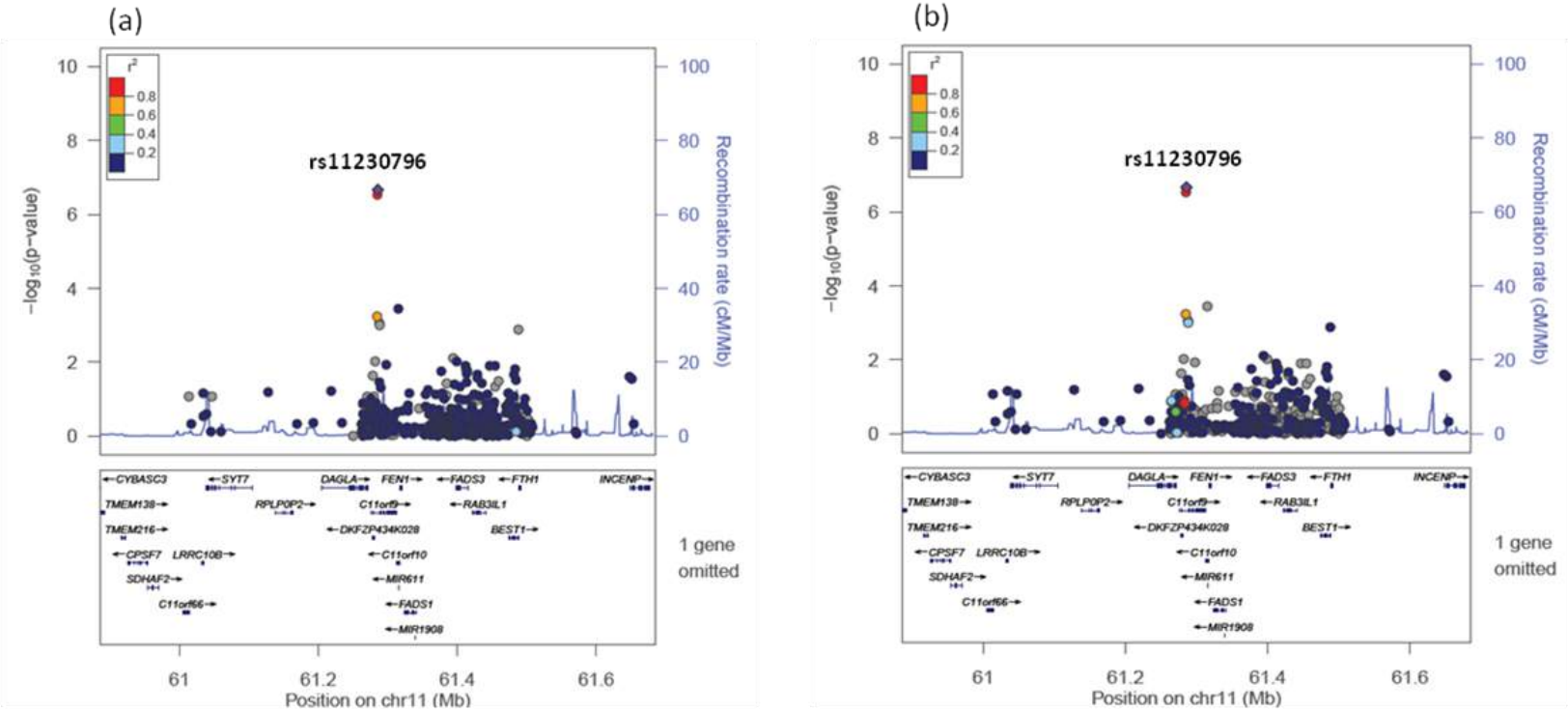
[e] Beta values are with respect to the minor allele in the sample. A positive beta indicates that the minor allele is associated with an increased blood pressure relative to the major allele, and *vice versa*.



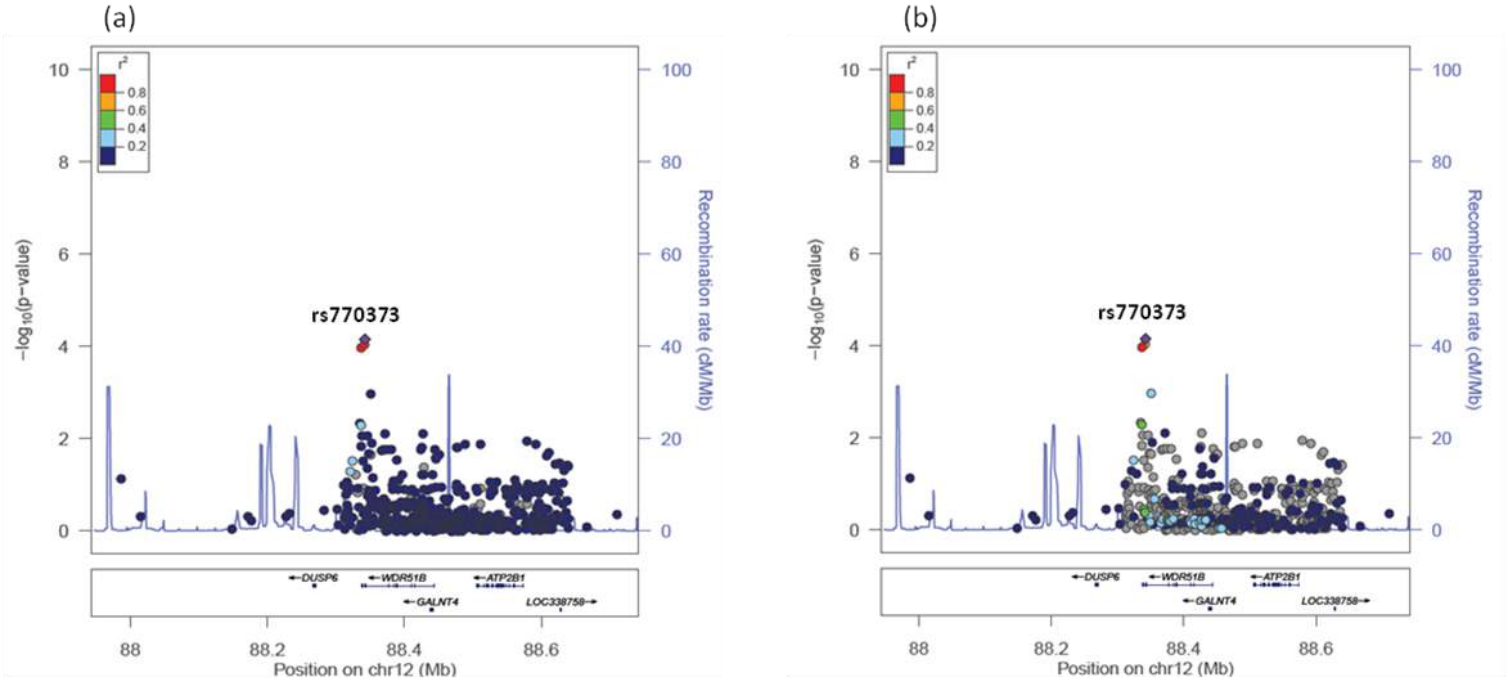
Supplemental Digital Content 2. LocusZoom plot for the association of rs112468105 (in *NOS1AP*) with DBP against a YRI LD background. rs112468105 is represented by a purple diamond. SNPs around this index SNP are coloured according to the LD between each SNP and the index SNP. SNPs with missing LD information are shown in grey. rs112468105 is monomorphic in the CEU population.



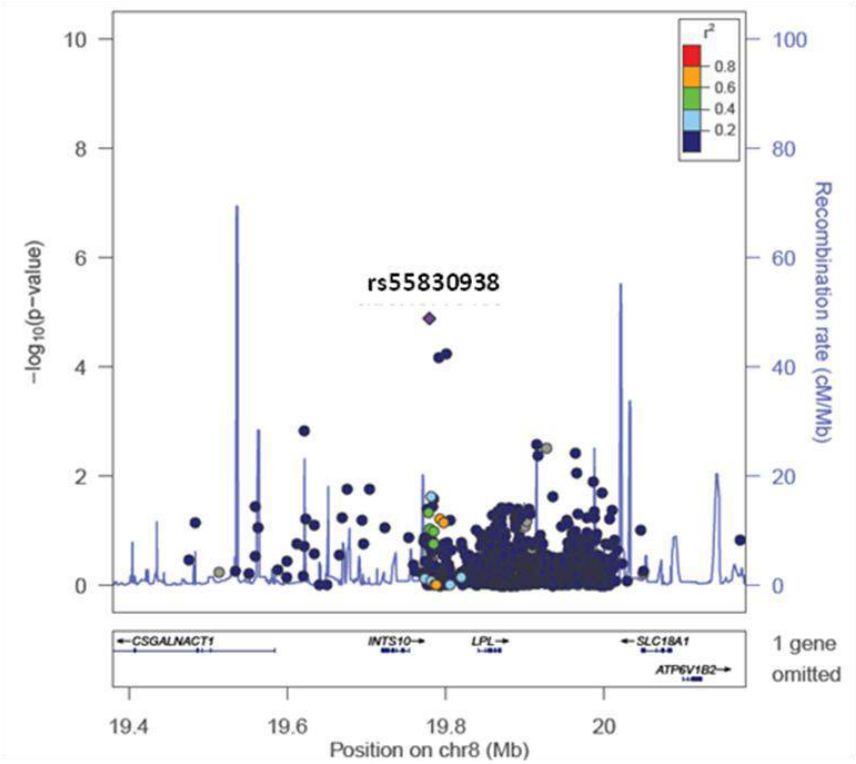
Supplemental Digital Content 3. LocusZoom plot for the association of rs4657181 (in *NOS1AP*) with SBP against a CEU LD background. rs4657181 is represented by a purple diamond. SNPs around this index SNP are coloured according to the LD between each SNP and the index SNP. SNPs with missing LD information are shown in grey. The plot shows evidence of high LD in this region in the CEU population. rs4657181 had completely missing LD information when drawn for the YRI population.



Supplemental Digital Content 4. LocusZoom plots for the association of rs11230796 (in *MYRF*) with SBP against (a) YRI and (b) CEU LD backgrounds. rs11230796 is represented by a purple diamond. SNPs around this index SNP are coloured according to the LD between each SNP and the index SNP. SNPs with missing LD information are shown in grey. *MYRF* is referred to by an alternative name (*C11orf9*) in this plot. The plot shows evidence of high LD in both the YRI and CEU populations between rs11230796 and the other SNP (rs400075) that has a strong association with SBP.



Supplemental Digital Content 5. LocusZoom plots for the association of rs770373 (in *POC1B*) with SBP against (a) YRI and (b) CEU LD backgrounds. rs770373 is represented by a purple diamond. SNPs around this index SNP are coloured according to the LD between each SNP and the index SNP. SNPs with missing LD information are shown in grey. *POC1B* is referred to by an alternative name (*WDR51B*) in this plot. The plot shows evidence of high LD in both the YRI and CEU populations between rs770373 and the other SNP (rs770374) that has a strong association with SBP.



Supplemental Digital Content 6. LocusZoom plot for the association of rs55830938 (intergenic to *INTS10* and *LPL*) with SBP against a YRI LD background. rs55830938 is represented by a purple diamond. SNPs around this index SNP are coloured according to the LD between each SNP and the index SNP. SNPs with missing LD information are shown in grey. rs55830938 is monomorphic in the CEU population.

Supplemental Digital Content 7. Identified regions of interest associated with DBP and SBP stratified into a) sex- and b) age-specific associations.

(a)	CHROMOSOME	GENE/REGION	SNP ID [a]	BASE PAIR POSITION [a]	ALLELES		DBP		SBP	
					A1 [b]	A2	P-value [c]	Beta [d]	P-value [c]	Beta [d]
FEMALES (female caregivers and female Bt20 participants merged)	1	*NOS1AP*	rs112468105	162195649	G	C	4.30×10^{-5}	8.91	9.27×10^{-6}	15.39
			rs4657181	162255385	T	A	-	-	8.08×10^{-5}	-6.66
	11	MYRF	rs11230796	61529267	G	T	-	-	6.76×10^{-7}	7.51
			rs400075	61528814	T	C	-	-	1.45×10^{-6}	7.22
	13	*DACH1 | LOC440145*	rs17240498	72965307	C	T	7.01×10^{-6}	9.16	3.40×10^{-5}	13.49

Supplemental Digital Content 7 (continued)

(b)	CHROMOSOME	GENE/REGION	SNP ID [a]	BASE PAIR POSITION [a]	ALLELES		DBP		SBP	
					A1 [b]	A2	P-value [c]	Beta [d]	P-value [c]	Beta [d]
OLDER INDIVIDUALS (female caregivers only)	1	*NOS1AP*	rs113559977	162202520	G	A	5.21×10^{-5}	10.66	1.37×10^{-5}	18.74
			rs112468105	162195649	G	C	8.20×10^{-5}	10.15	2.41×10^{-5}	17.80
	12	*POC1B*	rs770374	89818022	T	G	-	-	2.83×10^{-5}	-4.63
			rs770373	89818289	T	C	-	-	3.06×10^{-5}	-5.01
		DUSP6 | POC1B	rs770370	89812993	G	A	-	-	3.88×10^{-5}	-4.92
YOUNGER INDIVIDUALS (Bt20 participants only)	12	*POC1B*	rs114077950	89879278	C	T	-	-	3.29×10^{-6}	7.97

[a] All SNP IDs and base pair positions are reported using *GRCh37* (Genome Reference Consortium human genome Build *37).*

[b] A1 corresponds to the minor allele in the dataset.

[c] p-value adjusted for age, sex, BMI and principal components. P-values that pass the "array-wide" significance threshold ($p<6.7 \times 10^{-7}$) are shown in bold.

[d] Beta values are with respect to the minor allele in the sample. A positive beta indicates that the minor allele is associated with an increased blood pressure relative to the major allele, and *vice versa*.

Supplemental Digital Content 8. All SNPs associated with high blood pressure at p ≤ 1 x 10^{-4} in the merged dataset.

CHROMOSOME	GENE/REGION	LOCATION	SNP ID [a]	BASE PAIR POSITION [a]	ALLELES		A1 FREQUENCY [c]			P-value [d]	OR [e]
					A1 [b]	A2	Bt20	YRI	CEU		
8	INTS10 | LPL	INTERGENIC	rs55830938	19735188	G	T	0.026	0.028	0.000	**$5.20x10^{-7}$**	1.17
8	INTS10 | LPL	INTERGENIC	rs73599609	19756974	C	G	0.050	0.051	0.000	$3.46x10^{-6}$	1.12
8	INTS10 | LPL	INTERGENIC	rs73667448	19747475	C	A	0.028	0.042	0.000	$4.13x10^{-6}$	1.15
11	KCNQ1	INTRON	rs1557647	2551363	A	G	0.255	0.315	0.682	$9.25x10^{-6}$	1.05
16	LITAF	INTRON	rs1345441	11679823	G	A	0.344	0.315	0.369	$1.40x10^{-5}$	1.05
22	FLJ46257 | FAM19A5	INTERGENIC	rs13056403	48727674	A	G	0.176	0.171	0.101	$1.47x10^{-5}$	0.94
1	FAM5C	INTRON	rs16832100	190116031	G	A	0.075	0.074	0.071	$2.29x10^{-5}$	0.92
10	LOC642666 | LOC727960	INTERGENIC	rs12761063	82533425	T	C	0.112	0.120	0.101	$2.53x10^{-5}$	1.07
4	MAN2B2	CODING	chr4:6594947	6594947	A	G	0.011			$3.39x10^{-5}$	1.23
1	FMO4 | BAT2D1	INTERGENIC	rs10798391	171389938	T	G	0.022	0.014	0.187	$4.05x10^{-5}$	1.16
15	RAB8B	INTRON	rs12593078	63517347	A	G	0.398	0.449	0.318	$4.08x10^{-5}$	0.96
22	FLJ46257 | FAM19A5	INTERGENIC	rs6519991	48725535	A	G	0.173	0.157	0.101	$4.18x10^{-5}$	0.95
11	KCNQ1	INTRON	rs179431	2552148	G	A	0.323	0.407	0.692	$4.20x10^{-5}$	1.05
11	FAT3	INTRON	rs11019985	92266200	C	T	0.402	0.458	0.071	$4.27x10^{-5}$	0.96
11	KCNQ1	INTRON	rs179434	2553601	A	C	0.208	0.384	0.692	$4.43x10^{-5}$	1.05

Supplemental Digital Content 8 (continued)

CHROMOSOME	GENE/REGION	LOCATION	SNP ID [a]	BASE PAIR POSITION [a]	ALLELES		A1 FREQUENCY [c]			P-value [d]	OR [e]
					A1 [b]	A2	Bt20	YRI	CEU		
8	NAT2 | PSD3	INTERGENIC	rs17126618	18278776	C	T	0.025	0.023	0.005	$4.79x10^{-5}$	1.14
11	PLEKHA7 | LOC729362	INTERGENIC	rs177544	16925046	T	C	0.030	0.028	0.157	$4.89x10^{-5}$	1.13
15	SGK269 | HMG20A	INTERGENIC	rs17385059	77637812	C	T	0.107	0.060	0.293	$5.72x10^{-5}$	1.07
8	NAT1 | NAT2	INTERGENIC	rs1565684	18246664	G	A	0.357	0.273	0.465	$5.75x10^{-5}$	1.04
10	CDC123	INTRON	rs117334547	12284397	T	A	0.037	0.028	0.025	$6.14x10^{-5}$	1.11
10	CDC123	INTRON	rs77611247	12285815	A	G	0.037	0.028	0.040	$6.14x10^{-5}$	1.11
11	ARHGEF12 | GRIK4	INTERGENIC	rs6589829	120530973	T	C	0.145	0.176	0.601	$6.99x10^{-5}$	0.95
13	SLC7A1	UTR	rs2490264	30084535	T	G	0.094	0.139	0.015	$7.11x10^{-5}$	1.07
11	SLC22A18	INTRON	rs60055747	2941104	A	G	0.016	0.037	0.000	$8.52x10^{-5}$	1.17
15	RAB8B	INTRON	rs72747098	63525249	A	G	0.464	0.380	0.682	$9.02x10^{-5}$	1.04

Possible regions of interest are shaded in grey.

[a] All SNP IDs and base pair positions are reported using *GRCh37* (Genome Reference Consortium human genome Build *37).*

[b] A1 corresponds to the minor allele in the dataset.

[c] Frequencies of allele 1 are recorded for the merged dataset used in this study (Bt20) and for an African and European 1000 Genomes population - the Yoruba in Ibadan, Nigeria (YRI) and the Utah Residents (CEPH) with Northern and Western Ancestry (CEU).

[d] p-value adjusted for age, sex, BMI and principal components. P-values that pass the "array-wide" significance threshold ($p<6.7x10^{-7}$) are shown in bold.

[e] Odds ratio (OR) values are with respect to the minor allele in the sample. An OR greater than one indicates that the minor allele is associated with an increased risk for high blood pressure relative to the major allele, and *vice versa*.

Supplemental Digital Content 9. Identified regions of interest associated with high blood pressure stratified into a) sex- and b) age-specific associations.

(a)

	CHROMOSOME	GENE/REGION	SNP ID [a]	BASE PAIR POSITION [a]	ALLELES		P-value [c]	OR [d]
					A1 [b]	A2		
FEMALES (female caregivers and female Bt20 participants merged)	11	KCNQ1	rs179434	2553601	A	C	1.77×10^{-5}	1.07
			rs1557647	2551363	A	G	7.17×10^{-6}	1.07
			rs179431	2552148	G	A	1.83×10^{-5}	1.06
	22	*FLJ46257 | FAM19A5*	rs13056403	48727674	A	G	1.94×10^{-5}	0.93
MALES (male Bt20 participants only)	8	*INTS10 | LPL*	rs73667448	19747475	C	A	2.97×10^{-5}	12.48

Supplemental Digital Content 9 (continued)

(b)	CHROMOSOME	GENE/REGION	SNP ID [a]	BASE PAIR POSITION [a]	ALLELES		P-value [c]	OR [d]
					A1 [b]	A2		
OLDER INDIVIDUALS (female caregivers only)	15	*RAB8B*	rs12593078	63517347	A	G	8.49×10^{-5}	0.59
	22	*FLJ46257 | FAM19A5*	rs13056403	48727674	A	G	7.66×10^{-6}	0.39
YOUNGER INDIVIDUALS (Bt20 participants only)	8	*INTS10 | LPL*	rs73667448	19747475	C	A	9.51×10^{-5}	4.91

[a] All SNP IDs and base pair positions are reported using *GRCh37* (Genome Reference Consortium human genome Build *37).*

[b] A1 corresponds to the minor allele in the dataset.

[c] p-value adjusted for age, sex, BMI and principal components. P-values that pass the "array-wide" significance threshold ($p<6.7 \times 10^{-7}$) are shown in bold.

[d] Odds ratio (OR) values are with respect to the minor allele in the sample. An OR greater than one indicates that the minor allele is associated with an increased risk for high blood pressure relative to the major allele, and *vice versa*.

Chapter 5: AN INVESTIGATION INTO GENOTYPE IMPUTATION OF METABOCHIP DATA IN BLACK SOUTH AFRICANS USING A MIXED REFERENCE PANEL

Genotype imputation is the inference of genotype data at SNPs not initially genotyped using information from reference panel haplotypes. It is a useful extension of GWAS and other association studies to provide a more detailed view of an associated region, to provide a wider range of SNPs for analysis and follow-up or replication and to point to possible causal variants. Imputation in African individuals can be challenging due to their genetic diversity and lower levels of LD [*See Chapter 1*].

In this investigation, the aim was to explore the success of imputation in black South African individuals using Metabochip data as a starting dataset and a mixed population 1000 Genomes Project Phase 3 reference panel (The 1000 Genomes Project Consortium, 2015). To do this, regions where association signals were observed in the merged dataset or individual datasets with either DBP (*NOS1AP*) or SBP (*NOS1AP, POC1B, MYRF*) were selected.

5.1 Methodology

5.1.1 Preparation of files for imputation

Each imputation run focused only on the specific gene of interest, rather than the entire genome. The first step therefore involved extracting the genotype data of the chromosome in which the gene of interest sits from the merged, pruned, Build 37 file [*See Chapter 2*].

The SNPs in the reference panel are all aligned to the sense strand whereas the SNPs on the Metabochip are of a mixed strand orientation (sense and antisense).

Therefore, in preparation for imputation, a strand-check was run using SHAPEIT (v2) (Delaneau et al., 2012) and a strand flip was carried out in PLINK for any SNPs in the genotype data that did not align with the reference panel data. Any SNPs that still presented with problems following a second strand check were excluded from the dataset.

Imputation involved two main steps – a pre-phasing step and the actual imputation step.

5.1.2 Pre-phasing

Pre-phasing [S*ee Chapter 1*] was carried out using SHAPEIT using the merged (caregivers and participants) dataset. This step involved processing the study genotypes to produce the best-guess haplotypes and was carried out over the entire chromosome containing the gene of interest.

5.1.3 Imputation

Alleles from the 1000 Genomes Project Phase 3 mixed population reference panel (The 1000 Genomes Project Consortium, 2015) were then imputed into the estimated haplotypes in IMPUTE (v2.3.1) (Howie et al., 2009) over the specific genes of interest in the merged or individual dataset only. IMPUTE2 generates an info metric for each SNP which gives an indication of the certainty with which each SNP is imputed. An info metric of 1 indicates that the SNP has been imputed with high certainty. Any SNPs with an info metric < 0.4 were removed before further analysis (*see below*).

5.1.4 Association analysis

For the merged dataset, post-imputation association analysis was carried out in GEMMA (v 0.94.1) (to be able to account for relatedness between individuals)

using univariate linear mixed models with correction for covariates (age, BMI and sex and PCs, where applicable). The imputation output files were first converted to PLINK files using fcGENE (Roshyara & Scholz, 2014) and all SNPs with an info metric < 0.4 were excluded. Creation of the relatedness matrix involved input of the phenotype data into the PLINK .fam files and then calculation of the matrix in GEMMA for each of the phenotypes under investigation.

For the individual datasets, post-imputation association analysis was carried out using SNPTEST (v2.5.1) (an appropriate tool to use in conjunction with IMPUTE2) using an additive model with correction for covariates (age, BMI and sex (where applicable)).

The significance threshold for post-imputation analysis to measure array-wide significance was also calculated here as 0.05 divided by the number of unlinked markers. The significance thresholds for the different datasets and genes investigated are shown in **Table 5.1**.

Table 5.1 Significance thresholds for post-imputation analysis.

		Gene		
		NOS1AP	*MYRF*	*POC1B*
Dataset	Merged	$p<2.9x10^{-5}$ (0.05/1724)	$p<3.0x10^{-4}$ (0.05/167)	$p<9.3x10^{-5}$ (0.05/538)
	Female Caregivers	$p<3.1x10^{-5}$ (0.05/1616)	-	$p<9.9x10^{-5}$ (0.05/505)
	Bt20 Participants	-	-	$p<9.9x10^{-5}$ (0.05/507)

5.1.5 Result visualisation

Post-imputation plots showing genotyped and imputed SNPs together for each gene were drawn in R.

5.2 Results

5.2.1 Overall assessment of accuracy and yield

To get an indication of the overall quality of the imputation, IMPUTE2 performs an internal cross-validation by masking and re-imputing genotyped SNPs one at a time and then comparing the imputed genotypes with the original genotypes to give an overall concordance value. *NOS1AP* (merged and female caregivers), *MYRF* (merged) and *POC1B* (merged, female caregivers and Bt20 participants) were all imputed with high confidence (98.1%, 96.9% and 99.4% concordance between genotyped and imputed SNPs for the three genes, respectively, before applying any filters).

Table 5.2 outlines the number of SNPs before and after imputation in the region under investigation and the number of SNPs with an info metric ≥ 0.4 (well-imputed SNPs) that were taken through to analysis. **Figure 5.1** shows the proportion of SNPs with an info metric ≥ 0.4 in different MAF bins. The bins are separated into MAF=0, rare SNPs (0<MAF≤0.01), less common SNPs (0.01<MAF≤0.05) and common SNPs (MAF>0.05). In all but one scenario, the highest proportion of SNPs with info metric ≥ 0.4 are common SNPs.

Table 5.2 Number of SNPs pre- and post-imputation and with info metric ≥ 0.4 for each imputation scenario considered.

Gene	Dataset	Number of SNPs pre-imputation	Number of SNPs post-imputation [a]	Number of SNPs (%) with info metric ≥0.4 [b]
NOS1AP	Merged	568	8633	3078 (35.7)
	Female Caregivers			2958 (34.3)
MYRF	Merged	53	966	246 (25.5)
POC1B	Merged	222	2920	987 (33.8)
	Female Caregivers			954 (32.7)
	Bt20 Participants			965 (33.0)

[a] This indicates the number of SNPs resulting from imputation, before the info metric filter was applied.

[b] SNPs with an info metric ≥0.4 were taken through to association analysis.

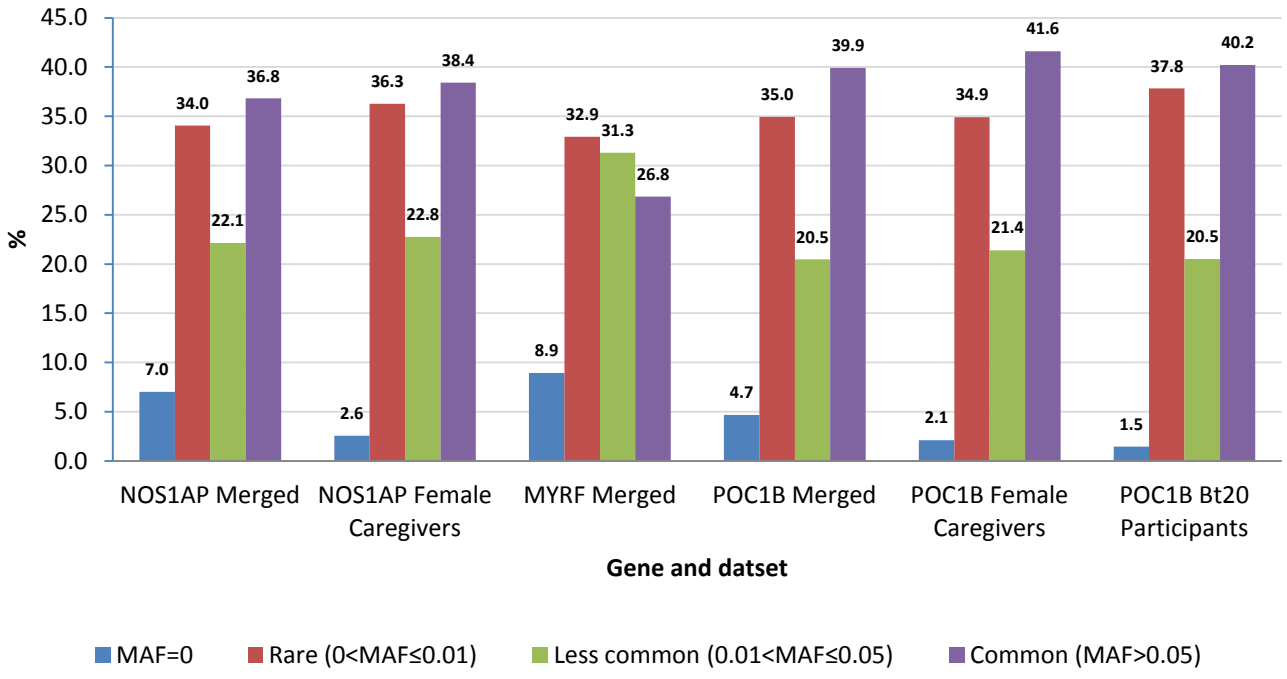


Figure 5.1 The proportion of well-imputed SNPs (info metric ≥ 0.4) in different MAF bins for each imputation scenario considered.

5.2.2 Merged dataset

For the merged dataset, a more detailed view/enrichment of the region, where an association signal was observed before imputation, was only evident in *NOS1AP* for association with both DBP and SBP (**Figure 5.2**). Top associated SNPs are shown in **Tables 5.3**. For both DBP and SBP, the pre-imputation associated SNPs had a higher p-value (DBP: rs112468105 $p=1.18x10^{-4}$; SBP: rs4657181 $p=1.01x10^{-4}$) than the top imputed SNPs and previously observed.

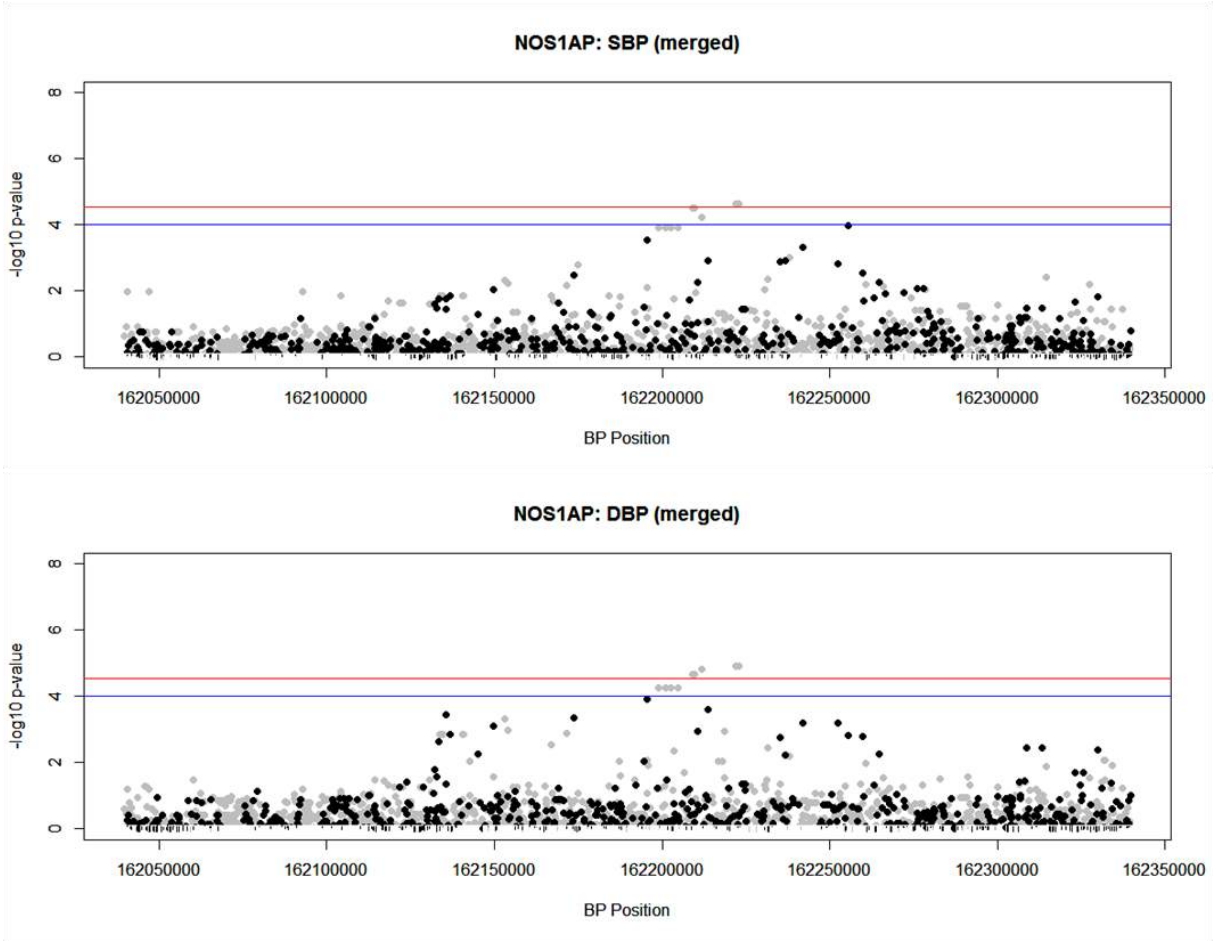


Figure 5.2a Imputation in the merged dataset resulted in a more detailed view/enrichment of the region where an association signal was observed before imputation for *NOS1AP* (SBP and DBP). Imputed SNPs are in grey and genotyped SNPs are in black. The blue significance line is at $p=1.0x10^{-4}$ and the red significance line is at $p=2.9x10^{-5}$.

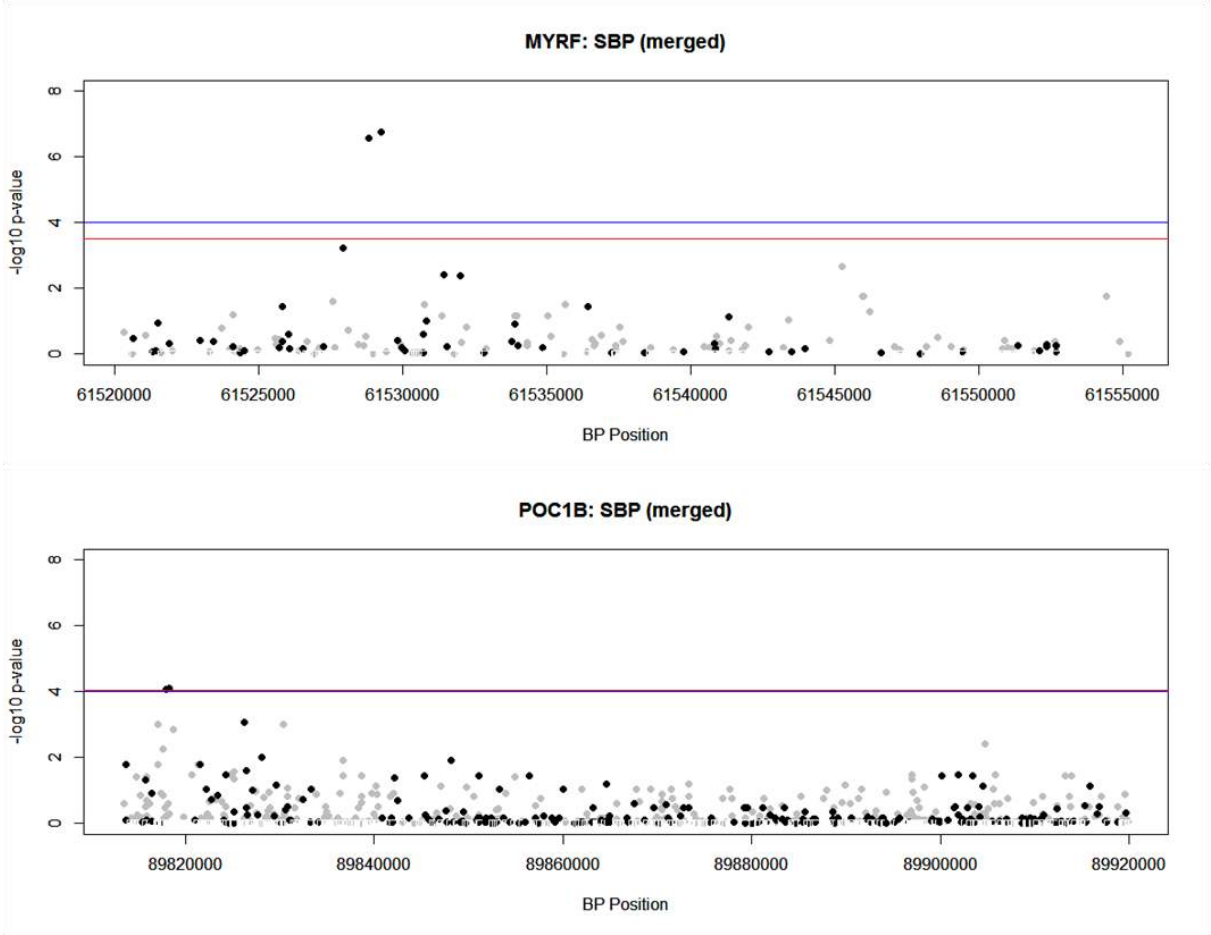


Figure 5.2b Imputation in the merged dataset didn't result in a more detailed view/enrichment of the region where an association signal was observed before imputation for *MYRF* or *POC1B*. Imputed SNPs are in grey and genotyped SNPs are in black. The blue significance line is at $p=1.0x10^{-4}$ and the red significance line is at $p=3.0x10^{-4}$ for *MYRF* and $p=9.3x10^{-5}$ for *POC1B*.

Table 5.3a Several imputed SNPs in the *NOS1AP* gene associated with DBP in the merged and female caregiver datasets. SNPs with p>1x10^{-4} are shown, with SNPs meeting the calculated significance threshold for multiple testing indicated in bold.

SNP ID[a]	Merged dataset			Female caregiver dataset		
	MAF	P-value[b]	Beta	MAF	P-value[b]	Beta
rs530351339	0.012	**1.21x10^{-5}**	7.25	0.013	**2.86x10^{-5}**	0.87
rs115986274	0.012	**1.21x10^{-5}**	7.25	0.013	**2.91x10^{-5}**	0.87
rs116537698	0.011	**1.50x10^{-5}**	7.26	0.012	**1.91x10^{-5}**	0.91
rs75879850	0.010	**2.07x10^{-5}**	7.51	0.011	**1.41x10^{-5}**	0.96
rs79382913	0.010	**2.07x10^{-5}**	7.51	0.011	**1.46x10^{-5}**	0.95
rs114010317	0.011	5.37x10^{-5}	7.03	0.012	6.43x10^{-5}	0.84
rs114372970	0.011	5.37x10^{-5}	7.03	0.012	5.48x10^{-5}	0.85
rs113559977	0.011	5.37x10^{-5}	7.03	0.011	5.43x10^{-5}	0.85
rs145140073	0.011	5.37x10^{-5}	7.03	0.011	4.23x10^{-5}	0.88
rs77776939	-	-	-	0.009	**1.93x10^{-8}**	4.06
rs17459307	-	-	-	0.002	**2.73x10^{-6}**	4.28
rs112468105*	-	-	-	0.012	8.50x10^{-5}	0.81

[a] All SNP IDs are reported using *GRCh37* (Genome Reference Consortium human genome Build *37).*

[b] p-value adjusted for age, sex (where appropriate), BMI and PCs (where appropriate).

*SNP associated before imputation.

Table 5.3b Several imputed SNPs in the *NOS1AP* gene associated with SBP in the merged and female caregiver datasets. SNPs with p>1x10^{-4} are shown, with SNPs meeting the calculated significance threshold for multiple testing indicated in bold.

SNP[a]	Merged dataset			Female caregiver dataset		
	MAF	P-value[b]	Beta	MAF	P-value[b]	Beta
rs530351339	0.012	**2.33x10^{-5}**	11.20	0.013	**4.03x10^{-6}**	0.94
rs115986274	0.012	**2.33x10^{-5}**	11.20	0.013	**4.18x10^{-6}**	0.94
rs75879850	0.010	3.17x10^{-5}	11.70	0.011	**2.19x10^{-6}**	1.03
rs79382913	0.010	3.17x10^{-5}	11.70	0.011	**2.38x10^{-6}**	1.02
rs116537698	0.011	5.75x10^{-5}	10.80	0.012	**2.38x10^{-6}**	0.99
rs77776939	-	-	-	0.009	**9.50x10^{-10}**	4.33
rs144974259	-	-	-	0.006	**2.18x10^{-7}**	3.40
rs188310846	-	-	-	0.001	**7.18x10^{-7}**	5.40
rs17459307	-	-	-	0.002	**5.88x10^{-6}**	4.29
rs145140073	-	-	-	0.011	**9.15x10^{-6}**	0.93
rs113559977	-	-	-	0.011	**1.45x10^{-5}**	0.90
rs114372970	-	-	-	0.012	**1.46x10^{-5}**	0.90
rs114010317	-	-	-	0.012	**1.78x10^{-5}**	0.88
rs112468105*	-	-	-	0.012	**2.53x10^{-5}**	0.85

[a] All SNP IDs are reported using *GRCh37* (Genome Reference Consortium human genome Build *37).*

[b] p-value adjusted for age, sex (where appropriate), BMI and PCs (where appropriate).

*SNP associated before imputation.

5.2.3 Individual datasets

For the individual datasets, a more detailed view/enrichment of the region, where an association signal was observed before imputation, was again evident in *NOS1AP* for association with both DBP and SBP in the female caregivers and to a lesser extent in *POC1B* for association with SBP in the Bt20 participants (**Figure 5.3**).

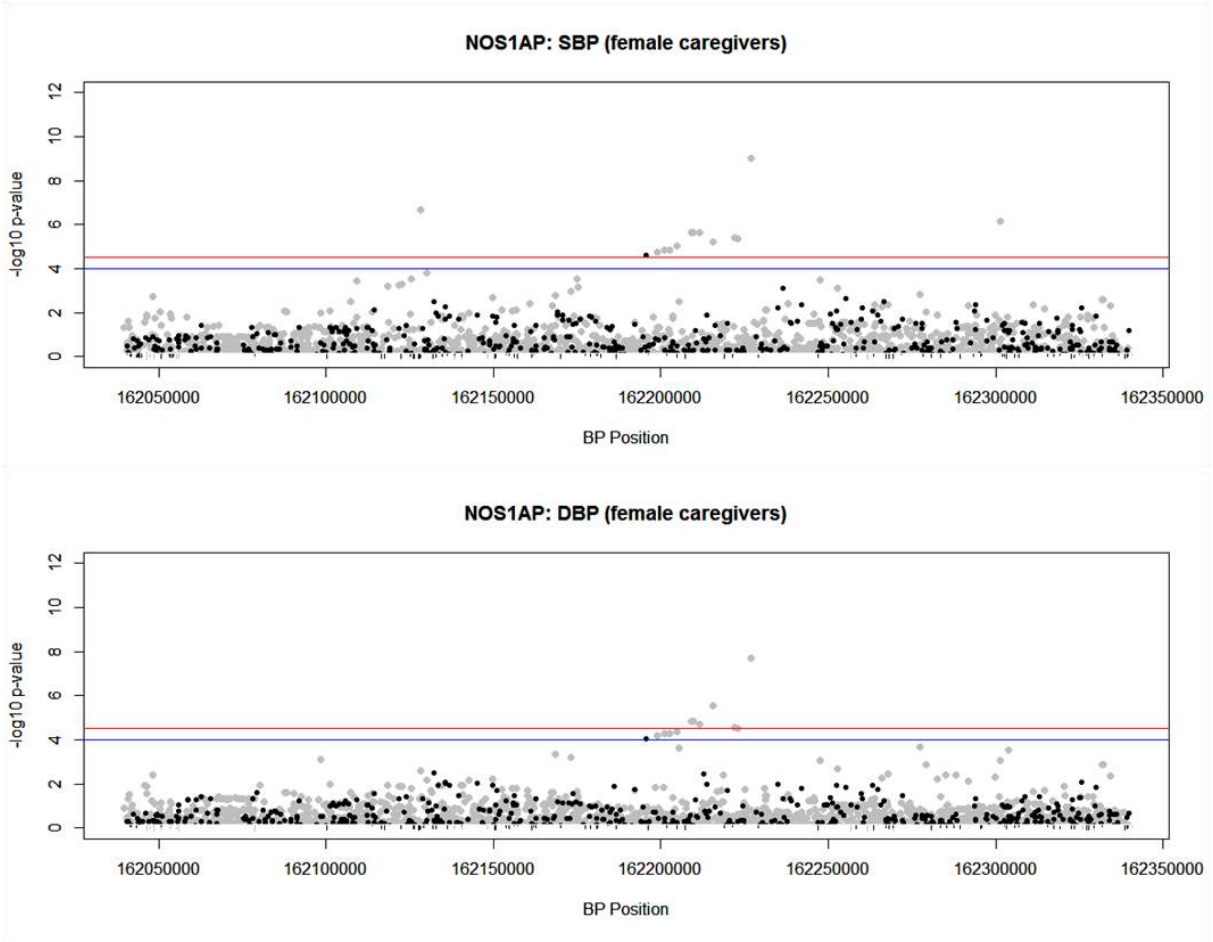


Figure 5.3a Imputation in the individual datasets resulted in a more detailed view/enrichment of the region where an association signal was observed before imputation for *NOS1AP* (SBP and DBP) in the female caregivers. Imputed SNPs are in grey and genotyped SNPs are in black. The blue significance line is at $p=1.0x10^{-4}$ and the red significance line is at $p=3.1x10^{-5}$.

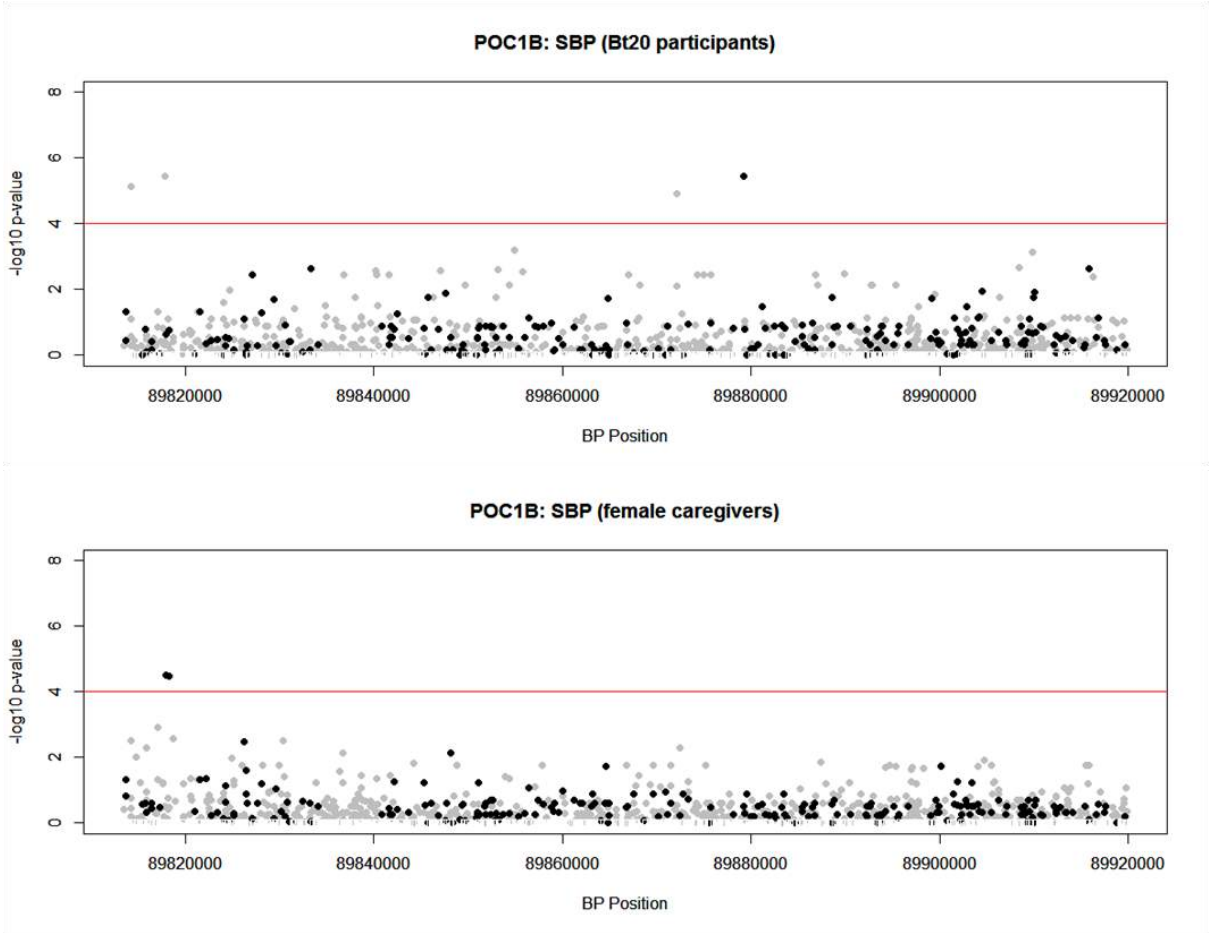


Figure 5.3b Imputation in the individual datasets resulted in a very slight enrichment of the region where an association signal was observed before imputation for *POC1B* (SBP) in the Bt20 participants and no enrichment for *POC1B* (SBP) in the female caregivers. Imputed SNPs are in grey and genotyped SNPs are in black. The blue significance line is at $p=1.0x10^{-4}$ and the red significance line is at $p=9.9x10^{-5}$.

Top associated SNPs are shown in **Table 5.3** (*above*) **and Table 5.4**. For *NOS1AP* for both DBP and SBP, the pre-imputation associated SNPs had a higher p-value (DBP: rs112468105 $p=8.50x10^{-5}$; SBP: rs112468105 $p= 2.53x10^{-5}$ and rs4657181 $p=2.24x10^{-3}$) than the top imputed SNPs and previously observed. For *POC1B*, the pre-imputation associated SNP was more significantly associated with SBP than any of the top imputed SNPs.

Table 5.4. A few imputed SNPs in the *POC1B* gene associated with SBP in the Bt20 participant dataset. SNPs with p>1x10^{-4} are shown, with SNPs meeting the calculated significance threshold for multiple testing indicated in bold.

SNP ID[a]	Bt20 participant dataset		
	MAF	P-value[b]	Beta
rs114077950*	0.022	**3.55x10^{-6}**	0.71
rs148043872	0.022	**3.55x10^{-6}**	0.71
rs544538503	0.003	**7.28x10^{-6}**	-4.04
rs146263687	0.007	**1.16x10^{-5}**	1.26

[a] All SNP IDs are reported using *GRCh37* (Genome Reference Consortium human genome Build *37).*

[b] p-value adjusted for age, sex, BMI and PCs (where appropriate).

*SNP associated before imputation.

5.3 Discussion

Imputation was carried out in few identified regions of interest using SHAPEIT (for pre-phasing) and IMPUTE2 (for imputation) and a 1000 Genomes mixed reference panel to investigate the success of imputation in our South African population that was genotyped using the Metabochip. Imputation was achieved with high confidence in all genes (>95% concordance between genotyped and imputed SNPs in each gene), but a more detailed view or improvement in the resolution of the region was only seen in *NOS1AP* (DBP and SBP in both the merged and female caregiver datasets) and *POC1B* (Bt20 participant dataset only). In each of these cases, imputation resulted in more SNPs associated with SBP and/or DBP, with the imputed SNPs in *NOS1AP* being more significantly associated with SBP/DBP than any significantly associated genotyped SNPs.

Despite the high concordance seen in each gene, the percentage of SNPs with info metric < 0.4 was quite high meaning that a large number of SNPs were removed prior to association analysis. This has, however, been seen before in a study comparing imputation performance using the 1000 Genomes pilot CEU (~8.5 million SNPs), interim EUR (~11.5 million SNPs) and Phase 1 ALL (~37.4 million SNPs) reference panels across different GWAS datasets. When using the 1000G Phase 1 ALL reference panel, on average only about 28% of the SNPs had an info metric ≥ 0.4, compared to the pilot and interim reference panels with an average of 87% and 67% with an info metric ≥ 0.4, respectively, across the different GWAS datasets (Zheng et al., 2015a). Despite the much lower percentage, the actual number of well-imputed SNPs was still higher when using the Phase 1 reference panel due to there being many more SNPs in the reference panel to start with. With the later 1000 Genomes Phase 3 reference panel, there may also be a large proportion of SNPs with an info metric < 0.4, but the actual number of well-imputed SNPs should still be reasonably large due to the even higher number of variants in the reference panel (> 88 million variants).

An assessment of the proportion of SNPs with info metric ≥ 0.4 in different MAF bins showed that in all but one scenario, the highest proportion of SNPs with info metric ≥ 0.4 are common SNPs. This is not surprising as common SNPs are generally more easily imputed than rare variants (Liu et al., 2012; Sung et al., 2012a; Chanda et al., 2012; Band et al., 2013; Zheng et al., 2015a). Compared to *NOS1AP* and *POC1B*, imputation in *MYRF* appeared to give a different picture in terms of the proportion of SNPs with info metric ≥ 0.4 in different MAF bins, with the highest proportion being rare variants. In addition, *MYRF* appears to be the least successfully imputed of the three genes as it has the greatest proportion of SNPs with info metric < 0.4. As it is known that a lower SNP density in the study sample negatively influences imputation accuracy and performance (Zhang et al., 2011; Wang et al., 2012; Howie et al., 2012), one might speculate that the number of SNPs present on the Metabochip for *MYRF* may be too small to allow

for successful imputation in this gene. The representation of SNPs for *each* of the genes, however, is relatively low on the Metabochip (less than 5% of the total variants present in each gene were available for imputation) and an increase in the SNP density in each case may improve imputation. The lack of an enhanced signal in *MYRF* may indicate that the gene is in fact not associated with SBP or the previously identified SNPs may be the actual causal variant(s).

In this study, imputation was investigated in the merged and individual caregiver and participant sets for comparison. In each scenario, however, the pre-phasing step was performed using the merged dataset. Larger study sample sizes can increase haplotype estimation accuracy during phasing (Howie et al., 2009). In addition, merging of the two datasets meant that there were a large number of related individuals present in one dataset. Related individuals can have a positive influence on the haplotype estimations as they have longer stretches of shared haplotypes (Li et al., 2009). If haplotypes are well estimated, the subsequent imputation can be fast and highly accurate (Howie et al., 2012). The actual imputation in the different scenarios was comparable with a number of overlapping associated SNPs in the merged and individual datasets, as is seen in *NOS1AP*.

Pre-imputation filtering of SNPs is believed to influence imputation accuracy, with little or no SNP filtering being favourable for imputing small to moderately sized datasets to keep the LD structure between SNPs intact (Roshyara et al., 2014). In this study, the data went through a strict QC process to remove all SNPs and samples that failed certain criteria. This was, however, deemed necessary for all parts of the study, regardless of any effect it may have on LD structure between SNPs and imputation accuracy.

Mixed/"cosmopolitan" reference panels have proven to be successful for imputation in several populations and are particularly useful in cases where no

clear reference panel matches exist (Roshyara et al., 2016). This was therefore the ideal choice in this study as no suitable publically available reference panels matching our South African population currently exists. Mixed reference panels are also thought to be able to improve imputation of rare variants (Howie et al., 2009; Liu et al., 2014; Howie et al., 2011). This could be particularly useful when dealing with a phenotype such as BP where many small contributions from risk alleles in multiple genes across the genome play a role in its aetiology (Doris, 2002). Many of these contributions may be from low frequency or rare variants which may only be identified following imputation. In fact, all of the imputed SNPs in *NOS1AP* or *POC1B* that were associated with SBP or DBP fell into the rare or less common SNP categories.

The choice of imputation tool is an important consideration. The chosen tool should, among other things, be fast, easy to install and handle, have meaningful default options and feed useful information back to the user (Ellinghaus et al., 2009). In addition, it should be accurate in its imputation. IMPUTE (1 and 2) is a sensible choice as it shows generally favourable performance compared to other commonly used imputation tools [*See Chapter 1*].

Some studies have reported on the merit of using a mixed or "cosmopolitan" 1000 Genomes reference panel in conjunction with IMPUTE 2 (Hancock et al., 2012; Liu et al., 2014). In addition, using IMPUTE2 combined with a diverse reference panel from HapMap was applicable to African populations (Band et al., 2013). Given the advantage of 1000 Genomes reference panels over HapMap reference panels, our use of IMPUTE2 and a diverse 1000 Genomes Phase 3 reference panel was an appropriate choice for imputation in our African population.

Pre-phasing has become a popular option when performing imputation as it speeds up imputation and allows for phasing of a particular gene/region to only

be performed once with multiple subsequent imputation runs. Pre-phasing was therefore chosen to be included as a step in this study and was carried out in SHAPEIT, a tool that is highly compatible with IMPUTE2. One recent study, however, suggests using the SHAPEIT-IMPUTE2 framework cautiously. The study ran pre-phasing using SHAPEIT and subsequent imputation with IMPUTE2 and using a HapMap reference panel. They reported that the SHAPEIT-IMPUTE2 framework can overestimate the certainty of genotype distributions and inflate the IMPUTE info metric. This leads to a low percentage of correctly imputed SNPs, thus decreasing imputation accuracy. This was reported to be particularly evident with smaller sample sizes. (Roshyara et al., 2016)

A different tool was used for association analysis in the merged and individual datasets. SNPTEST is considered an appropriate choice when using IMPUTE2, as it works well with its output. In analysis of the merged dataset, however, GEMMA was used to be able to account for the relatedness between individuals through incorporation of a relatedness matrix. Given the overlapping results in the merged and individual datasets using the two different tools, GEMMA appears to handle imputed data and genotype uncertainty as well as SNPTEST and is therefore probably a suitable enough tool to use following conversion of the IMPUTE2 output files into a format useable by GEMMA.

Imputation in Africans, in general, can be a challenge due to their high genetic diversity and lower levels of LD, which could lead to a reduced imputation accuracy (Howie et al., 2011; Huang et al., 2011). A bigger sample size is often needed to maintain the power in imputation based studies in Africans (Huang et al., 2011). The size of the sample used in this study is relatively small and increasing the sample size may improve imputation performance in the future.

This current investigation into imputation in our Metabochip-genotyped black South African population using a mixed 1000 Genomes reference panel showed

some promising results, especially in the *NOS1AP* gene, despite the high number of SNPs removed after imputation. Imputation is a useful addition to association studies to increase the SNP set for association analysis, thus increasing the power of the study and possibly helping to identify novel or subtle associations. Imputation in any population is not perfect and can produce inaccuracies. All imputation-based results should therefore be replicated, preferably through actual genotyping of the imputed SNPs in a separate dataset (Browning, 2008). Until now, many of the imputation-based studies in individuals of African ancestry have been carried out in African Americans, who are admixed and therefore not a true representation of Africans. More imputation-based studies need to be carried out in African populations to inform the most appropriate reference panel and parameter settings and to determine whether or not more exact matching reference panels, as they become available, may in fact perform better than the mixed panel.

Chapter 6: GENERAL CONCLUSIONS

Bioinformatics is an area of ever-increasing importance. It is an interdisciplinary field involving research into biological areas of interest using various computer- or programming-based tools or software (both pre-existing and self-developed) and often involves the handling of large, complex datasets. Many areas of Bioinformatics require that a researcher has insight into both the biological and computational aspects necessary to tackle a scientific problem. Marrying these two areas provides the potential for significant discoveries. All aspects of the research presented here falls into the broader field of Bioinformatics.

A major part of this study involved the development of a queryable cardiometabolic database for the current longitudinal project-specific data. The underlying motivation for this is the ever-increasing volume and complexity of biological data and the need for this data to be effectively stored, managed and used for biological knowledge discovery. Something which could be useful is to store project-specific data similar to how publically available data is stored with access to the data via an internet-based user interface. The phenotype (from multiple data collection time points), SNP annotation and association analysis data from this study was moved from basic Excel spreadsheets to structured tables making up a relational MySQL database. A user-friendly web interface linking to the database was designed using PHP, HTML and CSS code. The interface has made the project-specific data more easily accessible and queryable for useful data and information from any computer with internet access. The previous method of storing the data in spreadsheets was limited by the size of data that could be stored, the requirement to manually manipulate the data to obtain useful information and the fact that the data was spread across multiple files. Having all the data stored in one central location as a relational database means better management of the data, less chance of data handling errors and easier extraction of data and information with more complex

queries possible. Well managed data allows for studies to be conducted smoothly and for the data to be utilised to its full potential for discovery within the biological field. Access to the database is restricted to specific users within the research group via a username/password login. This and the session timeout feature ensure that the data remains protected and won't be accessed by anyone without permission to use the data. If necessary, access to some tables containing sensitive data can also be restricted. The stored association analysis results also provide a means by which future follow-up or replication studies can be better informed. A useful extension of the current database will be to include more participants from the Bt20 cohort, data from future data collection time points, additional phenotypes for the same individuals and the data or information resulting from genotype imputation. New features for more complex queries and extraction of additional useful information can also be added. This can specifically include longitudinal queries for extraction of useful information about changes in or stability of phenotypes over time.

The database and interface do not only serve a purpose in the current study or research area, but can serve as a tool for other research groups to implement their own similar databases to improve management and analysis of data in various biological settings. The developed database and interface, therefore, have room for growth and improvement. As a project that makes use of a database such as the one described here expands, the database itself can expand in terms of its functionality and available information. Careful consideration must be given to the needs of the researchers involved and what information is required by them to carry out their investigations. The current design is able to include the basic information required for a genetic association study involving a relatively small set of individuals. One possible future use of this database could be in the greater AWI-Gen study. The database will need to be expanded to include many more individuals from multiple project sites. The individuals will also be genotyped using a different technology with many more SNPs.

The biological focus of this study was cardiovascular or cardiometabolic diseases and more specifically BP/hypertension. The burden of CVDs and hypertension was detailed in Chapter 1, with the overall message being that of high and increasing prevalence, particularly in black or African individuals. The need to better understand the aetiology of CVDs and to introduce better management and treatment options to decrease the associated morbidity and mortality is therefore essential.

Hypertension is estimated to be 30-50% heritable (Munroe et al., 2013). Therefore, a large part of understanding its aetiology is discovering the underlying genetic factors. Blood pressure and hypertension are polygenic in nature with multiple genes and variants with small effect sizes contributing to BP variation or hypertension risk. It has also been shown that epistatic effects might exist where one gene(s) interacts with another gene(s) to exert its effect (Cicila et al., 2009; Norton et al., 2010; Wei et al., 2015; Scurrah et al., 2017). As the variants and loci identified so far explain less than 2.5% of the phenotypic variance for SBP and DBP (Ehret et al., 2011), researchers are a long way from elucidating the entire genetic architecture of BP/hypertension.

Most of the studies into the genetics of BP/hypertension have been performed in non-African or African-American individuals and more needs to still be discovered about the genetics of BP/hypertension in native Africans, who remain understudied. African populations are genetically more diverse than non-African populations and tend to have a greater SNP density and lower LD between SNPs (Remm & Metspalu, 2002; Tishkoff & Verrelli, 2003). Studies have also suggested that there is great genetic diversity between subgroups within the African population itself (Tishkoff et al., 2009). Therefore conducting genetic studies in Africans, and more specifically different groups within Africa itself, is necessary. The burden of hypertension in SSA is also of increasing concern (Ogah & Rayner,

2013) and as with the rest of Africa, little is known about its genetics in SSA individuals.

The second part of the study, therefore, was to identify genetic markers for SBP and DBP in black South African individuals and to record the findings in the developed database. The available cardiometabolic-related genotype data was harmonised with the rich phenotype data in the database, which is not disease specific, to enable easier investigation into disease-related genetics. The investigation mainly looked at SBP and DBP, two common measurements of hypertension, with an additional investigation into high versus normal/low BP and possible sex- and age-differences. The current study was exploratory in nature and used a set of SNPs with a previous association with cardiometabolic-related traits or diseases. This allowed us to relax the significance threshold slightly and potentially carry out a replication of previously identified variants/loci in our African population. An additive model was used for this analysis, although future analyses could explore using a genotypic, dominant or recessive model. None of the variants or genes on the Metabochip chosen based on previous associations with BP/hypertension in Europeans were replicated in this study. Instead, novel associations between variants in other cardiometabolic loci and SBP and/or DBP were found. The analysis pointed to regions of interest in the *NOS1AP* (DBP and SBP), *MYRF* (SBP) and *POC1B* (SBP) genes as well as two intergenic regions (*DACH1*|*LOC440145* (DBP and SBP) and *INTS10*|*LPL* (SBP)). Two SNPs in the *MYRF* gene met the calculated "array-wide" significance threshold ($p<6.7x10^{-7}$ for the merged dataset) for multiple testing. Of all the regions of interest identified, *NOS1AP* could have the most plausible functional link to blood pressure regulation or hypertension risk. A lack of replication of previously identified variants/loci in this study and the identification of different or novel variants/loci could be due to the lack of power to detect previous associations with the available sample size, the different patterns of LD and allele frequencies in different populations or differences in environmental exposures. It

could also simply reflect a lack of association in non-Africans of the variants/loci identified here and lack of association in Africans of the previously identified variants/loci (Fox et al., 2011; Ehret et al., 2016). It is important to carry out genetic studies in populations of different ancestry to discover new genetic mechanisms underlying a phenotype (Zhao et al., 2013).

In general, it is often difficult to relate identified SNPs to actual causal genes. Many of the identified variants in our and other studies are in non-coding regions and finding the exact causal variant and mechanism is often a challenge (Padmanabhan et al., 2015). Therefore, a post-study or post-GWAS analysis is usually required to test the effects of identified variants/loci on BP regulation or hypertension risk. Prioritisation of potential causal variants could narrow the list down to possible functional variants which can be further analysed by integration analysis to identify affected pathways. Further in vitro and in vivo experiments can also help to identify possible disease mechanisms. (Wang et al., 2011)

Genotype imputation is a useful addition to genetic association studies to increase the SNP panel for association testing and to help identify SNPs or genes/loci that may otherwise be missed. Imputation accuracy varies among populations of different ethnicities and is affected by several factors including the size and composition of the reference panel used, study sample size and SNP density of the region to be imputed, LD and MAF of the SNPs to be imputed [*See Chapter 1*]. As Africans are generally more challenging to impute (Howie et al., 2011) and no ideal matching reference panel for black South Africans exists, it was unclear from the onset whether or not imputation would be successful in our sample. The third aim of this research was therefore to investigate the effectiveness of genotype imputation in this black South African dataset and possibly provide a more detailed view of identified association signals. The investigation looked at regions of interest identified in the association analysis.

Imputation was achieved with high confidence in all genes, but a more detailed view or improvement in the resolution of the region was only seen in *NOS1AP* (DBP and SBP in both the merged and female caregiver datasets) and *POC1B* (Bt20 participant dataset only). The findings of the investigation offer encouraging prospects for using imputation in future studies using the same or other South African data, with a mixed population reference panel being a good reference panel choice.

Future studies in the field of disease genetics need to move beyond simple association studies or GWAS where mainly common variants are identified. Rare variants are a possible source of much of the missing heritability of many traits and need to be investigated in more detail (Jones et al., 2012). Genotype imputation is useful in this situation. In addition, as the environment and combinations of genes may play a significant role in modifying BP and other traits, gene-environment and gene-gene studies are an important future area to explore (Kidambi et al., 2012). Some of the variance in BP regulation, and probably other traits, is also likely due to epigenetic factors. Future studies could include epigenetic approaches looking at micro-ribonucleic acids (miRNAs), histone modifications and methylation (Wang et al., 2011). Longitudinal studies may also be key in predicting hypertension risk (Zhao et al., 2013). In addition, one could explore risk score analyses and SNP-set level testing using, for example, SNP-set (Sequence) Kernel Association Test (SKAT) (Ionita-Laza et al., 2013).

A better understanding of the genetics and general aetiology of BP/hypertension is important. Identifying the exact genes and pathways involved in BP regulation may highlight new ways to reduce BP and CVD risk (Ganesh et al., 2013). In addition it will highlight possible new prevention strategies (Padmanabhan et al., 2010) or targets for existing anti-hypertensive drugs or preclinical compounds (Tragante et al., 2014) and will allow for personalised prevention and treatment

(Zheng et al., 2015b). This is, however, a challenge in developing countries, such as those in Africa, where lack of funds, infrastructure or experience often prevent adequate diagnosis and treatment of hypertension (Opie & Seedat, 2005). A public health intervention is vital to improve access to the care and treatment needed for each individual (Ogah & Rayner, 2013). Until this happens, there needs to be an increased awareness of the implications of hypertension and other CVD risk factors and low-cost alternatives need to be implemented, such as changes in lifestyle including a decrease in dietary salt and increase in potassium intake, an increase in exercise, and a decrease in obesity and smoking (Opie & Seedat, 2005).

The main limitations of this study are linked to a small sample size available for the genetic association analysis, despite some positive novel associations being found. As the variants associated with BP/hypertension have small effect sizes, a significantly larger sample is needed to identify more associated variants, including those that are rare. With our available sample, we were only powered to detect variants with higher effect sizes, with smaller effect sizes only detectable at higher MAFs. We were also under-powered to carry out accurate sex- and age-stratified analyses as the sample consists of only younger males and more older females compared to younger females. An accurate investigation into the genetics of hypertension as a binary trait was also not possible. Although we could classify individuals into those with high BP and those with normal/low BP, the number of individuals with normal/low BP outnumbered the number of individuals with high blood pressure making the case-control study biased. Another limitation of the study is in the measurement of SBP and DBP. An automated machine was used to take the readings, which eliminates some of the problems arising from manual measurement of BP (Kaczorowski et al., 2012). The readings were, however, still taken at one particular sitting and the state that the patient was in and the conditions under which the readings were taken could still have influenced the quality and accuracy of the readings. It has also been

suggested that SBP and DBP are not the best measures to use to better understand the genetic architecture of BP and hypertension and that newer phenotype parameters that can accurately reflect underlying mechanisms or subphenotypes may be valuable to study. Pulse wave velocity and central BP are two examples that are potentially better markers of hypertension risk, but the progress in using these in large scale studies has been slow or limited (Padmanabhan et al., 2015).

The use of the Metabochip in this study also has some limitations including the questionable efficacy of using a genotyping tool developed from data on European populations in African populations. A large number of SNPs were in fact removed during QC for being monomorphic in the black South Africans. The Metabochip is also a relatively "old" tool having been developed in 2009. Several new BP/hypertension associated variants and regions have been identified since then, therefore further limiting the capacity to replicate in our population what has previously been found. The Metabochip was, however, one of the most cost-effective options at the time of commencement of the broader project and its contents are specific to cardiometabolic traits, making it a good starting point for the investigation in our black South African sample. The contents of the chip may, however, be a limitation in itself as the chances of identifying novel associations in Africans is reduced. Low coverage sequencing may be a possible alternative for future investigations, but this can also be costly.

The Human Heredity and Health in Africa (H3Africa) initiative and its latest developments provide a promising and exciting future for genetic studies in Africa. An Illumina African-specific genotyping chip containing 2.5 million variants specific to African populations has been designed under the H3Africa initiative and is expected to be available for use later in 2017 (https://www.illumina.com). This chip is the first to specifically target African populations, with most existing tools, such as the Metabochip, being of European origin and therefore most

suitable for use in European populations. This advancement will hopefully improve genomic and epidemiological research in Africa and ultimately improve the health and well-being of people in Africa. This is in line with the main goal of H3Africa which has been to bring infrastructure and genetic studies to Africa and to help in developing capacity and networks among researchers in Africa (http://h3africa.org). Understanding the underlying genetics of disease in Africans is critical to better understand the high burden of many diseases in Africa, but can also provide clues to the mechanisms underlying diseases across the globe.

A follow-on and replication of this current study can therefore include a more powered and accurate investigation into BP genetics in Africans using the new African chip and a larger sample from the H3Africa consortium. Replication studies are important in any genetic study to confirm identified associations. The chip may also be useful in future imputation studies and can be tested for its efficiency as a reference panel compared to the mixed population reference panel used in this study.

In summary, the current research has allowed for the development of MetaboBTT, a useful database, with accompanying interface, for the storage and querying of project-specific data currently being used in an investigation into identifying risk factors for cardiometabolic disease in South Africans. The database is easy to use and provides efficient access to the stored data. Some of this data was used to investigate the genetics of BP/hypertension in these individuals. The analysis and genotype imputation, which proved to be fairly successful in this dataset, pointed to several regions of interest and provided some insight into the genetics of blood pressure and hypertension in black South Africans. Further studies in larger samples and using a more African-specific genotyping tool are, however, required to confirm the identified associations and whether or not the genetic links are African-specific, followed by functional

studies to determine the role of the genes identified in blood pressure regulation.

REFERENCES

Abegunde, D.O., Mathers, C.D., Adam, T., Ortegon, M. & Strong, K. (2007). The burden and costs of chronic diseases in low-income and middle-income countries. *Lancet*. 370 (9603). 1929–1938.

Adams, J. (2008). DNA Sequencing Technologies. *Nature Education*. 1 (1). 193.

Addo, J., Smeeth, L. & Leon, D.A. (2007). Hypertension in sub-Saharan Africa: A systematic review. *Hypertension*. 50 (6). 1012–1018.

Adeyemo, A., Gerry, N., Chen, G., Herbert, A., Doumatey, A., Huang, H., Zhou, J., Lashley, K., Chen, Y., Christman, M., et al. (2009). A genome-wide association study of hypertension and blood pressure in African Americans. *PLoS Genetics*. 5 (7). e1000564.

Anderson, C.A., Pettersson, F.H., Barrett, J.C., Zhuang, J.J., Ragoussis, J., Cardon, L.R. & Morris, A.P. (2008). Evaluating the effects of imputation on the power, coverage, and cost efficiency of genome-wide SNP platforms. *American Journal of Human Genetics*. 83 (1). 112–119.

Anderson, C.A., Pettersson, F.H., Clarke, G.M., Cardon, L.R., Morris, A.P. & Zondervan, K.T. (2010). Data quality control in genetic case-control association studies. *Nature Protocols*. 5 (9). 1564–1573.

Anderson, N.R., Lee, E.S., Brockenbrough, J.S., Minie, M.E., Fuller, S., Brinkley, J. & Tarczy-Hornoch, P. (2007). Issues in biomedical research data management and analysis: needs and barriers. *Journal of the American Medical Informatics Association*. 14 (4). 478–488.

Aviv, A., Hollenberg, N.K. & Weder, A. (2004). Urinary potassium excretion and sodium sensitivity in blacks. *Hypertension*. 43 (4). 707–713.

Baker, E.H., Dong, Y.B., Sagnella, G.A., Rothwell, M., Onipinla, A.K., Markandu, N.D., Cappuccio, F.P., Cook, D.G., Persu, A., Corvol, P., et al. (1998). Association of hypertension with T594M mutation in beta subunit of epithelial sodium channels in black people resident in London. *Lancet*. 351 (9113). 1388–1392.

de Bakker, P.I.W., Ferreira, M.A.R., Jia, X., Neale, B.M., Raychaudhuri, S. & Voight, B.F. (2008). Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Human Molecular Genetics*. 17 (R2). R122-128.

Band, G., Le, Q.S., Jostins, L., Pirinen, M., Kivinen, K., Jallow, M., Sisay-Joof, F., Bojang, K., Pinder, M., Sirugo, G., et al. (2013). Imputation-based meta-

analysis of severe malaria in three African populations. *PLoS Genetics*. 9 (5). e1003509.

Banerjee, S. (2013). Hypertension in children. *Clinical Queries: Nephrology*. 2 (2). 78–83.

Barlassina, C., Norton, G.R., Samani, N.J., Woodwiss, A.J., Candy, G.C., Radevski, I., Citterio, L., Bianchi, G. & Cusi, D. (2000). Alpha-adducin polymorphism in hypertensives of South African ancestry. *American Journal of Hypertension*. 13 (6 Pt 1). 719–723.

Bärnighausen, T., Welz, T., Hosegood, V., Bätzing-Feigenbaum, J., Tanser, F., Herbst, K., Hill, C. & Newell, M.-L. (2007). Hiding in the shadows of the HIV epidemic: obesity and hypertension in a rural population with very high HIV prevalence in South Africa. *Journal of Human Hypertension*. 22 (3). 236–239.

Biernacka, J.M., Tang, R., Li, J., McDonnell, S.K., Rabe, K.G., Sinnwell, J.P., Rider, D.N., de Andrade, M., Goode, E.L. & Fridley, B.L. (2009). Assessment of genotype imputation methods. *BMC Proceedings*. 3 (Suppl 7). S5.

Boutin-Foster, C., Ogedegbe, G., Ravenell, J.E., Robbins, L. & Charlson, M.E. (2007). Ascribing meaning to hypertension: a qualitative study among African Americans with uncontrolled hypertension. *Ethnicity & Disease*. 17 (1). 29–34.

Browning, S.R. (2008). Missing data imputation and haplotype phase inference for genome-wide association studies. *Human Genetics*. 124 (5). 439–450.

Browning, S.R. & Browning, B.L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *American Journal of Human Genetics*. 81 (5). 1084–1097.

Buermans, H.P. & den Dunnen, J.T. (2014). Next generation sequencing technology: Advances and applications. *Biochimica et Biophysica Acta*. 1842 (10). 1932–1941.

Candy, G., Samani, N., Norton, G., Woodiwiss, A., Radevski, I., Wheatley, A., Cockcroft, J. & Hall, I.P. (2000). Association analysis of beta2 adrenoceptor polymorphisms with hypertension in a Black African population. *Journal of Hypertension*. 18 (2). 167–172.

Cardon, L.R. & Palmer, L.J. (2003). Population stratification and spurious allelic association. *Lancet*. 361 (9357). 598–604.

Chalmers, J., MacMahon, S., Mancia, G., Whitworth, J., Beilin, L., Hansson, L., Neal, B., Rodgers, A., Ni Mhurchu, C. & Clark, T. (1999). 1999 World Health

Organization-International Society of Hypertension Guidelines for the management of hypertension. Guidelines sub-committee of the World Health Organization. *Clinical and Experimental Hypertension*. 21 (5–6). 1009–1060.

Chanda, P., Yuhki, N., Li, M., Bader, J.S., Hartz, A., Boerwinkle, E., Kao, W.H. & Arking, D.E. (2012). Comprehensive evaluation of imputation performance in African Americans. *Journal of Human Genetics*. 57 (7). 411–421.

Chang, Y.P., Liu, X., Kim, J.D., Ikeda, M.A., Layton, M.R., Weder, A.B., Cooper, R.S., Kardia, S.L., Rao, D.C., Hunt, S.C., et al. (2007). Multiple genes for essential-hypertension susceptibility on chromosome 1q. *American Journal of Human Genetics*. 80 (2). 253–264.

Cicila, G., Morgan, E., Lee, S., Farms, P., Yerga-Woolwine, S., Toland, E., Ramdath, R., Gopalakrishnan, K., Bohman, K., Nestor-Kalinoski, A., et al. (2009). Epistatic genetic determinants of blood pressure and mortality in a salt-sensitive hypertension model. *Hypertension*. 53 (4). 725–732.

Cooper, R. & Rotimi, C. (1997). Hypertension in blacks. *American Journal of Hypertension*. 10 (7 Pt 1). 804–812.

Cooper, R. & Rotimi, C. (1994). Hypertension in populations of West African origin: is there a genetic predisposition? *Journal of Hypertension*. 12 (3). 215–227.

Cooper, R.S., Forrester, T.E., Plange-Rhule, J., Bovet, P., Lambert, E. V, Dugas, L.R., Cargill, K.E., Durazo-Arvizu, R.A., Shoham, D.A., Tong, L., et al. (2015). Elevated hypertension risk for African-origin populations in biracial societies: modeling the Epidemiologic Transition Study. *Journal of Hypertension*. 33 (3). 473–481.

Council on High Blood Pressure Research (2003). Lifestyle factors and blood pressure. In: J. L. Izzo & H. R. Black (eds.). *Hypertension Primer*. 274–296.

Delaneau, O., Marchini, J. & Zagury, J.-F. (2012). A linear complexity phasing method for thousands of genomes. *Nature Methods*. 9 (2). 179–181.

Dickson, S.P., Wang, K., Krantz, I., Hakonarson, H. & Goldstein, D.B. (2010). Rare variants create synthetic genome-wide associations. *PLoS Biology*. 8 (1). e1000294.

Doris, P.A. (2002). Hypertension genetics, single nucleotide polymorphisms, and the common disease:common variant hypothesis. *Hypertension*. 39 (2 Pt 2). 323–331.

Douglas, J.G., Bakris, G.L., Epstein, M., Ferdinand, K.C., Ferrario, C., Flack, J.M.,

Jamerson, K.A., Jones, W.E., Haywood, J., Maxey, R., et al. (2003). Management of high blood pressure in African Americans: consensus statement of the Hypertension in African Americans Working Group of the International Society on Hypertension in Blacks. *Archives of Internal Medicine*. 163 (5). 525–541.

Egan, B.M., Zhao, Y. & Axon, R.N. (2010). US trends in prevalence, awareness, treatment, and control of hypertension, 1988-2008. *JAMA : The Journal of the American Medical Association*. 303 (20). 2043–2050.

Ehret, G.B., Ferreira, T., Chasman, D.I., Jackson, A.U., Schmidt, E.M., Johnson, T., Thorleifsson, G., Luan, J., Donnelly, L.A., Kanoni, S., et al. (2016). The genetics of blood pressure regulation and its target organs from association studies in 342,415 individuals. *Nature Genetics*. 48 (10). 1171–1184.

Ehret, G.B., Munroe, P.B., Rice, K.M., Bochud, M., Johnson, A.D., Chasman, D.I., Smith, A. V, Tobin, M.D., Verwoert, G.C., Hwang, S.J., et al. (2011). Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature*. 478 (7367). 103–109.

Ellinghaus, D., Schreiber, S., Franke, A., Nothnagel, M., Marchini, J., Howie, B., Myers, S., McVean, G., Servin, B., Stephens, M., et al. (2009). Current software for genotype imputation. *Human Genomics*. 3 (4). 371–380.

Faruque, M.U., Chen, G., Doumatey, A., Huang, H., Zhou, J., Dunston, G.M., Rotimi, C.N. & Adeyemo, A.A. (2011). Association of ATP1B1, RGS5 and SELE polymorphisms with hypertension and blood pressure in African-Americans. *Journal of Hypertension*. 29 (10). 1906–1912.

Forrester, T. (2004). Historic and early life origins of hypertension in Africans. *Journal of Nutrition*. 134 (1). 211–216.

Fox, E.R., Young, J.H., Li, Y., Dreisbach, A.W., Keating, B.J., Musani, S.K., Liu, K., Morrison, A.C., Ganesh, S., Kutlar, A., et al. (2011). Association of genetic variation with systolic and diastolic blood pressure among African Americans: the Candidate Gene Association Resource study. *Human Molecular Genetics*. 20 (11). 2273–2284.

Franceschini, N., Carty, C.L., Lu, Y., Tao, R., Sung, Y.J., Manichaikul, A., Haessler, J., Fornage, M., Schwander, K., Zubair, N., et al. (2016). Variant discovery and fine mapping of genetic loci associated with blood pressure traits in Hispanics and African Americans. *PLoS ONE*. 11 (10). e0164132.

Franceschini, N., Fox, E., Zhang, Z., Edwards, T.L., Nalls, M.A., Sung, Y.J., Tayo, B.O., Sun, Y. V, Gottesman, O., Adeyemo, A., et al. (2013). Genome-wide association analysis of blood-pressure traits in African-ancestry individuals reveals common associated genes in African and non-African populations.

American Journal of Human Genetics. 93 (3). 545–554.

Frazer, K., Murray, S., Schork, N. & Topol, E. (2009). Human genetic variation and its contribution to complex traits. *Nature Reviews Genetics*. 10 (4). 241–251.

Fuchsberger, C., Abecasis, G.R. & Hinds, D.A. (2015). minimac2: faster genotype imputation. *Bioinformatics*. 31 (5). 782–784.

Ganesh, S.K., Tragante, V., Guo, W., Guo, Y., Lanktree, M.B., Smith, E.N., Johnson, T., Castillo, B.A., Barnard, J., Baumert, J., et al. (2013). Loci influencing blood pressure identified using a cardiovascular gene-centric array. *Human Molecular Genetics*. 22 (8). 1663–1678.

Gómez-Olivé, F.X., Ali, S.A., Made, F., Kyobutungi, C., Nonterah, E., Micklesfield, L., Alberts, M., Boua, R., Hazelhurst, S., Debpuur, C., et al. (2016). Stark regional and sex differences in the prevalence and awareness of hypertension across six sites in sub-Saharan Africa: an H3Africa AWI-Gen study. *Under review*.

Gordon, C., Bachrach, L., Carpenter, T., Crabtree, N., El-Hajj Fuleihan, G., Kutilek, S., Lorenc, R., Tosi, L., Ward, K., Ward, L., et al. (2008). Dual energy X-ray absorptiometry interpretation and reporting in children and adolescents: the 2007 ISCD Pediatric Official Positions. *Journal of Clinical Densitometry*. 11 (1). 43–58.

Grim, C.E. & Robinson, M. (1996). Blood pressure variation in blacks: genetic factors. *Seminars in Nephrology*. 16 (2). 83–93.

Gross, M. (2011). Riding the wave of biological data. *Current Biology*. 21 (6). R204-206.

Guwatudde, D., Nankya-Mutyoba, J., Kalyesubula, R., Laurence, C., Adebamowo, C., Ajayi, I., Bajunirwe, F., Njelekela, M., Chiwanga, F.S., Reid, T., et al. (2015). The burden of hypertension in sub-Saharan Africa: a four-country cross sectional study. *BMC Public Health*. 15 (1211). 1–8.

Hancock, D.B., Levy, J.L., Gaddis, N.C., Bierut, L.J., Saccone, N.L., Page, G.P. & Johnson, E.O. (2012). Assessment of genotype imputation performance using 1000 Genomes in African American studies. *PLoS ONE*. 7 (11). e50610.

Hao, K., Chudin, E., McElwee, J. & Schadt, E.E. (2009). Accuracy of genome-wide imputation of untyped markers and impacts on statistical power for association studies. *BMC Genetics*. 10 (27).

Helmy, M., Crits-Christoph, A. & Bader, G. (2016). Ten Simple Rules for Developing Public Biological Databases. *PLoS Computational Biology*. 12 (11). e1005128.

Hirschman, L., Burns, G.A., Krallinger, M., Arighi, C., Cohen, K.B., Valencia, A., Wu, C.H., Chatr-Aryamontri, A., Dowell, K.G., Huala, E., et al. (2012). Text mining for the biocuration workflow. *Database (Oxford)*. 2012. bas020.

Hoffmann, T.J. & Witte, J.S. (2015). Strategies for imputing and analyzing rare variants in association studies. *Trends in Genetics*. 31 (10). 556–563.

Hollier, J., Martin, D., Bell, D., Li, J., Chirachanchai, M., Menon, D., Leonard, D., Wu, X., Cooper, R., McKenzie, C., et al. (2006). Epithelial sodium channel allele T594M is not associated with blood pressure or blood pressure response to amiloride. *Hypertension*. 47 (3). 428–433.

Howe, D., Costanzo, M., Fey, P., Gojobori, T., Hannick, L., Hide, W., Hill, D.P., Kania, R., Schaeffer, M., St Pierre, S., et al. (2008). Big data: The future of biocuration. *Nature*. 455 (7209). 47–50.

Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G.R. (2012). Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nature Genetics*. 44 (8). 955–959.

Howie, B., Marchini, J. & Stephens, M. (2011). Genotype imputation with thousands of genomes. *G3: Genes, Genomes, Genetics*. 1 (6). 457–470.

Howie, B.N., Donnelly, P. & Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genetics*. 5 (6). e1000529.

Huang, G.H. & Tseng, Y.C. (2014). Genotype imputation accuracy with different reference panels in admixed populations. *BMC Proceedings*. 8 (Suppl 1). S64.

Huang, L., Jakobsson, M., Pemberton, T.J., Ibrahim, M., Nyambo, T., Omar, S., Pritchard, J.K., Tishkoff, S.A. & Rosenberg, N.A. (2011). Haplotype variation and genotype imputation in African populations. *Genetic Epidemiology*. 35 (8). 766–780.

Huang, L., Li, Y., Singleton, A.B., Hardy, J.A., Abecasis, G., Rosenberg, N.A. & Scheet, P. (2009). Genotype-imputation accuracy across worldwide human populations. *American Journal of Human Genetics*. 84 (2). 235–250.

Ionita-Laza, I., Lee, S., Makarov, V., Buxbaum, J.D. & Lin, X. (2013). Sequence kernel association tests for the combined effect of rare and common variants. *American Journal of Human Genetics*. 92 (6). 841–853.

Johnson, E.O., Hancock, D.B., Levy, J.L., Gaddis, N.C., Saccone, N.L., Bierut, L.J. & Page, G.P. (2013). Imputation across genotyping arrays for genome-wide association studies: Assessment of bias and a correction strategy. *Human*

Genetics. 132 (5). 509–522.

Johnson, T., Gaunt, T.R., Newhouse, S.J., Padmanabhan, S., Tomaszewski, M., Kumari, M., Morris, R.W., Tzoulaki, I., O'Brien, E.T., Poulter, N.R., et al. (2011). Blood pressure loci identified with a gene-centric array. *American Journal of Human Genetics*. 89 (6). 688–700.

Jones, E.S., Owen, E.P. & Rayner, B.L. (2012). The association of the R563Q genotype of the ENaC with phenotypic variation in Southern Africa. *American Journal of Hypertension*. 25 (12). 1286–1291.

Kaczorowski, J., Dawes, M. & Gelfer, M. (2012). Measurement of blood pressure: New developments and challenges. *British Columbia Medical Journal*. 54 (8). 399–403.

Kagura, J., Adair, L.S., Musa, M.G., Pettifor, J.M. & Norris, S.A. (2015). Blood pressure tracking in urban black South African children: birth to twenty cohort. *BMC Pediatrics*. 15 (78). 1–7.

Kato, N., Loh, M., Takeuchi, F., Verweij, N., Wang, X., Zhang, W., Kelly, T.N., Saleheen, D., Lehne, B., Mateo Leach, I., et al. (2015). Trans-ancestry genome-wide association study identifies 12 genetic loci influencing blood pressure and implicates a role for DNA methylation. *Nature Genetics*. 47 (11). 1282–1293.

Kato, N., Takeuchi, F., Tabara, Y., Kelly, T.N., Go, M.J., Sim, X., Tay, W.T., Chen, C.H., Zhang, Y., Yamamoto, K., et al. (2011). Meta-analysis of genome-wide association studies identifies common variants associated with blood pressure variation in east Asians. *Nature Genetics*. 43 (6). 531–538.

Kidambi, S., Ghosh, S., Kotchen, J.M., Grim, C.E., Krishnaswami, S., Kaldunski, M.L., Cowley, A.W., Patel, S.B. & Kotchen, T.A. (2012). Non-replication study of a genome-wide association study for hypertension and blood pressure in African Americans. *BMC Medical Genetics*. 13 (27). 1–8.

Larranaga, P., Calvo, B., Santana, R., Bielza, C., Galdiano, J., Inza, I., Lozano, J.A., Armananzas, R., Santafe, G., Perez, A., et al. (2006). Machine learning in bioinformatics. *Briefings in Bioinformatics*. 7 (1). 86–112.

Levy, D., Ehret, G.B., Rice, K., Verwoert, G.C., Launer, L.J., Dehghan, A., Glazer, N.L., Morrison, A.C., Johnson, A.D., Aspelund, T., et al. (2009). Genome-wide association study of blood pressure and hypertension. *Nature Genetics*. 41 (6). 677–687.

Lewis, C.M. & Knight, J. (2012). Introduction to genetic association studies. *Cold Spring Harbor Protocols*. 2012 (3). 297–306.

Li, J., Guo, Y., Pei, Y. & Deng, H.-W. (2012). The impact of imputation on meta-analysis of genome-wide association studies. *PloS ONE*. 7 (4). e34486.

Li, Y., Willer, C., Sanna, S. & Abecasis, G. (2009). Genotype imputation. *Annual Review of of Genomics and Human Genetics*. 10. 387–406.

Li, Y., Willer, C.J., Ding, J., Scheet, P. & Abecasis, G.R. (2010). MaCH: Using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic Epidemiology*. 34 (8). 816–834.

Lifton, R. (2004). Genetic dissection of human blood pressure variation: common pathways from rare phenotypes. *Harvey Lectures*. 100. 71–101.

Liu, E.Y., Buyske, S., Aragaki, A.K., Peters, U., Boerwinkle, E., Carlson, C., Carty, C., Crawford, D.C., Haessler, J., Hindorff, L.A., et al. (2012). Genotype imputation of Metabochip SNPs using a study-specific reference panel of ~4,000 haplotypes in African Americans from the Women's Health Initiative. *Genetic Epidemiology*. 36 (2). 107–117.

Liu, Q., Cirulli, E.T., Han, Y., Yao, S., Liu, S. & Zhu, Q. (2014). Systematic assessment of imputation performance using the 1000 Genomes reference panels. *Briefings in Bioinformatics*. 16 (4). 549–562.

Lombard, Z., Crowther, N.J., Van Der Merwe, L., Pitamber, P., Norris, S.A. & Ramsay, M. (2012). Appetite regulation genes are associated with body mass index in black South African adolescents: A genetic association study. *BMJ Open*. 2 (3). e000873.

Lu, X., Wang, L., Lin, X., Huang, J., Charles gu, C., He, M., Shen, H., He, J., Zhu, J., Li, H., et al. (2015). Genome-wide association study in Chinese identifies novel loci for blood pressure and hypertension. *Human Molecular Genetics*. 24 (3). 865–874.

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. *Nature*. 461 (7265). 747–753.

Marchini, J. & Howie, B. (2008). Comparing algorithms for genotype imputation. *American Journal of Human Genetics*. 83 (4). 535–539.

Marchini, J., Howie, B.N., Myers, S., McVean, G. & Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genetics*. 39 (7). 906–913.

Mathers, C.D. & Loncar, D. (2006). Projections of global mortality and burden of disease from 2002 to 2030. *PLoS Medicine*. 3 (11). 2011–2030.

Merelli, I., Calabria, A., Cozzi, P., Viti, F., Mosca, E. & Milanesi, L. (2010). SNPranker 2.0: a gene-centric data mining tool for diseases associated SNP prioritization in GWAS. *BMC Bioinformatics*. 14 (Suppl 1). S9.

Miller, S.A., Dykes, D.D. & Polesky, H.F. (1988). A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Research*. 16 (3). 1215.

Munroe, P.B., Barnes, M.R. & Caulfield, M.J. (2013). Advances in blood pressure genomics. *Circulation Research*. 112 (10). 1365–1379.

Newton-Cheh, C., Johnson, T., Gateva, V., Tobin, M.D., Bochud, M., Coin, L., Najjar, S.S., Zhao, J.H., Heath, S.C., Eyheramendy, S., et al. (2009a). Genome-wide association study identifies eight loci associated with blood pressure. *Nature Genetics*. 41 (6). 666–676.

Newton-Cheh, C., Larson, M.G., Vasan, R.S., Levy, D., Bloch, K.D., Surti, A., Guiducci, C., Kathiresan, S., Benjamin, E.J., Struck, J., et al. (2009b). Association of common variants in NPPA and NPPB with circulating natriuretic peptides and blood pressure. *Nature Genetics*. 41 (3). 348–353.

Nho, K., Shen, L., Kim, S., Swaminathan, S., Risacher, S.L., Saykin, A.J. & Alzheimer's Disease Neuroimaging, I. (2011). The effect of reference panels and software tools on genotype imputation. *AMIA Annual Symposium Proceedings*. 2011. 1013–1018.

Nissinen, A., Böthig, S., Granroth, H. & Lopez, A. (1988). Hypertension in developing countries. *World Health Statistics Quarterly*. 41 (3–4). 141–154.

Nkeh, B., Samani, N.J., Badenhorst, D., Libhaber, E., Sareli, P., Norton, G.R. & Woodiwiss, A.J. (2003). T594M variant of the epithelial sodium channel beta-subunit gene and hypertension in individuals of African ancestry in South Africa. *American Journal of Hypertension*. 16 (10). 847–852.

Nkeh, B., Tiago, A., Candy, G.P., Woodiwiss, A.J., Badenhorst, D., Luker, F., Netjhardt, M., Brooksbank, R., Libhaber, C., Sareli, P., et al. (2002). Association between an atrial natriuretic peptide gene polymorphism and normal blood pressure in subjects of African ancestry. *Cardiovascular Journal of South Africa*. 13 (3). 97–101.

Norton, G.R., Brooksbank, R. & Woodiwiss, A.J. (2010). Gene variants of the renin-angiotensin system and hypertension: from a trough of disillusionment to a welcome phase of enlightenment? *Clinical Science*. 118 (8). 487–506.

Ogah, O.S. & Rayner, B.L. (2013). Recent advances in hypertension in sub-Saharan Africa. *Heart*. 99 (19). 1390–1397.

Opie, L.H. & Seedat, Y.K. (2005). Hypertension in sub-Saharan African populations. *Circulation*. 112 (23). 3562–3568.

Org, E., Eyheramendy, S., Juhanson, P., Gieger, C., Lichtner, P., Klopp, N., Veldre, G., Doring, A., Viigimaa, M., Sober, S., et al. (2009). Genome-wide scan identifies CDH13 as a novel susceptibility locus contributing to blood pressure determination in two European populations. *Human Molecular Genetics*. 18 (12). 2288–2296.

Padmanabhan, S., Caulfield, M. & Dominiczak, A.F. (2015). Genetic and molecular aspects of hypertension. *Circulation Research*. 116 (6). 937–959.

Padmanabhan, S., Melander, O., Johnson, T., Di Blasio, A.M., Lee, W.K., Gentilini, D., Hastie, C.E., Menni, C., Monti, M.C., Delles, C., et al. (2010). Genome-wide association study of blood pressure extremes identifies variant near UMOD associated with hypertension. *PLoS Genetics*. 6 (10). 1–11.

Parmar, P., Taal, H., Timpson, N., Thiering, E., Lehtimäki, T., Marinelli, M., Lind, P., Howe, L., Verwoert, G., Aalto, V., et al. (2016). International Genome-Wide Association Study Consortium Identifies Novel Loci Associated With Blood Pressure in Children and Adolescents. *Circulation: Cardiovascular Genetics*. 9 (3). 266–278.

Pasaniuc, B., Avinery, R., Gur, T., Skibola, C.F., Bracci, P.M. & Halperin, E. (2010). A generic coalescent-based framework for the selection of a reference panel for imputation. *Genetic Epidemiology*. 34 (8). 773–782.

Patterson, N., Price, A.L. & Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genetics*. 2 (12). e190.

Pei, Y.F., Li, J., Zhang, L., Papasian, C.J. & Deng, H.W. (2008). Analyses and comparison of accuracy of different genotype imputation methods. *PLoS ONE*. 3 (10). e3551.

Pei, Y.F., Zhang, L., Li, J. & Deng, H.W. (2010). Analyses and comparison of imputation-based association methods. *PLoS ONE*. 5 (5). e10827.

Persu, A., Evenepoel, L., Jin, Y., Mendola, A., Ngueta, G., Yang, W., Gruson, D., Horman, S., Staessen, J. & Vikkula, M. (2016). STK39 and WNK1 Are Potential Hypertension Susceptibility Genes in the BELHYPGEN Cohort. *Medicine*. 95 (15). e2968.

Pillay, V., Crowther, N., Ramsay, M., Smith, G., Norris, S. & Lombard, Z. (2015). Exploring genetic markers of adult obesity risk in black adolescent South Africans-the Birth to Twenty Cohort. *Nutrition and Diabetes*. 5. e157.

Purcell, S. & Chang, C. (2014). *PLINK 1.9*. [Online]. 2014. Available from:

https://www.cog-genomics.org/plink2.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*. 81 (3). 559–575.

Ramsay, M., Crowther, N., Tambo, E., Agongo, G., Baloyi, V., Dikotope, S., Gómez-Olivé, X., Jaff, N., Sorgho, H., Wagner, R., et al. (2016). H3Africa AWI-Gen Collaborative Centre: a resource to study the interplay between genomic and environmental risk factors for cardiometabolic diseases in four sub-Saharan African countries. *Global Health, Epidemiology and Genomics*. 1 (e20). 1–13.

Rayner, B.L., Owen, E.P., King, J. a, Soule, S.G., Vreede, H., Opie, L.H., Marais, D. & Davidson, J.S. (2003). A new mutation, R563Q, of the beta subunit of the epithelial sodium channel associated with low-renin, low-aldosterone hypertension. *Journal of Hypertension*. 21 (5). 921–926.

Rayner, B.L. & Spence, J.D. (2017). Hypertension in blacks: insights from Africa. *Journal of Hypertension*. 35 (2). 234–239.

Reckelhoff, J.F. (2001). Gender differences in the regulation of blood pressure. *Hypertension*. 37 (5). 1199–1208.

Redmond, N., Baer, H.J. & Hicks, L.S. (2011). Health behaviors and racial disparity in blood pressure control in the national health and nutrition examination survey. *Hypertension*. 57 (3). 383–389.

Remm, M. & Metspalu, A. (2002). High-density genotyping and linkage disequilibrium in the human genome using chromosome 22 as a model. *Current Opinion in Chemical Biology*. 6 (1). 24–30.

Richter, L., Norris, S., Pettifor, J., Yach, D. & Cameron, N. (2007). Cohort Profile: Mandela's children: the 1990 Birth to Twenty study in South Africa. *International Journal of Epidemiology*. 36 (3). 504–511.

Rigden, D.J., Fernández-Suárez, X.M. & Galperin, M.Y. (2016). The 2016 database issue of nucleic acids research and an updated molecular biology database collection. *Nucleic Acids Research*. 44 (D1). D1–D6.

Roger, V.L., Go, A.S., Lloyd-Jones, D.M., Benjamin, E.J., Berry, J.D., Borden, W.B., Bravata, D.M., Dai, S., Ford, E.S., Fox, C.S., et al. (2012). Heart disease and stroke statistics--2012 update: a report from the American Heart Association. *Circulation*. 125 (1). e2–e220.

Rosamond, W., Flegal, K., Friday, G., Furie, K., Go, A., Greenlund, K., Haase, N.,

Ho, M., Howard, V., Kissela, B., et al. (2007). Heart disease and stroke statistics--2007 update: a report from the American Heart Association Statistics Committee and Stroke Statistics Subcommittee. *Circulation*. 115 (5). e69-171.

Roshyara, N.R., Horn, K., Kirsten, H., Ahnert, P., Scholz, M., An, P., Leeuwen, E.M. van, Zeggini, E., Lambert, J.C., Olama, A.A. Al, et al. (2016). Comparing performance of modern genotype imputation methods in different ethnicities. *Scientific Reports*. 6 (34386). 1–12.

Roshyara, N.R., Kirsten, H., Horn, K., Ahnert, P. & Scholz, M. (2014). Impact of pre-imputation SNP-filtering on genotype imputation results. *BMC Genetics*. 15 (88). 1–11.

Roshyara, N.R. & Scholz, M. (2014). fcGENE: A versatile tool for processing and transforming SNP datasets. *PLoS ONE*. 9 (7). e97589.

Roshyara, N.R. & Scholz, M. (2015). Impact of genetic similarity on imputation accuracy. *BMC Genetics*. 16 (90). 1–16.

Sahibdeen, V. (2016). *The identification of genetic markers of obesity risk in a South African black population (Unpublished doctoral thesis)*. University of the Witwatersrand, Johannesburg, South Africa.

Salako, B.L., Ogah, O.S., Adebiyi, a a, Adedapo, K.S., Bekibele, C.O., Oluleye, T.S. & Okpechi, I. (2007). Unexpectedly high prevalence of target-organ damage in newly diagnosed Nigerians with hypertension. *Cardiovascular Journal of Africa*. 18 (2). 77–83.

Salvi, E., Kutalik, Z., Glorioso, N., Benaglio, P., Frau, F., Kuznetsova, T., Arima, H., Hoggart, C., Tichet, J., Nikitin, Y.P., et al. (2012). Genome-wide association study using a high-density SNP-array and case-control design identifies a novel essential hypertension susceptibility locus in the promoter region of eNOS. *Hypertension*. 59 (2). 248–255.

Scheet, P. & Stephens, M. (2006). A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *American Journal of Human Genetics*. 78 (4). 629–644.

Schultheiss, S.J. (2011). Ten simple Rules for Providing a Scientific Web Resource. *PLoS Computational Biology*. 7 (5). e1001126.

Scurrah, K., Lamantia, A., Ellis, J. & Harrap, S. (2017). Familial Analysis of Epistatic and Sex-Dependent Association of Genes of the Renin-Angiotensin-Aldosterone System and Blood Pressure. *Circulation: Cardiovascular Genetics*. 10 (3). e001595.

Seedat, Y.K. (1999). Hypertension in black South Africans. *Journal of Human Hypertension*. 13 (2). 96–103.

Seedat, Y.K. (1983). Race, environment and blood pressure: the South African experience. *Journal of Hypertension*. 1 (1). 7–12.

Shetty, P.B., Tang, H., Tayo, B.O., Morrison, A.C., Hanis, C.L., Rao, D.C., Young, J.H., Fox, E.R., Boerwinkle, E., Cooper, R.S., et al. (2012). Variants in CXADR and F2RL1 are associated with blood pressure and obesity in African-Americans in regions identified through admixture mapping. *Journal of Hypertension*. 30 (10). 1970–1976.

Singh, M., Singh, A., Pandey, P., Chandra, S., Singh, K. & Gambhir, I. (2016). Molecular genetics of essential hypertension. *Clinical and Experimental Hypertension*. 38 (3). 268–277.

Sliwa, K., Ojji, D., Bachelier, K., Böhm, M., Damasceno, A. & Stewart, S. (2014). Hypertension and hypertensive heart disease in African women. *Clinical Research in Cardiology*. 103 (7). 515–523.

Spencer, C.C.A., Su, Z., Donnelly, P. & Marchini, J. (2009). Designing genome-wide association studies: Sample size, power, imputation, and the choice of genotyping chip. *PLoS Genetics*. 5 (5). e1000477.

Sung, Y.J., Gu, C.C., Tiwari, H.K., Arnett, D.K., Broeckel, U. & Rao, D.C. (2012a). Genotype imputation for African Americans using data from HapMap phase II versus 1000 genomes projects. *Genetic Epidemiology*. 36 (5). 508–516.

Sung, Y.J., Wang, L., Rankinen, T., Bouchard, C. & Rao, D.C. (2012b). Performance of genotype imputations using data from the 1000 Genomes Project. *Human Heredity*. 73 (1). 18–25.

Takeuchi, F., Isono, M., Katsuya, T., Yamamoto, K., Yokota, M., Sugiyama, T., Nabika, T., Fujioka, A., Ohnaka, K., Asano, H., et al. (2010). Blood pressure and hypertension are associated with 7 loci in the Japanese population. *Circulation*. 121 (21). 2302–2309.

Taylor, J.Y., Sampson, D., Taylor, A.D., Caldwell, D. & Sun, Y. V (2011). Genetic and BMI risks for predicting blood pressure in three generations of West African Dogon women. *Biological Research for Nursing*. 15 (1). 105–111.

Tchelougou, D., Kologo, J.K., Karou, S.D., Yaméogo, V.N., Bisseye, C., Djigma, F.W., Ouermi, D., Compaoré, T.R., Assih, M., Pietra, V., et al. (2015). Renin-angiotensin system genes polymorphisms and essential hypertension in Burkina Faso, West Africa. *International Journal of Hypertension*. 2015 (979631). 1–7.

The 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. *Nature*. 526 (7571). 68–74.

The International HapMap Consortium (2005). A haplotype map of the human genome. *Nature*. 437 (7063). 1299–1320.

The International HapMap Consortium (2003). The International HapMap Project. *Nature*. 426 (6968). 789–796.

Tiago, A.D., Badenhorst, D., Nkeh, B., Candy, G.P., Brooksbank, R., Sareli, P., Libhaber, E., Samani, N.J., Woodiwiss, A.J. & Norton, G.R. (2003). Impact of renin-angiotensin-aldosterone system gene variants on the severity of hypertension in patients with newly diagnosed hypertension. *American Journal of Hypertension*. 16 (12). 1006–1010.

Tiago, A.D., Samani, N.J., Candy, G.P., Brooksbank, R., Libhaber, E.N., Sareli, P., Woodiwiss, A.J. & Norton, G.R. (2002). Angiotensinogen gene promoter region variant modifies body size-ambulatory blood pressure relations in hypertension. *Circulation*. 106 (12). 1483–1487.

Tishkoff, S.A., Reed, F.A., Friedlaender, F.R., Ehret, C., Ranciaro, A., Froment, A., Hirbo, J.B., Awomoyi, A.A., Bodo, J.M., Doumbo, O., et al. (2009). The genetic structure and history of Africans and African Americans. *Science*. 324 (5930). 1035–1044.

Tishkoff, S.A. & Verrelli, B.C. (2003). Role of evolutionary history on haplotype block structure in the human genome: implications for disease mapping. *Current Opinion in Genetics and Development*. 13 (6). 569–575.

Topaloglou, T. (2004). Biological Data Management: Research, Practice and Opportunities. In: *Thirtieth International Conference on Very Large Data Bases*. 2004, 1233–1236.

Tragante, V., Barnes, M.R., Ganesh, S.K., Lanktree, M.B., Guo, W., Franceschini, N., Smith, E.N., Johnson, T., Holmes, M. V., Padmanabhan, S., et al. (2014). Gene-centric meta-analysis in 87,736 individuals of European ancestry identifies multiple blood-pressure-related loci. *American Journal of Human Genetics*. 94 (3). 349–360.

Tu, W. & Pratt, J.H. (2013). A consideration of genetic mechanisms behind the development of hypertension in blacks. *Current Hypertension Reports*. 15 (2). 108–113.

Voight, B.F., Kang, H.M., Ding, J., Palmer, C.D., Sidore, C., Chines, P.S., Burtt, N.P., Fuchsberger, C., Li, Y., Erdmann, J., et al. (2012). The Metabochip, a Custom Genotyping Array for Genetic Studies of Metabolic, Cardiovascular, and Anthropometric Traits. *PLoS Genetics*. 8 (8). e1002793.

Wang, X., Prins, B.P., Sõber, S., Laan, M. & Snieder, H. (2011). Beyond genome-wide association studies: New strategies for identifying genetic determinants of hypertension. *Current Hypertension Reports*. 13 (6). 442–451.

Wang, Y., Cai, Z., Stothard, P., Moore, S., Goebel, R., Wang, L. & Lin, G. (2012). Fast accurate missing SNP genotype local imputation. *BMC Research Notes*. 5 (404). 1–12.

Wang, Y., O'Connell, J.R., McArdle, P.F., Wade, J.B., Dorff, S.E., Shah, S.J., Shi, X., Pan, L., Rampersaud, E., Shen, H., et al. (2009). From the Cover: Whole-genome association study identifies STK39 as a hypertension susceptibility gene. *Proceedings of the National Academy of Sciences*. 106 (1). 226–231.

Wei, L.K., Menon, S., Griffiths, L.R. & Gan, S.H. (2015). Signaling pathway genes for blood pressure, folate and cholesterol levels among hypertensives: an epistasis analysis. *Journal of Human Hypertension*. 29 (2). 99–104.

Weinberger, M., Miller, J., Luft, F., Grim, C. & Fineberg, N. (1986). Definitions and characteristics of sodium sensitivity and blood pressure resistance. *Hypertension*. 8 (6 Pt 2). II127-134.

Wood, A.R., Perry, J.R.B., Tanaka, T., Hernandez, D.G., Zheng, H.F., Melzer, D., Gibbs, J.R., Nalls, M.A., Weedon, M.N., Spector, T.D., et al. (2013). Imputation of variants from the 1000 Genomes Project modestly improves known associations and can identify low-frequency variant - phenotype associations undetected by HapMap based imputation. *PLoS ONE*. 8 (5). e64343.

World Health Organization (2014a). *Global Health Estimates : Death by cause, age, sex and country, 2000 -2012. WHO*. 2014.

World Health Organization (2014b). *Global status report on noncommunicable diseases 2014*. 2014.

Yang, Q., Zhang, Z., Kuklina, E. V, Fang, J., Ayala, C., Hong, Y., Loustalot, F., Dai, S., Gunn, J.P., Tian, N., et al. (2012). Sodium intake and blood pressure among US children and adolescents. *Pediatrics*. 130 (4). 611–619.

Zhang, B., Zhi, D., Zhang, K., Gao, G., Limdi, N.N. & Liu, N. (2011). Practical Consideration of Genotype Imputation: Sample Size, Window Size, Reference Choice, and Untyped Rate. *Statistics and Its Interface*. 4 (3). 339–352.

Zhang, H., Sun, Z.Q., Liu, S.S. & Yang, L.N. (2015). Association between GRK4 and DRD1 gene polymorphisms and hypertension: A meta-analysis. *Clinical Interventions in Aging*. 2016 (11). 17–27.

Zhao, Q., Kelly, T.N., Li, C. & He, J. (2013). Progress and future aspects in genetics of human hypertension. *Current Hypertension Reports*. 15 (6). 676–686.

Zhao, Z., Timofeev, N., Hartley, S.W., Chui, D.H., Fucharoen, S., Perls, T.T., Steinberg, M.H., Baldwin, C.T. & Sebastiani, P. (2008). Imputation of missing genotypes: an empirical evaluation of IMPUTE. *BMC Genetics*. 9 (85).

Zheng-Bradley, X. & Flicek, P. (2016). Applications of the 1000 Genomes Project resources. *Briefings in Functional Genomics*. pii (elw027). 1–8.

Zheng, H.F., Rong, J.J., Liu, M., Han, F., Zhang, X.W., Richards, J.B. & Wang, L. (2015a). Performance of genotype imputation for low frequency and rare variants from the 1000 genomes. *PLoS ONE*. 10 (1). e0116487.

Zheng, J., Rao, D. & Shi, G. (2015b). An update on genome-wide association studies of hypertension. *Applied Informatics*. 2 (10). 1–20.

Zhu, X. & Cooper, R.S. (2007). Admixture mapping provides evidence of association of the VNN1 gene with hypertension. *PLoS ONE*. 2 (11). e1244.

Zhu, X., Feng, T., Tayo, B.O., Liang, J., Young, J.H., Franceschini, N., Smith, J.A., Yanek, L.R., Sun, Y. V., Edwards, T.L., et al. (2015). Meta-analysis of correlated traits via summary statistics from GWASs with an application in hypertension. *American Journal of Human Genetics*. 96 (1). 21–36.

Zhu, X., Young, J.H., Fox, E., Keating, B.J., Franceschini, N., Kang, S., Tayo, B., Adeyemo, A., Sun, Y. V, Li, Y., et al. (2011). Combined admixture mapping and association analysis identifies a novel blood pressure genetic locus on 5p13: contributions from the CARe consortium. *Human Molecular Genetics*. 20 (11). 2285–2295.

WEB REFERENCES

Genesis software for PCA plots: http://www.bioinf.wits.ac.za/software/genesis (Robert Buchmann, 2014)

H3Africa: http://h3africa.org/

Lucidchart: https://www.lucidchart.com

Illumina. (2016). *H3Africa Consortium Array Available Soon.* [online] Available at: https://www.illumina.com/company/news-center/feature-articles/h3africa-consortium-array-available-soon-.html [Accessed 30 January 2017]

APPENDIX A – Ethics certificates, relevant agreements/ letters and consent forms

R14/49 Miss Liesl Mary Hendry

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)

CLEARANCE CERTIFICATE NO. M1411116

NAME: (Principal Investigator)	Miss Liesl Mary Hendry
DEPARTMENT:	Molecular and Cell Biology Sydney Brenner Institute for Molecular Bioscience
PROJECT TITLE:	Bioinformatics-Driven Development of a Queryable Cardiovascular Database and its Application in a Biological Setting
DATE CONSIDERED:	Adhoc
DECISION:	Approved unconditionally
CONDITIONS:	
SUPERVISOR:	Dr Zane Lombard
APPROVED BY:	Professor P Cleaton-Jones, Chairperson, HREC (Medical)
DATE OF APPROVAL:	14/01/2015

This clearance certificate is valid for 5 years from date of approval. Extension may be applied for.

DECLARATION OF INVESTIGATORS

To be completed in duplicate and **ONE COPY** returned to the Secretary in Room 10004, 10th floor, Senate House, University.
I/we fully understand the conditions under which I am/we are authorized to carry out the above-mentioned research and I/we undertake to ensure compliance with these conditions. Should any departure be contemplated, from the research protocol as approved, I/we undertake to resubmit the application to the Committee. **I agree to submit a yearly progress report**.

Principal Investigator Signature Date

PLEASE QUOTE THE PROTOCOL NUMBER IN ALL ENQUIRIES

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG
Division of the Deputy Registrar (Research)

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)
R14/49 Ms Venesa Pillay et al

CLEARANCE CERTIFICATE	**M120647**
PROJECT	Identification of Genetic Markers of Obesity Risk and Body Composition in a South African Black Population
INVESTIGATORS	Ms Venesa Pillay et al.
DEPARTMENT	School of Pathology/Div. of Human Genetics
DATE CONSIDERED	29/06/2012
DECISION OF THE COMMITTEE*	Approved unconditionally

Unless otherwise specified this ethical clearance is valid for 5 years and may be renewed upon application.

DATE 29/06/2012 **CHAIRPERSON** .....
(Professor PE Cleaton-Jones)

*Guidelines for written 'informed consent' attached where applicable
cc: Supervisor : Dr Zane Lombard

DECLARATION OF INVESTIGATOR(S)
To be completed in duplicate and **ONE COPY** returned to the Secretary at Room 10004, 10th Floor, Senate House, University.
I/We fully understand the conditions under which I am/we are authorized to carry out the abovementioned research and I/we guarantee to ensure compliance with these conditions. Should any departure to be contemplated from the research procedure as approved I/we undertake to resubmit the protocol to the Committee. **I agree to a completion of a yearly progress report.**

PLEASE QUOTE THE PROTOCOL NUMBER IN ALL ENQUIRIES...

Human Research Ethics Committee (Medical)

Research Office Secretariat: Senate House Room SH 10005, 10th floor. Tel +27 (0)11-717-1252
Medical School Secretariat: Medical School Room 10M07, 10th Floor. Tel +27 (0)11-717-2700
Private Bag 3, Wits 2050, www.wits.ac.za. Fax +27 (0)11-717-1265

25 November 2013

Ms Venesa Pillay et al
School of pathology
Division of Human Genetics
University of the Witwatersrand

Sent by email to: venesap@gmail.com

Dear Ms Pillay

Protocol no: M120647
Protocol Title: Identification of Genetic Markers of Obesity Risk and Body Composition in South African Black Population
Principal Investigator: Ms Venesa Pillay et al
Protocol Amendment

This letter serves to confirm that the Chairman of the Human Research Ethics Committee (Medical) has approved the following amendments as detailed in your letter dated 16 November 2013:

- Extension of the research to include 1260 additional samples from study no M010556

Thank you for keeping us informed and updated,

Yours Sincerely,

.....
Ms Zanele Ndlovu
Administrative Officer
Human Research Ethics Committee (Medical)

health

Department:
Health
REPUBLIC OF SOUTH AFRICA

Private Bag X828, PRETORIA, 0001
Civitas Building Corner Andries and Struben Street, PRETORIA, 0001
Tel (012) 395 0922 • Fax (012) _____

Tel (012) 395-8366/9197 Ms Lineo Motopi

Fax 086 632 6815/2606 J1/2/4/2 No ' '13

Prof Michele Ramsay
Medical Technologist
NHLS and Wits University
Division of Human Genetics
P O Box 1038
JOHANNESBURG
2000

Dear Prof Ramsay

EXPORT PERMIT

Attached please receive one export permit as requested by your fax dated 30 April 2013.

Please note that the Department is committed to processing all requests for permits as soon as possible. However, the Department cannot guarantee the issuing of a permit within a specified time. You are therefore advised that applications for permits reach the Department well in advance of shipping dates and/or requirements.

The Department will not be responsible for any losses due to applications that are not received timeously.

Kind regards

DIRECTOR-GENERAL: HEALTH
Date:
Ms P Netshidzivhani

health

Department:
Health
REPUBLIC OF SOUTH AFRICA

Private Bag X828, PRETORIA, 0001
Civitas Building Corner Andries and Struben Street, PRETORIA, 0001
Tel (012) 395 0922 • Fax (012) _____

Reference : J1/2/4/14 No1/13
Enquiry : Ms Lineo Motopi
Tel : (012) 395-8366/9197
Fax : (086) 632 6815/2306

EXPORT PERMIT

In terms of Section 68 of the National Health Act 2003 (Act No. 61 of 2003) –

Prof Michele Ramsay
Medical Scientist
NHLS and Wits University
Division of Human Genetics
P O Box 1038
JOHANNESBURG
2000
Tel. No.: (011) 489 9214 Fax. No.: (011) 489 9226

is hereby authorised to export from the Republic of South Africa –
12 x PCR plates (1034 samples) Human genomic DNA samples

to –

Dr Vanessa
DNA Technologies Core,
University of Crlifornia
Davis 4212A GBSF
475 Health Sciences Drive
Davis, CA, 95616
UNITED STATES OF AMERICA
Tel:. No: 530-754-5821 Fax. 530-754-9658

for – Research.

This export permit is subject to the following conditions:

1. The substance shall be imported into the country specified above, within the legal requirements of that country.
2. The substance shall be exported from South Africa and handled in accordance with the provisions of Section 68 of the National Health Act 2003 (Act No. 61 of 2003), and the regulations made in terms of the Act.
3. The export permit shall not be used for any trade or advertising purposes.
4. **This export permit shall expire on 31 May 2014.**

DIRECTOR-GENERAL: HEALTH
Date:
Ms P Netshidzivhani

Human Materials Transfer Agreement
Wits-INDEPTH partnership – AWI-Gen Study
NIH - H3Africa funded research

In agreement between the RECIPIENT and the PROVIDER this is a request to transfer MATERIAL:

MATERIAL: Human Genomic DNA

Definitions:

PROVIDER: Principal Investigator responsible for the project from which the material originates

RECIPIENT: Principal Investigator responsible for the project where the research data will be generated

RECIPIENT SCIENTISTS: Scientists working in the group of the RECIPIENT to generate the data for the specified project

MATERIAL: Refers to the specific biological samples as detailed in Appendix A to this agreement which describes the nature of the project

The PROVIDER asks that the RECIPIENT and the RECIPIENT SCIENTISTS agree to the following before the RECIPIENT receives the MATERIAL:

1. The above MATERIAL is the property of the PROVIDER and is made available for a specific purpose (Generating CardioMetabochip data on the samples provided).

2. THIS MATERIAL IS NOT FOR USE IN HUMAN SUBJECTS.

3. The MATERIAL will be used for the stated purpose only.

4. The MATERIAL will be used as described in Appendix A only. Any other use of the MATERIAL requires the PROVIDER's written consent and an amendment from the relevant Research Ethics Committee.

5. The data generated by the MATERIAL, once it has been used as described in Appendix A, shall be retained by the RECIPIENT for a limited time only AND will be sent to the PROVIDER for use in the public domain according to a stipulated process of the Wits-INEPTH H3Africa funded partnership program. The RECIPIENT will refer requests for access to the data to the appropriate Wits-INDEPTH partnership committee on which the PROVIDER will have representation.

6. The MATERIAL will not be further distributed to others. Any MATERIAL left over after the stipulated experiments have been concluded will be kept until completion of the project and will then be destroyed, as agreed between the parties.

7. The RECIPIENT will not publish any of the data generated.

8. Any MATERIAL delivered pursuant to this Agreement is understood to be experimental in nature and may have hazardous properties. The MATERIAL is of human origin and may pose unknown and unrecognized health or safety risks. The RECIPIENT will handle the MATERIAL using appropriate procedures for human-source materials. THE PROVIDER MAKES NO REPRESENTATIONS AND EXTENDS NO WARRANTIES OF ANY KIND, EITHER EXPRESSED OR IMPLIED. THERE ARE NO EXPRESS OR IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE, OR THAT THE USE OF THE MATERIAL WILL NOT INFRINGE ANY

PATENT, COPYRIGHT, TRADEMARK, OR OTHER PROPRIETARY RIGHTS. Unless prohibited by law, RECIPIENT assumes all liability for claims for damages against it by third parties which may arise from the use, storage or disposal of the MATERIAL except that, to the extent permitted by law, the PROVIDER shall be liable to the RECIPIENT when the damage is caused by the gross negligence or willful misconduct of the PROVIDER.

9. The Material provided by the PROVIDER Institution will be de-identified. The RECIPIENT Institution will not be provided with any information that could be used to identify the subjects from whom the MATERIAL was collected, although the PROVIDER Institution may retain a confidential link to the subject's identity. Neither RECIPIENT Institution nor RECIPIENT Scientists shall make any attempts to determine the identity of those subjects, or to contact the subjects. Should a human subject from whom MATERIAL was collected object to the use set forth herein, then the RECIPIENT Institution agrees to promptly comply with PROVIDER Institution's request to return or destroy any such sample.

10. RECIPIENT agrees to use the MATERIAL in a safe manner and in compliance with all applicable laws and regulations. PROVIDER warrants that it has obtained any Institutional Review Board or Ethics Committee approval required for the transfer of this MATERIAL.

11. The RECIPIENT and PROVIDER take responsibility to the authenticity of the ethics approval certificates (Appendix B) and export, import certificates (Appendix C) as required by and in accordance with the laws of their respective countries.

12. This Agreement will terminate upon completion of RECIPIENT Institution's use of the Material. The Parties agree that the provisions of this Agreement are intended to be interpreted and implemented so as to comply with all applicable laws, governmental rules and regulations; however, if it is determined that any provision of this Agreement is not in such compliance, or if an Institutional Review Board or other comparable body objects to the terms of this Agreement or the provision or use of the Materials, then the Parties agree to modify that provision, or this Agreement so as to be in compliance or to be acceptable to such Institutional Review Board. If such modification is not possible, or practical, or if the Parties are unable to agree upon the modification to be made, then either Party may immediately terminate this Agreement. PROVIDER Institution shall have the right to terminate this Agreement for any reason including RECIPIENT Institution's breach of any of its obligations or responsibilities under this Agreement, and such breach is not cured within thirty days of receipt of written notice from PROVIDER Institution. Upon termination for any reason the RECIPIENT Institution will, at PROVIDER Institution's discretion, either store, or return or destroy any remaining Material in accordance with applicable laws and regulations.

The PROVIDER and RECIPIENT must sign both copies of this letter and return one signed copy to the PROVIDER. The PROVIDER will then send the MATERIAL.

PROVIDER INFORMATION and AUTHORIZED SIGNATURE

Provider: Prof. Michèle Ramsay
Provider contact details: Michele.ramsay@nhls.ac.za, 0114899214
Provider Organisation: National Health Laboratory Services
Provider Address: Department of Human Genetics, Room 109 Watkins Pitchford Building, National Health Laboratory Services, Braamfontein, 2000, South Africa

M. Ramsay

29 April 2013

Signature of Provider Date

Name of Authorised Official:
Title of Authorised Official:

03/05/2013

Signature of Authorised Official Date

Chair Human Research Ethics Committee (Medical)

RECIPIENT INFORMATION and AUTHORIZED SIGNATURE

Recipient: Dr Vanessa Rashbrook
Recipient contact details: tel: 530-754-5281, vkrashbrook@ucdavis.edu, fax: 530-754-9658
Recipient Organisation: University of California, Davis
Recipient Address: DNA Technologies Core
4212A GBSF
451 Health Sciences Drive
University of California - Davis
Davis, CA 95616

Vanessa Rashbrook 9 May 2013

Signature of Recipient Date

Name of Authorised Official:
Title of Authorised Official:

9 May 2013

Signature of Authorised Official Date

Appendix A

Aim and objectives of the project

Title of the Research Project: Identification of Genetic Markers of Obesity Risk and Body Composition in a South African Black Population.

The overall aim of the proposed research is to identify genetic markers of obesity risk in a South African Black population (caregivers of the Bt20 cohort)

To achieve this, we will study the genetic population contributions of the Bt20 caregiver cohort, to aid optimal design of the association studies. We will then perform an association study, focusing on loci associated with obesity measures (e.g. Body mass index (BMI), body fat %, waist to hip circumference) in Europeans, to confirm if these loci are similarly associated with obesity risk in Africans, using the Human CardioMetabochip (Illumina).

AIM: Fine map loci previously associated with body composition to assess the association in an African population, and identify the candidate regions at these loci using the data obtained from the cardioMetabochip.

Brief description of the project

Obesity is a considerable risk factor for the development of several chronic, non-communicable diseases (CNDC), and is becoming more prevalent in developing countries such a South Africa. Although environmental factors have a considerable impact on obesity risk it also has a genetic component (Bodurtha et al., 1990; Wardle et al., 2008). The study of syndromic obesity has provided some insight into the genetic causes of obesity and genome-wide association studies (GWAS) have provided further evidence for common obesity risk loci. All current GWAS for body composition and obesity are based on non-African populations, with an explicit deficit of African-centric research. Also the identification of causal variants in genetic association studies in African groups is more efficient, because linkage disequilibrium (LD) exists over a shorter genomic distance. Furthermore, the metabolic consequences of obesity are highly dependant on body fat distribution and ethnic differences in body composition are well documented (Crowther and Ferris, 2010). The factors underlying these differences are not fully understood and there is an explicit shortfall in African-focused research into genetic factors for obesity and body composition. Although

variation within and around numerous genetic loci has been associated with obesity, heritability is not fully explained by our current knowledge of genetic variation. This study will utilize available longitudinal data to search for molecular obesity risk factors across the life-course and will be the first study of its kind focusing on a black South African cohort (Bt20).

Participants in the project
Ethnicity: SE Bantu (South African Black)
Relevant phenotype data: Body mass index (BMI), body fat %, waist to hip circumference

Precise Description of the Material (could be in table format):
Number of samples: 1034
Quantity of material per sample: 25-50µl of 50ng/µl DNA samples in 11 Deep well- round bottom PCR plates.

Ethics approval (Institutional code and number): Clearance certificate for WITS_INDEPTH partnership: Genomic and Environment Risk Factors for Cardiometabolic Disease in Africans (Certificate number; M1210209).
Clearance certificate for research project: "Identification of Genetic Markers of Obesity Risk and Body Composition in a South African Black population." (Certificate number; M120647).
Both certificates attached.

Data analysis and data sharing strategy: To be decided once data is received, data will be analysed using PLINK (Harvard University freeware).

Ownership of data and IP: Letter of intent to publish attached.

Intent to publish (in accordance with Wits-INDEPTH H3Africa funded research protocols): refer above

Financial implications – Provider will pay for services rendered by DNA Technologies Core

Appendix B

Copy of ethics approval certificates

1. From provider:
    a. Approval to collect the samples
    b. Approval for their use in this project
2. From recipient:
    a. Not required (Recipient is a Service Provider – all data generated will be returned to Provider and after an agreed time period the data will be removed from the recipient database)

Appendix C

Legal documents as required from the DONOR and RECIPIENT countries
e.g. Export permit from the South African National Department of Health

MRC/Wits Developmental Pathways for Health Research Unit
Department of Paediatrics
School of Clinical Medicine
Faculty of Health Sciences
University of the Witwatersrand, Johannesburg
17th July 2011

The Chair
Human Ethics Research Committee
University of the Witwatersrand, Johannesburg

Dear Professor Cleaton-Jones,

Re-consenting Birth to Twenty cohort participants around historical blood and DNA samples

When Birth to Ten transitioned to Birth to Twenty and the Bt20 cohort and the Bone Health sub-cohort was formed, the HERC approved protocols that entailed the collection of blood and DNA samples for the analyses linked to non-communicable disease (obesity, bone health, diabetes, etc). The participants are now 21 years of age and we wish to re-consent them around the management and use of biological samples collected.

Below are the information and consent sheets. We are asking the participants to consent to the following:

- I acknowledge that all procedure/tests on the stored blood and DNA samples have been or will be approved by the Human Research Ethics Committee of the University of the Witwatersrand.
- I understand that every time a new study is done on my DNA, permission will be obtained from the ethics committee for the study to make sure that it is used appropriately.
- I am in agreement that my DNA may be stored and used for the purposes described.
- I am in agreement that the data generated from my DNA may be made available in a public domain without any identifiers.
- I agree that a small bit of my DNA may be sent out of the country if the research cannot easily be done in South Africa.
- I understand that I will not benefit directly from the research done on my DNA.
- I understand that I may withdraw from the study at any time.

We piloted the re-consent procedure with 10 biological parents of the Bt20 cohort. The re-consent procedure included:

- Power-Point on genetics and core concepts presented by a masters-level research assistant in both English and Zulu.
- Question & Answer session.
- Information sheet distributed and discussed as a group.
- Question & Answer session.
- Individual consent in private with a trained research assistant.

The response from the pilot study was extremely positive. The participants understood and appreciated the presentation; they understood the information sheet and 100% consented to all of the points above. I seek permission from the HERC to approve the re-consent documents and procedure.

Shane Norris

INFORMATION SHEET
Birth to Twenty participants
Re-consenting around historical blood and DNA samples

Historical blood samples

You have been a part of the Birth to Twenty cohort since 1990, now over 20 years. We have collected data on the Birth to Twenty participants at 16 time points. At some of these time points you provided consent for us to collect blood samples from you to test for health indicators such as blood sugar (glucose), cholesterol, vitamin D, calcium, etc. These blood samples are stored at the University of the Witwatersrand with only a unique number identifier so your name is not attached to the sample. The information linking you name to the unique number identifier is locked away in the study offices with strict access control and is not made available to the people working on your blood sample.

In the future we may run additional tests on the stored blood samples, but we will not do so without the approval of the Human Research Ethics Committee of the University of the Witwatersrand.

Historical DNA samples

From the blood samples we collected from you in 2003/2004 you consented to us extracting DNA (the inherited material in your cells) so that this DNA could be used for studies that look for genes/changes in DNA that are involved in causing diabetes, obesity and cardiovascular diseases, and for studies to understand how the DNA works (we call this epigenetic studies). DNA studies give us information about how your body works and also whether you are more likely to get certain conditions or diseases. It will also tell us about your ancestry (your family) and about the way that your body uses medications (we call this pharmacogenetics).

Just as for the blood samples, all DNA samples are stored in a safe place at the University of the Witwatersrand and strict control access. If at some future date we realize there are other studies that we would like to carry out on your DNA samples, such tests would only be performed if permission is given to us by the Human Research Ethics Committee of the University of the Witwatersrand. Your identity will be anonymous as your sample will be identified by a number (as described above).

Blood and DNA sample management

The scientists who do the research will generate information (data) which will be placed in databases. Some of the information in the databases will be shared with other scientists, possibly around the world. These scientists will not have access to your name and therefore the data will not be linked back to you. It is possible that we may send small amounts of the DNA out of the country where tests will be done that we cannot easily do in South Africa, but that would give us valuable information for our studies.

Benefits

The discoveries that come from the studies on your DNA will not be of direct benefit to you and will not be communicated back to you. The discoveries may lead to information that will help us in the future to diagnoses disease, understand who is most likely to get ill and how different people behave when they are given medication. The DNA belongs to you and we are the keepers of the DNA. You may withdraw your sample at any time.

Risks

Since the DNA of every person is different, it is possible that if someone tested your blood and compared it to the data in the databases, they could conclude that the two samples come from the same person. However, this is unlikely given the procedures in place that protect your samples.

CONSENT SHEET

The information around the blood and DNA samples taken from me in the past is clear and that the purpose is for me to inform the study what they can or cannot do with these samples.

I acknowledge that all procedure/tests on the stored blood and DNA samples have been or will be approved by the Human Research Ethics Committee of the University of the Witwatersrand.

YES ☐ NO ☐

I am in agreement that my DNA may be stored and used for the purposes described above.

YES ☐ NO ☐

I am in agreement that the data generated from my DNA may be made available in a public domain without any identifiers.

YES ☐ NO ☐

I agree that a small bit of my DNA may be sent out of the country if the research cannot easily be done in South Africa.

YES ☐ NO ☐

I understand that every time a new study is done on my DNA, permission will be obtained from the ethics committee for the study to make sure that it is used only for the purposes stated above.

YES ☐ NO ☐

I understand that I will not benefit directly from the research done on my DNA.

YES ☐ NO ☐

I understand that I may withdraw from the study at any time.

YES ☐ NO ☐

RESEARCH ASSISTANT:

_____

Printed Name Signature/Mark or Thumbprint
Date and Time

PARTICIPANTS:

_____

Printed Name Signature/Mark or Thumbprint
Date and Time

WITNESS: (If applicable)

_____

Printed Name Signature/Mark or Thumbprint
Date and Time

University
of the Witwatersrand,
Johannesburg

Human Research Ethics Committee (Medical)
(formerly Committee for Research on Human Subjects (Medical)

Secretariat: Research Office, Room SH10005, 10th floor, Senate House • Telephone: +27 11 717-1234 • Fax: +27 11 339-5708
Private Bag 3, Wits 2050, South Africa

01 August 2011

Professor Shane Norris
Director
MRC/Wits Developmental
Pathways for Health Research Unit
Department of Paediatrics
School of Clinical Medicine
Faculty of Health Sciences
University

Dear Professor Norris

RE: Re-Consenting Birth to Twenty Cohort Participants around historical blood and DNA Samples

This letter serves to confirm that the Human Research Ethics Committee (Medical) at its meeting on 29 August 2011 discussed your request to "Re-consenting Birth to Twenty Cohort Participants around historical blood and DNA samples" as detailed in your letter dated 17 July 2011.

I am happy to inform you that the Committee unanimously approved your request to "Re-consenting the Birth to Twenty Cohort" and included in this approval are the Information Sheet and Consent Form.

Thank you for keeping us informed and updated.

Yours sincerely,

Professor PE Cleaton-Jones
Chairman
Human Research Ethics Committee (Medical)

APPENDIX B – Script for conversion of GenomeStudio forward report files into transposed PLINK files

```
#!/usr/local/bin/python
import os

# Arrays
snpdata = {}
csvfiles = []
tfamarray = []
tpedarray = {}
tfamfilename = "batch1.tfam"
tpedfilename = "batch1.tped"

# Create a list of CSV files
files=os.listdir("./")
for eachfile in files:
      if eachfile[-3:] == "csv":
            csvfiles.append(eachfile)

# Read in metabochip SNP annotations
annotationfile = open("Metabochip_Gene_Annotation.txt",'r')
annotations = annotationfile.readlines()
for annotation in annotations:
      annotation = annotation.strip().split("\t")
      snpdata[annotation[0]]                                    =
[annotation[1],annotation[0],"0",annotation[2]]
annotationfile.close()

# Read in individual CSV files
for eachfile in csvfiles:
      datafile = open(eachfile, 'r')
      FRdata = datafile.readlines()
      samplerow = ""
```

```
      for i in range(10):
            samplerow = FRdata.pop(0)
      samplelist = samplerow.strip().strip(",").split(",")
      print samplelist
      tfamarray.extend(samplelist)
      for row in FRdata:
            row = row.strip().split(",")
            snpid = row.pop(0)
            for genotype in row:
                  try:
                        if genotype != "--":

      tpedarray[snpid].append(genotype[0])

      tpedarray[snpid].append(genotype[1])
                        else:
                              tpedarray[snpid].append("0")
                              tpedarray[snpid].append("0")
                  except:
                        if genotype != "--":
                              tpedarray[snpid]                =
[(genotype[0])]

      tpedarray[snpid].append(genotype[1])
                        else:
                              tpedarray[snpid] = ["0"]
                              tpedarray[snpid].append("0")



      datafile.close()

# Write out tfam file
tfamfile = open(tfamfilename, 'w')
for individual in tfamarray:
      tfamfile.write("\t".join([individual,individual,"0","0
","0","0"])+"\n")
```

```
tfamfile.close()

# Write out tped file
tpedfile = open(tpedfilename, 'w')
for snp in tpedarray.keys():
      snpcols = "\t".join(snpdata[snp])
      tpedfile.write(snpcols+"\t"+"\t".join(tpedarray[snp])+
"\n")
tpedfile.close()

print tfamarray
```

APPENDIX C – PLINK, SMARTPCA GEMMA, SHAPEIT, IMPUTE2, SNPTEST and R commands for chapters 2, 4 and 5

Chapter 2 - QC Steps

- Conversion from transposed to binary PLINK files:

```
plink --tfile File1 --make-bed --out File1 --allow-no-sex
```

--tfile: Specifies the PLINK input file prefix (in this case it reads a transposed fileset).
--make-bed: Directs PLINK to create a binary fileset.
--out: Specifies the output file prefix.
--allow-no-sex: Disables the automatic setting of the phenotype to missing if the individual has an ambiguous sex code.

- Removal of SNPs with complete missing data:

```
plink --bfile File1 --exclude NaNsnps.txt --make-bed --out
File2 --allow-no-sex
```

--bfile: Specifies the PLINK input file prefix (in this case it reads a binary fileset).
--exclude: Directs PLINK to remove the SNPs listed in the specified .txt file.

- Initial removal of poorly genotyped samples:

```
plink --bfile  File2 --mind 0.20 --make-bed --out File3 --
allow-no-sex
```

--mind: Directs PLINK to remove individuals with more than 20% missing genotypes.

- SNP QC (missingness, MAF, HWE):

```
plink --bfile  File3  --geno 0.02 --maf 0.01--hwe 0.00001 --make bed --out File4 --allow-no-sex
```

--geno: Directs PLINK to remove SNPs with more than 2% missing data.
--maf: Directs PLINK to remove SNPs with MAF less than 0.01.
--hwe: Directs PLINK to remove SNPs with HWE p-value less than $1x10^{-5}$.

- Removal of samples with high **missingness** rate:

```
plink --bfile  File4 --mind 0.02 --make-bed --out File5 --allow-no-sex
```
[female caregivers]

AND

```
link --bfile  File4 --mind 0.03 --make-bed --out File5 --allow-no-sex
```
[Bt20 participants]

- **Sex check** and removal of samples with discordant sex:

```
plink --bfile File1 --check-sex --out File1
```

--check-sex: Directs PLINK to perform a sex check and flag any problem individuals.

```
plink --bfile File5 --remove sexremoved.txt --make-bed --out
File6
```

--remove: Directs PLINK to remove the individuals listed in the specified .txt file.

- Removal of **related** samples:

LD pruning:

```
plink --bfile File6 --indep-pairwise 50 5 0.2 --out LDpruned
```

```
plink --bfile File6 --extract LDpruned.prune.in --make-bed -
-out File6_LDpruned
```

--indep-pairwise: Directs PLINK to generate a pruned subset of SNPs that are in approximate linkage equilibrium with each other based on pairwise genotypic correlation. At each step it considers a window of 50 SNPs, shifts the window 5 SNPs forward each time, and calculates the LD between each pair of SNPs in the window and removes one of a pair of SNPs if the LD is greater than 0.5.
--extract: Directs PLINK to extract the SNPs which are not pruned out.

Generation of IBD scores:

```
plink --bfile File6_LDpruned --genome --min 0.05
```

--genome: Directs PLINK to calculate IBD.
--min: Directs PLINK to only output pairs where `PI_HAT` *is greater than 0.05 to the* `.genome` *file.*

Removal of one of each of the pair of individuals:

```
plink --bfile File6 --remove IBDremoved.txt --make-bed --out
File7
```

- Removal of duplicates:

```
plink --bfile File7 --remove duplicates.txt --make-bed --out
File8
```

- Removal of samples based on **extreme heterozygosity**:

Generation of .het file:

```
plink --bfile File2 --het --out File2
```

--het: Directs PLINK to generate an output file containing the observed and expected number of homozygotes and number of non-missing genotypes.

Calculation of missingness:

```
plink --bfile File2 --missing --out File2
```

--missing: Directs PLINK to generate genotyping/missingness rate statistics.

Exclusion of individuals with a heterozygosity rate ± 3 standard deviations from the mean:

```
plink --bfile File8 --remove extremehet.txt --make-bed --out
File9
```

- **Principal component analysis (PCA)** and removal of population outliers:

LD pruning of the datasets:

```
plink --bfile File9 --indep-pairwise 50 5 0.2 --out LDpruned
```

```
plink --bfile File9 --extract LDpruned.prune.in --make-bed --out File9_LDpruned
```

Running of PCA using SMARTPCA:

```
smartpca.perl -i $1.bed -a $1.bim -b $1.fam -p $1.pca -e $1.eval -o $1.pca -q NO -l $1.log
```

-i: Specifies input genotype file.

-a: Specifies input SNP file

-b: Specifies input individual file.

-p: Specifies prefix of output plot files of top two PCs.

-e: Specifies output file of all eigenvalues.

-o: Specifies output file of PCs.

-q: Specifies output log file.

-l: Set to NO to indicate that there is not a single population and that the population field doesn't contain real-valued phenotypes.

Removal of population outlier:

```
plink --bfile File9 --remove popoutliers.txt --make-bed --out File10
```

** File10_1 = Final QC'ed female caregiver genotype files*

** File10_2 = Final QC'ed Bt20 participant genotype files*

Chapter 2 - Merging of datasets and update to Build 37

```
plink --bfile File10_1 --bmerge File10_2.bed File10_2.bim
File10_2.fam --make-bed –-out merged_ALL --allow-no-sex

plink --bfile merged_ALL --exclude Same_pos_SNPs.txt --
make-bed --out merged_ALL2 --allow-no-sex

plink --bfile merged_ALL2 --geno 0.05 --make-bed --out
merged_pruned_ALL --allow-no-sex
```

--bmerge: Directs PLINK to merge binary PLINK files.

Chapter 4 – Association analysis of merged dataset in GEMMA

Creation of relatedness matrix:

```
gemma -bfile merged_pruned_ALL_DBP -gk 1 –o
Merged_ALL_DBP_RM                    [DBP]

gemma -bfile merged_pruned_ALL_SBP -gk 1 -o
Merged_ALL_SBP_RM                    [SBP]

gemma -bfile merged_pruned_ALL_HT -gk 1 -o Merged_ALL_HT_RM
                                     [high versus normal/low BP]
```

-bfile: Specifies the PLINK binary file prefix.

-gk 1: Directs GEMMA to calculate the centered relatedness matrix.

-o: Specifies the output file prefix.

Association analysis (with correction for covariates):

```
gemma \
-bfile merged_pruned _ALL_DBP \
-k Merged_ALL_DBP_RM.cXX.txt \
-c Merged_ALL_covariates.txt \
-lmm 4 \
-o Merged_ALL_DBP_ULMM_covar
```
[DBP]

```
gemma \
-bfile merged_pruned _ALL_SBP \
-k Merged_ALL_SBP_RM.cXX.txt \
-c Merged_ALL_covariates.txt \
-lmm 4 \
-o Merged_ALL_SBP_ULMM_covar
```
[SBP]

```
gemma \
-bfile merged_pruned _ALL_HT \
-k Merged_ALL_HT_RM.cXX.txt \
-c Merged_ALL_covariates.txt \
-lmm 4 \
-o Merged_ALL_HT_ULMM_covar
```
[high versus normal/low BP]

-k: Specifies the relatedness matrix file name.

-lmm4: Directs GEMMA to perform a Wald test, likelihood ratio test and score test. P-values generated from these three tests were almost identical and only the Wald test p-values were therefore recorded.

-c: Specifies the covariate file name in the case of multivariate analysis. The covariate file is in the format: Column 1: intercept term (1's), Columns 2 to n: covariates and headings are removed.

Chapter 4 – Association analysis of individual datasets in PLINK (with correction for covariates)

```
plink \
--bfile File10 \
--pheno Caregiver_Phenotypes.txt \
--pheno-name DBP \
--linear \
--allow-no-sex \
--covar Caregiver_Phenotypes.txt \
--covar-name AGE,BMI \
--out CG_ALL_DBP_multi_linear    [DBP]
```

```
plink \
--bfile File10 \
--pheno Caregiver_Phenotypes.txt \
--pheno-name SBP \
--linear \
--allow-no-sex \
--covar Caregiver_Phenotypes.txt \
--covar-name AGE,BMI \
--out CG_ALL_SBP_multi_linear    [SBP]
```

```
plink \
--bfile File10 \
--pheno Caregiver_Phenotypes.txt \
--pheno-name HT \
--logistic \
--ci 0.95 \
--allow-no-sex \
--covar Caregiver_Phenotypes.txt \
--covar-name AGE,BMI \
--out CG_ALL_HT_multi_logistic   [high versus normal/low BP]
```

--bfile: Specifies the PLINK binary file prefix.

--pheno: Specifies the phenotype file name.

--pheno-name: Specifies the phenotype to use in the analysis.

--linear/logistic: Directs PLINK to perform linear/logistic regression analysis.

--ci: Gives the specified confidence intervals for the estimated parameters.

--allow-no-sex: Disables the automatic setting of the phenotype to missing if the individual has an ambiguous sex code.

--out: Specifies the output file prefix.

--covar: Specifies the covariate file name.

--covar-name: Specifies the covariate(s) to use in the analysis.

The PLINK commands above are for the female caregiver dataset. The same commands applied for analysis in each of the individual datasets, with inclusion of appropriate covariates in each case.

Chapter 4 – Manhattan, Q-Q plots and calculation of genomic inflation factors (R)

Manhattan plots:

```
Plot = read.table("Merged_ALL_DBP_ULMM_covar.assoc.txt",
header = T,as.is=T)

manhattan(Plot, chr = "chr", bp = "ps", p = "p_wald", snp =
"rs", col = c("gray10", "gray60"), chrlabs = NULL, main =
"Merged_ALL_DBP_multi",suggestiveline = -log10(1e-04),
genomewideline = -log10(6.7e-07),ylim=c(0,8), highlight =
NULL, logp = TRUE)
```

[Merged, DBP]

Q-Q plots:

```
Plot = read.table("Merged_ALL_DBP_ULMM_covar.assoc.txt",
header = T,as.is=T)
```

```
qq(Plot$p_wald, main="Merged (all): DBP (with correction for
covariates)")
```

[Merged, DBP]

Genomic inflation factor calculation:

```
median(qchisq(Plot1a[,9],df=1,lower.tail=F),na.rm=T)/0.456
```

Chapter 5 – Preparation of files for imputation (PLINK and SHAPEIT)

Extraction of chromosome containing gene of interest:

```
plink --bfile merged_pruned_ALL_37 --chr 1 --make-bed --out
merged_pruned_ALL_37_chr1 --allow-no-sex
```

Strand check, flip and removal of problem SNPs:

```
shapeit \
-check \
-B merged_pruned_ALL_37_chr1 \
-M genetic_map_chr1_combined_b37.txt \
--input-ref 1000GP_Phase3_chr1.hap 1000GP_Phase3_chr1.legend
1000GP_Phase3.sample \
--output-log merged_pruned_37_ALL_chr1.alignments
```

-B: Specifies the input filename prefix of unphased genotypes in PLINK binary file format.

-M: Specifies the input genetic map file.

--input-ref: Specifies the input reference panel of haplotypes in the IMPUTE2 file format (.hap contains the reference haplotypes, .legend contains the SNP map, .sample contains information about the individuals).

--output-log: Specifies the log file where the screen output is copied and gives the prefix of all the files generated by SHAPEIT when checking input data.

```
plink –bfile merged_pruned_ALL_37_chr1 –flip
Chr1_strand_issue_SNPs.txt --make-bed –out
merged_pruned_ALL_37_chr1_2

plink --bfile merged_pruned_ALL_37_chr1_2 --exclude
Strand_issue_SNPs_to_exclude.txt --make-bed --out
merged_pruned_ALL_37_chr1_3 --allow-no-sex
```

Chapter 5 – Pre-phasing (SHAPEIT)

```
shapeit \
-B merged_pruned_ALL_37_chr1_3 \
-M genetic_map_chr1_combined_b37.txt \
--duohmm \
-W 5 \
--output-max merged_pruned_ALL_37_chr1_3.phased.haps
merged_pruned_ALL_37_chr1_3.phased.sample
```

--duohmm and -W: Take into account the relatedness of the individuals in the sample to improve phasing.
--output-max: The most likely pair of haplotypes for each individual in IMPUTE2 format is added to the .haps file and all extra information present in the input files is added to the .sample file.

Chapter 5 – Imputation (IMPUTE2)

For the merged dataset:

```
impute2 \
-use_prephased_g \
```

```
-known_haps_g merged_pruned_ALL_37_chr1_3.phased.haps \
-m genetic_map_chr1_combined_b37.txt \
-h 1000GP_Phase3_chr1.hap \
-l 1000GP_Phase3_chr1.legend \
-int 162039564 162340265 \
-Ne 20000 \
-buffer 500 \
-o NOS1AP_MERGED.gen
```

For the individual datasets:

```
impute2 \
-use_prephased_g \
-known_haps_g merged_pruned_ALL_37_chr1_3.phased.haps \
-m genetic_map_chr1_combined_b37.txt \
-h 1000GP_Phase3_chr1.hap \
-l 1000GP_Phase3_chr1.legend \
-sample_g merged.sample \
-exclude_samples_g Children_IDs.txt \
-int 162039564 162340265 \
-Ne 20000 \
-buffer 500 \
-o NOS1AP_CG.gen
```

-use_prephased_g: Tells IMPUTE2 to perform imputation with pre-phased haplotypes (generated in the pre-phasing step).

-known_haps_g: Specifies the file containing the known pre-phased haplotypes (generated in the pre-phasing step).

-m: Specifies fine-scale recombination map for the region to be analysed.

-h: Specifies file of known haplotypes.

-l: Specifies legend file with information about the SNPs in the '-h' file.

-sample_g: File of sample IDs for the individuals in the haplotypes input files (SPECIFIC TO INDIVIDUAL DATASET IMPUTATION).

-exclude_samples_g: Indicates which samples to exclude from the pre-phased files during imputation (e.g. if we just want to impute genotypes in the female caregiver dataset, the exclude file would contain all IDs of the Bt20 participants to exclude) (SPECIFIC TO INDIVIDUAL DATASET IMPUTATION).
-int: Specifies the region/genomic interval to be imputed.
-Ne: Specifies the "effective size" of the population from which the dataset was sampled. 20000 recommended for most modern imputation analyses.
-buffer: Specifies the length of the buffer region (in kilobases) to include on each side of the analysis interval specified by the '-int' option.
-o: Specifies the name of the main output file.

Chapter 5 – Association analysis of merged imputed dataset (GEMMA)

Conversion of imputation output files to PLINK files:

```
fcgene --gens NOS1AP_MERGED.gen --info
NOS1AP_MERGED.gen_info --info-thresh 0.4 --oformat plink --
out NOS1AP_MERGED
```

Creation of relatedness matrix:

```
gemma -bfile NOS1AP_MERGED_SBP -gk 1 -o NOS1AP_MERGED_SBP_RM
```

[SBP, NOS1AP]

Association analysis:

```
gemma -bfile NOS1AP_MERGED_SBP -k
NOS1AP_MERGED_SBP_RM.cXX.txt -c Covariates.txt -lmm 4 -o
NOS1AP_MERGED_SBP_ULMM_multi
```

[SBP, NOS1AP]

Chapter 5 – Association analysis of individual imputed datasets (SNPTEST)

```
snptest_v2.5.1 \
-data NOS1AP_CG.gen CG.sample \
-frequentist 1 \
-method score \
-pheno pheno1 \
-cov_names cov_1  cov_4 \
-o NOS1AP_CG_SBP _multi.txt
```

[SBP, NOS1AP]

-data: Specifies the input genotype (.gen) and the sample files. The sample file stores the IDs and associated phenotype and covariate information for each individual.

-frequentist: Specifies which model of association to test. 1 represents an additive model.

-method: Controls the way genotype uncertainty is taken into account. 'score' uses a missing data likelihood score test.

-pheno: Specifies which phenotype in the sample file to test.

-cov_names: Specifies which covariate(s) in the sample file to test.

-o: Specifies the name of the output file.

Chapter 5 – Result visualisation (R)

```
NOS1AP_SBP =
read.table("NOS1AP_MERGED_SBP_ULMM_multi.assoc.txt", header
= T,as.is=T)

plot(NOS1AP_SBP$ps, -log10(NOS1AP_SBP
$p_wald)*(NOS1AP_SBP[,1] == "---"), ylim = c(0, 8), main =
"NOS1AP Typed + Imputed SNPs: SBP (Merged, with correction
for covariates)", col = 8*(NOS1AP_SBP[,1] == "---"), pch =
16*(NOS1AP_SBP[,1] == "---"), ylab = "-log10 p-value", xlab=
"BP Position")
```

```
par(new=T)

plot(NOS1AP_SBP$ps, -
log10(NOS1AP_SBP$p_wald)*(NOS1AP_SBP[,1] == "1"), ylim =
c(0, 8), main = "NOS1AP Typed + Imputed SNPs: SBP (Merged,
with correction for covariates)", col = 1*(NOS1AP_SBP[,1] ==
"1"), pch = 16*(NOS1AP_SBP[,1] == "1"), ylab = "-log10 p-
value", xlab= "BP Position")

par(new=T)

abline(h = -log10(2.9e-05), untf = FALSE, col=2)
abline(h = -log10(1e-04), untf = FALSE, col=4)

par(new=F)
```

APPENDIX D – MySQL code for creation of database tables

```
CREATE DATABASE metabobtt;

USE metabobtt;
```

#Phenotype data tables:

```
CREATE TABLE female_caregivers_basic (
      IndividualID          VARCHAR(5) NOT NULL PRIMARY KEY,
      Gender                VARCHAR(1),
      Ethnicity             INTEGER(1),
      RelationshipToChild   VARCHAR(1) ) ;

CREATE TABLE bt20_participants_basic (
      IndividualID          VARCHAR(4) NOT NULL PRIMARY KEY,
      Gender                VARCHAR(1),
      Ethnicity             INTEGER(1),
      CaregiversID          VARCHAR(5));

CREATE TABLE phenotype_yr13_female_caregivers (
      IndividualID          VARCHAR (5)NOT NULL PRIMARY KEY,
      Age_yr13              VARCHAR(6),
      Height_yr13           VARCHAR (5),
      Weight_yr13           VARCHAR (5),
      BMI_yr13              VARCHAR (6),
      HipCircumference_yr13 VARCHAR (5),
      WaistCircumference_yr13    VARCHAR (5),
      WaistToHipRatio_yr13  VARCHAR (5),
      AverageSBP_yr13       VARCHAR (5),
      AverageDBP_yr13       VARCHAR (5),
      SubtotalBodyFat_yr13  VARCHAR (7),
      SubtotalLeanMass_yr13 VARCHAR (7),
      PercentageBodyFat_yr13     VARCHAR (4));
```

```
CREATE TABLE phenotype_yrN_bt20_participants (
     IndividualID          VARCHAR(4) NOT NULL PRIMARY KEY,
     Age_yrN               VARCHAR(6),
     Height_yrN            VARCHAR (5),
     Weight_yrN            VARCHAR (5),
     BMI_yrN               VARCHAR (6),
     HipCircumference_yrN  VARCHAR (5),
     WaistCircumference_yrN     VARCHAR (5),
     WaistToHipRatio_yrN   VARCHAR (5),
     AverageSBP_yrN        VARCHAR (5),
     AverageDBP_yrN        VARCHAR (5),
     SubtotalBodyFat_yrN   VARCHAR (7),
     SubtotalLeanMass_yrN  VARCHAR (7),
     PercentageBodyFat_yrN VARCHAR (4));
```

(Separate table were created for each data collection time point. N=5, 7, 9/10, 11/12, 13, 14, 15, 16, 17/18, 19, 20)

#Metabochip data tables:

```
CREATE TABLE metabochip_snp_information_1 (
     SNPID36               VARCHAR(17)NOT NULL PRIMARY KEY,
     SNPID37               VARCHAR(17)NOT NULL UNIQUE,
     Chromosome            VARCHAR(2),
     BasePairPosition36    VARCHAR(9),
     BasePairPosition37    VARCHAR(9),
     NearestGene           VARCHAR(143),
     LocationWithinGene    VARCHAR(20));

CREATE TABLE metabochip_snp_information_extra(
     SNPID36          VARCHAR(17)  NOT NULL  PRIMARY KEY,
     Allele1          CHAR(1),
     Allele2          CHAR(1));

CREATE TABLE metabochip_snp_information_extra2 (
     SNPID36          VARCHAR(17)  NOT NULL  PRIMARY KEY,
```

```
    AfterQC         VARCHAR(6));

CREATE TABLE metabochip_snp_information_2 AS (SELECT
metabochip_snp_information_1.*,
metabochip_snp_information_extra.Allele1,
metabochip_snp_information_extra.Allele2 FROM
metabochip_snp_information_1 LEFT JOIN
metabochip_snp_information_extra ON
metabochip_snp_information_1.SNPID36=
metabochip_snp_information_extra.SNPID36);

CREATE TABLE metabochip_snp_information AS (SELECT
metabochip_snp_information_2.*,
metabochip_snp_information_extra2.AfterQC FROM
metabochip_snp_information_2 LEFT JOIN
metabochip_snp_information_extra2 ON
metabochip_snp_information_2.SNPID36=
metabochip_snp_information_extra2.SNPID36);

ALTER TABLE metabochip_snp_information ADD PRIMARY
KEY(SNPID36);
```

#Association analysis data tables:

```
CREATE TABLE metabochip_snp_associations_batch1_all (
    SNPID36          VARCHAR (17)          NOT NULL ,
    Dataset          VARCHAR (14),
    Phenotype        CHAR (3),
    P_valueUncorrected    VARCHAR(20),
    Beta_ORUncorrected    VARCHAR(20),
    P_valueCorrected      VARCHAR(20),
    Beta_ORCorrected      VARCHAR(20),
    Covariate1       CHAR (3),
    Covariate2       CHAR (3),
    Covariate3       CHAR (3));
```

```
ALTER TABLE metabochip_snp_associations_batch1_all ADD
PRIMARY KEY(SNPID36,Dataset,Phenotype);

CREATE TABLE metabochip_snp_associations_batch2_all(
     SNPID36          VARCHAR (17)          NOT NULL ,
     Dataset          VARCHAR (14),
     Phenotype        CHAR (3),
     P_valueUncorrected    VARCHAR(20),
     Beta_ORUncorrected    VARCHAR(20),
     P_valueCorrected      VARCHAR(20),
     Beta_ORCorrected      VARCHAR(20),
     Covariate1       CHAR (3),
     Covariate2       CHAR (3),
     Covariate3       CHAR (3));

ALTER TABLE metabochip_snp_associations_batch2_all ADD
PRIMARY KEY(SNPID36,Dataset,Phenotype);

CREATE TABLE metabochip_snp_associations_batch2_females(
     SNPID36          VARCHAR (17)          NOT NULL ,
     Dataset          VARCHAR (14),
     Phenotype        CHAR (3),
     P_valueUncorrected    VARCHAR(20),
     Beta_ORUncorrected    VARCHAR(20),
     P_valueCorrected      VARCHAR(20),
     Beta_ORCorrected      VARCHAR(20),
     Covariate1       CHAR (3),
     Covariate2       CHAR (3),
     Covariate3       CHAR (3));

ALTER TABLE metabochip_snp_associations_batch2_females ADD
PRIMARY KEY(SNPID36,Dataset,Phenotype);

CREATE TABLE metabochip_snp_associations_batch2_males(
     SNPID36          VARCHAR (17)          NOT NULL ,
     Dataset          VARCHAR (14),
```

```
      Phenotype        CHAR (3),
      P_valueUncorrected    VARCHAR(20),
      Beta_ORUncorrected    VARCHAR(20),
      P_valueCorrected      VARCHAR(20),
      Beta_ORCorrected      VARCHAR(20),
      Covariate1       CHAR (3),
      Covariate2       CHAR (3),
      Covariate3       CHAR (3));

ALTER TABLE metabochip_snp_associations_batch2_males ADD
PRIMARY KEY(SNPID36,Dataset,Phenotype);

CREATE TABLE metabochip_snp_associations_pre2_all(
      SNPID36          VARCHAR (17)          NOT NULL ,
      Dataset          VARCHAR (14),
      Phenotype        CHAR (3),
      P_valueUncorrected    VARCHAR(20),
      Beta_ORUncorrected    VARCHAR(20),
      PRIMARY KEY(SNPID36,Dataset,Phenotype));

CREATE TABLE metabochip_snp_associations_pre3_all (
      SNPID36          VARCHAR (17)          NOT NULL ,
      Dataset          VARCHAR (14),
      Phenotype        CHAR (3),
      P_valueCorrected      VARCHAR(20),
      Beta_ORCorrected      VARCHAR(20),
      Covariate1       CHAR (3),
      Covariate2       CHAR (3),
      Covariate3       CHAR (3),
      PRIMARY KEY(SNPID36,Dataset,Phenotype));

CREATE TABLE metabochip_snp_associations_merged_all AS
(SELECT metabochip_snp_associations_pre3_all.SNPID36,
metabochip_snp_associations_pre3_all.Dataset,
metabochip_snp_associations_pre3_all.Phenotype,
metabochip_snp_associations_pre2_all.P_valueUncorrected,
```

```
metabochip_snp_associations_pre2_all.Beta_ORUncorrected,
metabochip_snp_associations_pre3_all.P_valueCorrected,
metabochip_snp_associations_pre3_all.Beta_ORCorrected,
metabochip_snp_associations_pre3_all.Covariate1,
metabochip_snp_associations_pre3_all.Covariate2,
metabochip_snp_associations_pre3_all.Covariate3 FROM
metabochip_snp_associations_pre2_all RIGHT JOIN
metabochip_snp_associations_pre3_all ON
metabochip_snp_associations_pre2_all.SNPID36 =
metabochip_snp_associations_pre3_all.SNPID36 AND
metabochip_snp_associations_pre2_all.Dataset =
metabochip_snp_associations_pre3_all.Dataset AND
metabochip_snp_associations_pre2_all.Phenotype =
metabochip_snp_associations_pre3_all.Phenotype);

ALTER TABLE metabochip_snp_associations_merged_all ADD
PRIMARY KEY(SNPID36,Dataset,Phenotype);

CREATE TABLE metabochip_snp_associations_pre2_females (
     SNPID36          VARCHAR (17)          NOT NULL ,
     Dataset          VARCHAR (14),
     Phenotype        CHAR (3),
     P_valueUncorrected    VARCHAR(20),
     Beta_ORUncorrected    VARCHAR(20),
     PRIMARY KEY(SNPID36,Dataset,Phenotype));

CREATE TABLE metabochip_snp_associations_pre3_females (
     SNPID36          VARCHAR (17)          NOT NULL ,
     Dataset          VARCHAR (14),
     Phenotype        CHAR (3),
     P_valueCorrected      VARCHAR(20),
     Beta_ORCorrected      VARCHAR(20),
     Covariate1       CHAR (3),
     Covariate2       CHAR (3),
     Covariate3       CHAR (3),
     PRIMARY KEY(SNPID36,Dataset,Phenotype));
```

```
CREATE TABLE metabochip_snp_associations_merged_females AS
(SELECT metabochip_snp_associations_pre2_females.*,
metabochip_snp_associations_pre3_females.P_valueCorrected,
metabochip_snp_associations_pre3_females.Beta_ORCorrected,
metabochip_snp_associations_pre3_females.Covariate1,
metabochip_snp_associations_pre3_females.Covariate2,
metabochip_snp_associations_pre3_females.Covariate3 FROM
metabochip_snp_associations_pre2_females LEFT JOIN
metabochip_snp_associations_pre3_females ON
metabochip_snp_associations_pre2_females.SNPID36 =
metabochip_snp_associations_pre3_females.SNPID36 AND
metabochip_snp_associations_pre2_females.Dataset =
metabochip_snp_associations_pre3_females.Dataset AND
metabochip_snp_associations_pre2_females.Phenotype =
metabochip_snp_associations_pre3_females.Phenotype);

ALTER TABLE metabochip_snp_associations_merged_females ADD
PRIMARY KEY(SNPID36,Dataset,Phenotype);
```

#Username and password table:

```
CREATE TABLE user (
     username   VARCHAR (40)     NOT NULL   PRIMARY KEY,
     password   VARCHAR (40));
```

APPENDIX E – Python code for input of data into MySQL tables

 # female_caregivers_basic

```
#!/usr/bin/python

import csv
import MySQLdb

db =
MySQLdb.connect("localhost","<USERNAME>","<PASSWORD>","metab
obtt")

cursor = db.cursor()

csv_data=csv.reader(file('female_caregivers_basic.csv'))
for row in csv_data:
     cursor.execute("INSERT INTO
female_caregivers_basic(IndividualID, Gender, Ethnicity,
RelationshipToChild) VALUES (%s, %s, %s, %s), row)

db.commit()

cursor.close()
```

#bt20_participants_basic

```
#!/usr/bin/python

import csv
import MySQLdb
```

```
db =
MySQLdb.connect("localhost","<USERNAME>","<PASSWORD>","metab
obtt")

cursor = db.cursor()

csv_data=csv.reader(file('bt20_participants_basic.csv'))
for row in csv_data:
     cursor.execute("INSERT INTO
bt20_participants_basic(IndividualID, Gender, Ethnicity,
CaregiversID) VALUES (%s, %s, %s, %s)", row)

db.commit()

cursor.close()
```

#phenotype_yr13_female_caregivers

```
#!/usr/bin/python

import csv
import MySQLdb

db =
MySQLdb.connect("localhost","<USERNAME>","<PASSWORD>","metab
obtt")

cursor = db.cursor()

csv_data=csv.reader(file('phenotype_yr13_female_caregivers.c
sv'))
for row in csv_data:
     cursor.execute("INSERT INTO
phenotype_yr13_female_caregivers(IndividualID, Age, Height,
Weight, BMI, HipCircumference, WaistCircumference,
WaistToHipRatio, AverageSBP, AverageDBP, SubtotalBodyFat,
```

```
SubtotalLeanMass, PercentageBodyFat) VALUES (%s, %s, %s, %s,
%s, %s, %s, %s, %s, %s, %s, %s, %s)", row)

db.commit()

cursor.close()
```

#phenotype_yr1718_bt20_participants

```
#!/usr/bin/python

import csv
import MySQLdb

db =
MySQLdb.connect("localhost","<USERNAME>","<PASSWORD>","metab
obtt")

cursor = db.cursor()

csv_data=csv.reader(file('phenotype_yr1718_bt20_participants
.csv'))
for row in csv_data:
     cursor.execute("INSERT INTO
phenotype_yr1718_bt20_participants(IndividualID, Age,
Height, Weight, BMI, HipCircumference, WaistCircumference,
WaistToHipRatio, AverageSBP, AverageDBP, SubtotalBodyFat,
SubtotalLeanMass, PercentageBodyFat) VALUES (%s, %s, %s, %s,
%s, %s, %s, %s, %s, %s, %s, %s, %s)", row)

db.commit()

cursor.close()
```

#metabochip_snp_information_1

```
#!/usr/bin/python

import csv
import MySQLdb

db =
MySQLdb.connect("localhost","<USERNAME>","<PASSWORD>","metab
obtt")

cursor = db.cursor()

csv_data=csv.reader(file('metabochip_snp_information_1.csv')
)
for row in csv_data:
     cursor.execute("INSERT INTO
metabochip_snp_information_1(SNPID36, SNPID37, Chromosome,
BasePairPosition36, BasePairPosition37, NearestGene,
LocationWithinGene) VALUES (%s, %s, %s, %s, %s, %s, %s)",
row)

db.commit()

cursor.close()
```

#metabochip_snp_information_extra

```
#!/usr/bin/python

import csv
import MySQLdb

db =
MySQLdb.connect("localhost","<USERNAME>","<PASSWORD>","metab
obtt")
```

```
cursor = db.cursor()

csv_data=csv.reader(file('metabochip_snp_informationEXTRA.csv'))
for row in csv_data:
     cursor.execute("INSERT INTO metabochip_snp_information_extra(SNPID36, Allele1, Allele2) VALUES (%s, %s, %s)", row)

db.commit()

cursor.close()
```

#metabochip_snp_information_extra2

```
#!/usr/bin/python

import csv
import MySQLdb

db = MySQLdb.connect("localhost","<USERNAME>","<PASSWORD>","metabobtt")

cursor = db.cursor()

csv_data=csv.reader(file('metabochip_snp_informationEXTRA2.csv'))
for row in csv_data:
     cursor.execute("INSERT INTO metabochip_snp_information_extra(SNPID36, AfterQC) VALUES (%s, %s)", row)

db.commit()

cursor.close()
```

#metabochip_snp_associations_batch1_all

```
#!/usr/bin/python

import csv
import MySQLdb

db =
MySQLdb.connect("localhost","<USERNAME>","<PASSWORD>","metab
obtt")

cursor = db.cursor()

csv_data=csv.reader(file('Assoc-Batch1_ALL.csv'))
for row in csv_data:
     cursor.execute("INSERT INTO
metabochip_snp_associations_batch1_all(SNPID36, Dataset,
Phenotype, P_valueUncorrected, Beta_ORUncorrected,
P_valueCorrected, Beta_ORCorrected, Covariate1, Covariate2)
VALUES (%s, %s, %s, %s, %s, %s, %s, %s, %s)", row)

db.commit()

cursor.close()
```

#metabochip_snp_associations_batch2_all

```
#!/usr/bin/python

import csv
import MySQLdb

db =
MySQLdb.connect("localhost","<USERNAME>","<PASSWORD>","metab
obtt")

cursor = db.cursor()
```

```
csv_data=csv.reader(file('Assoc-Batch2_ALL.csv'))
for row in csv_data:
     cursor.execute("INSERT INTO
metabochip_snp_associations_batch2_all(SNPID36, Dataset,
Phenotype, P_valueUncorrected, Beta_ORUncorrected,
P_valueCorrected, Beta_ORCorrected, Covariate1, Covariate2,
Covariate3) VALUES (%s, %s, %s, %s, %s, %s, %s, %s, %s,
%s)", row)

db.commit()

cursor.close()
```

#metabochip_snp_associations_batch2_females

```
#!/usr/bin/python

import csv
import MySQLdb

db =
MySQLdb.connect("localhost","<USERNAME>","<PASSWORD>","metab
obtt")

cursor = db.cursor()

csv_data=csv.reader(file('Assoc-Batch2_FEMALES.csv'))
for row in csv_data:
     cursor.execute("INSERT INTO
metabochip_snp_associations_batch2_females(SNPID36, Dataset,
Phenotype, P_valueUncorrected, Beta_ORUncorrected,
P_valueCorrected, Beta_ORCorrected, Covariate1, Covariate2)
VALUES (%s, %s, %s, %s, %s, %s, %s, %s, %s)", row )

db.commit()
```

```
cursor.close()
```

#metabochip_snp_associations_batch2_males

```
#!/usr/bin/python

import csv
import MySQLdb

db =
MySQLdb.connect("localhost","<USERNAME>","<PASSWORD>","metab
obtt")

cursor = db.cursor()

csv_data=csv.reader(file('Assoc-Batch2_MALES.csv'))
for row in csv_data:
     cursor.execute("INSERT INTO
metabochip_snp_associations_batch2_males(SNPID36, Dataset,
Phenotype, P_valueUncorrected, Beta_ORUncorrected,
P_valueCorrected, Beta_ORCorrected, Covariate1, Covariate2)
VALUES (%s, %s, %s, %s, %s, %s, %s, %s, %s)", row)

db.commit()

cursor.close()
```

#metabochip_snp_associations_pre2_all

```
#!/usr/bin/python

import csv
import MySQLdb
```

```
db =
MySQLdb.connect("localhost","<USERNAME>","<PASSWORD>","metab
obtt")

cursor = db.cursor()

csv_data=csv.reader(file('Assoc-Merged_ALL_uni.csv'))
for row in csv_data:
     cursor.execute("INSERT INTO
metabochip_snp_associations_pre2_all(SNPID36, Dataset,
Phenotype, P_valueUncorrected, Beta_ORUncorrected) VALUES
(%s, %s, %s, %s, %s)", row)

db.commit()

cursor.close()
```

#metabochip_snp_associations_pre3_all

```
#!/usr/bin/python

import csv
import MySQLdb

db =
MySQLdb.connect("localhost","<USERNAME>","<PASSWORD>","metab
obtt")

cursor = db.cursor()

csv_data=csv.reader(file('Assoc-Merged_ALL_multi.csv'))
for row in csv_data:
     cursor.execute("INSERT INTO
metabochip_snp_associations_pre3_all(SNPID36, Dataset,
Phenotype, P_valueCorrected, Beta_ORCorrected, Covariate1,
```

```
Covariate2, Covariate3) VALUES (%s, %s, %s, %s, %s, %s, %s,
%s)", row)

db.commit()

cursor.close()
```

#metabochip_snp_associations_pre2_females

```
#!/usr/bin/python

import csv
import MySQLdb

db =
MySQLdb.connect("localhost","<USERNAME>","<PASSWORD>","metab
obtt")

cursor = db.cursor()

csv_data=csv.reader(file('Assoc-Merged_FEMALES_uni.csv'))
for row in csv_data:
     cursor.execute("INSERT INTO
metabochip_snp_associations_pre2_females(SNPID36, Dataset,
Phenotype, P_valueUncorrected, Beta_ORUncorrected) VALUES
(%s, %s, %s, %s, %s)", row)

db.commit()

cursor.close()
```

#metabochip_snp_associations_pre3_females

```
#!/usr/bin/python

import csv
import MySQLdb
```

```
db =
MySQLdb.connect("localhost","<USERNAME>","<PASSWORD>","metab
obtt")

cursor = db.cursor()

csv_data=csv.reader(file('Assoc-Merged_FEMALES_multi.csv'))
for row in csv_data:
     cursor.execute("INSERT INTO
metabochip_snp_associations_pre3_females(SNPID36, Dataset,
Phenotype, P_valueCorrected, Beta_ORCorrected, Covariate1,
Covariate2) VALUES (%s, %s, %s, %s, %s, %s, %s)", row)

db.commit()

cursor.close()
```

APPENDIX F – MetaboBTT README

MetaboBTT Database

What is the MetaboBTT Database?

It is a MySQL database that stores project-specific data related to a current study in the field of genetics. The database is easily accessible and queryable by all members of the research group via a user-friendly web interface.

Where can the database be accessed?

The database can be accessed via a web interface available at http://www.bioinf.wits.ac.za/software/metabobtt and must be accessed with a username and password.

What data does the database contain?

The MetaboBTT Database houses phenotype, SNP annotation and association analysis data from an ongoing project focused on identifying risk factors for cardiometabolic disease in South Africans. The data is from participants and their female caregivers from the Birth to Twenty (Bt20) cohort and DNA samples were genotyped using the Metabochip.

Each individual recorded in the database has a unique Individual ID (with a suffix 'C' or 'CG' for the Bt20 participants and female caregivers, respectively).

The Bt20 cohort is a longitudinal cohort consisting of data collected at multiple time points since its inception. **PHENOTYPE DATA** currently present in the

database is from the year 17/18 data collection time point for the Bt20 participants and the year 13 data collection time point for the female caregivers. Tables have been constructed for other data collection time points for the Bt20 participants (year 5, year 7, year 9/10, year 11/12, year 13, year 14, year 15, year 16, year 19, year 20) and can be populated with the data when available. Phenotype data only exists in the database for individuals with available genotype data and includes:

- Gender (1 = males, 2 = females)
- Ethnicity (2 = Africans)
- Relationship to Child (1 = Mother , 2 = Aunt, 3 = Grandmother, 4 = Sister, 5 = Other) *[only relevant to the female caregivers]*
- Caregiver's ID *[only relevant to the Bt20 participants]*
- Age (in years)
- Height (in metres)
- Weight (in kilograms)
- Body Mass Index (BMI) (in kg/m^2)
- Hip Circumference (in metres)
- Waist Circumference (in metres)
- Waist to Hip Ratio
- Average Systolic Blood Pressure (SBP) (in millimetres of mercury) *[average of the 2^{nd} and 3^{rd} of three readings taken]*
- Average Diastolic Blood Pressure (DBP) (in millimetres of mercury) *[average of the 2^{nd} and 3^{rd} of three readings taken]*
- Subtotal Body Fat (in grams)
- Subtotal Lean Mass (in grams)
- Percentage Body Fat (%)

SNP ANNOTATION/METABOCHIP DATA exists for all 196725 SNPs on the Metabochip and includes:

- Build 36 SNPID

- Build 37 SNPID
- Chromosome
- Build 36 Base Pair Position
- Build 37 Base Pair Position
- Nearest Gene *[refers to the gene(s) in which the SNP lies or is intergenic to]*
- Location Within Gene (CODING, COMPLEX, INTERGENIC, INTRON or UTR)
- Allele 1
- Allele 2
- After QC (BOTH = remained in both the Bt20 participant and female caregiver datasets after QC, BATCH1 = remained in only the female caregiver dataset after QC, BATCH2 = remained in only the Bt20 participant dataset after QC)

All **ASSOCIATION ANALYSIS DATA** for the available phenotypes under investigation are recorded and includes:

- Build 36 SNPID
- Dataset* (BATCH1_ALL, BATCH2_ALL, BATCH2_FEMALES, BATCH2_MALES, MERGED_ALL, MERGED_FEMALES)
- Phenotype
- Uncorrected P-value
- Uncorrected Beta/OR
- Corrected P-value *[corrected for the covariates listed]*
- Corrected Beta/OR *[corrected for the covariates listed]*
- Covariate 1
- Covariate 2
- Covariate 3

**The datasets available can be analysed as individual or merged datasets – results for all possible scenarios are recorded.*

BATCH1_ALL = yr13 female caregivers
BATCH2_ALL = all yr 17/18 Bt20 participants

BATCH2_FEMALES = female yr 17/18 Bt20 participants

BATCH2_MALES = male yr 17/18 Bt20 participants

MERGED_ALL = yr13 female caregivers and Bt20 participants merged

MERGED_FEMALES = yr13 female caregivers and female yr17/18 Bt20 participants merged

The **GENOTYPE DATA** has undergone an extensive quality control (QC) process. The data exists as cleaned/QC'ed binary PLINK format files *(.bed/.bim/.fam*) for both the Bt20 participants and the female caregivers and these can be accessed on request.

What can users do?

Users can access the database from the user interface to generate **summary statistics** (basic and complex counts and average/minimum/maximum) on the phenotype data, **download** relevant phenotype, Metabochip and association analysis data that match certain user-supplied criteria and get information on how to work with the genotype files in PLINK.

1) File uploads

When specifying a list of individuals/SNPs etc., the uploaded file must be a text file containing a list of Individual IDs/SNPIDs etc. each on a separate line.

2) Specifying criteria

When phenotype criteria can be specified, up to three criteria can be added (in the form <PHENOTYPE> <OPERATOR> <VALUE>).

e.g. Height (m) >= 1.6

3) Output

Summary statistics will always be printed to the screen. Phenotype, Metabochip and association analysis data download outputs can be printed to the screen ('Print to screen') or downloaded as a CSV file ('Save to File').

Additional features

- A reset button is present on each page to clear all previously selected fields.
- Each page has a footer with links to useful websites (NCBI and Ensembl) and the home and parent pages.

Contact

To request access to the database or for any queries pertaining to the database or data contained in the database can be directed to Liesl Hendry (Sydney Brenner Institute for Molecular Bioscience/University of the Witwatersrand) at lieslmaryhendry@gmail.com.

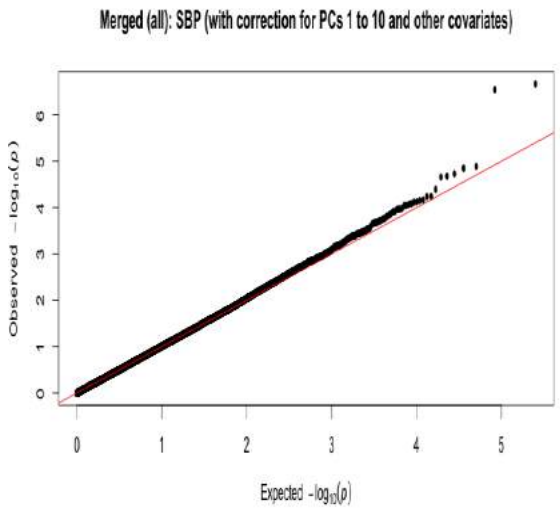
APPENDIX G – Post-analysis QC: Q-Q plots and genomic inflation factors

MAIN ANALYSIS:

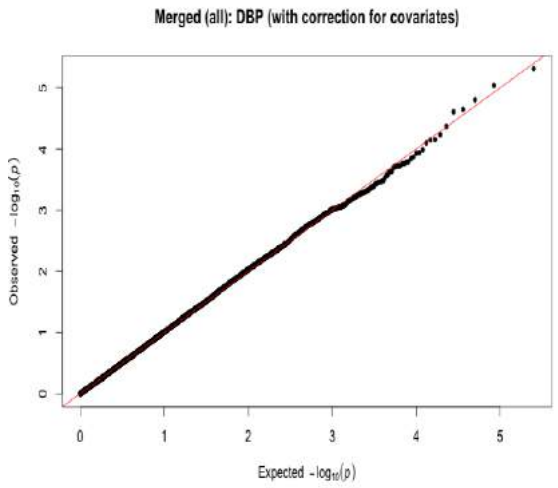
All merged dataset:

Phenotype	Before correction for PCs	After correction for PCs
SBP	λ = 0.9982	λ = 0.9997
DBP	λ = 0.9988	NA
High versus normal/low	λ = 0.9977	λ = 1.0117

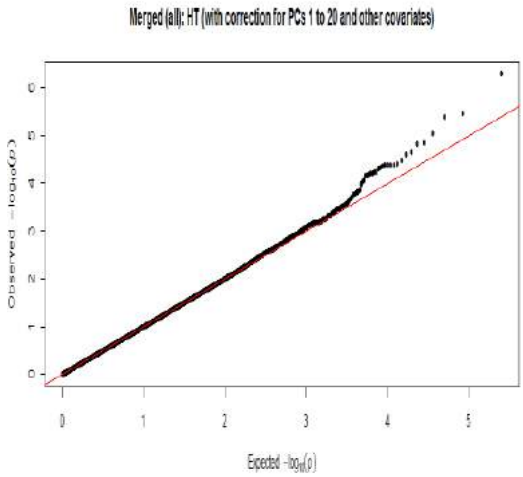
SBP (corrected for first 10 PCs):



DBP (no correction for PCs):



High versus normal/low BP (correction for first 20 PCs):

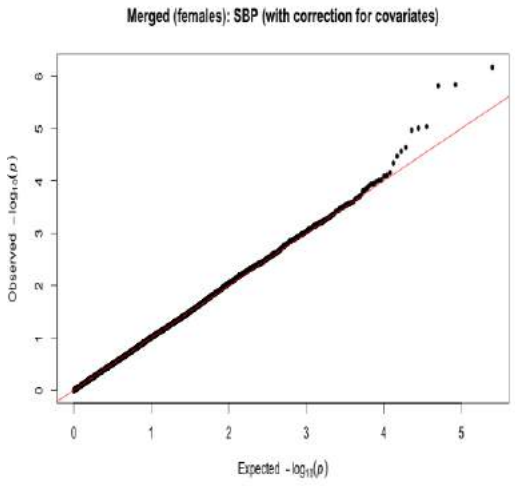


SEX-SPECIFIC ANALYSIS:

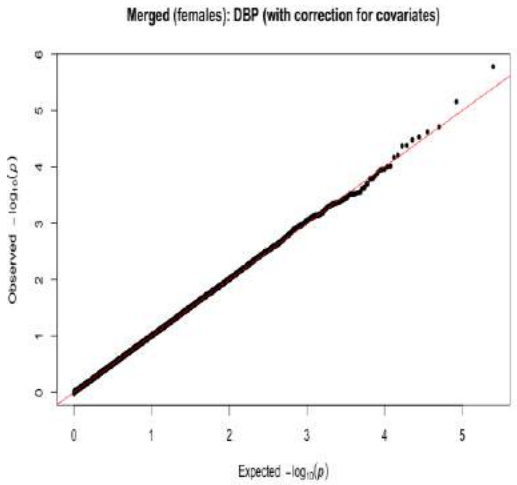
Merged female dataset:

Phenotype	Before correction for PCs	After correction for PCs
SBP	λ = 1.0120	NA
DBP	λ = 1.0073	NA
High versus normal/low	λ = 0.9930	λ = 1.0117

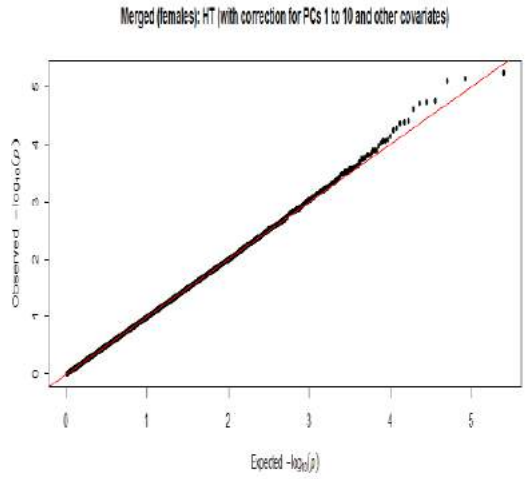
SBP (no correction for PCs):



DBP (no correction for PCs):



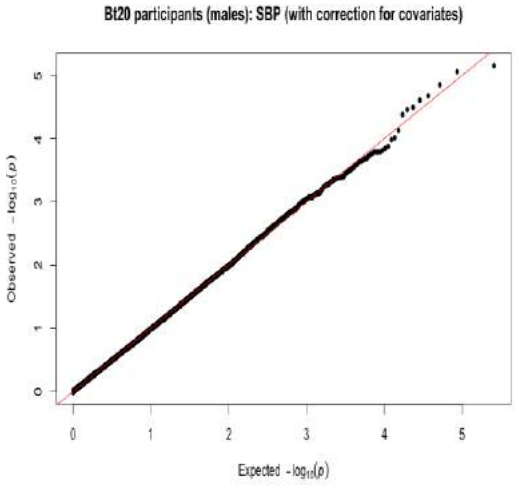
High versus normal/low BP (correction for first 10 PCs):



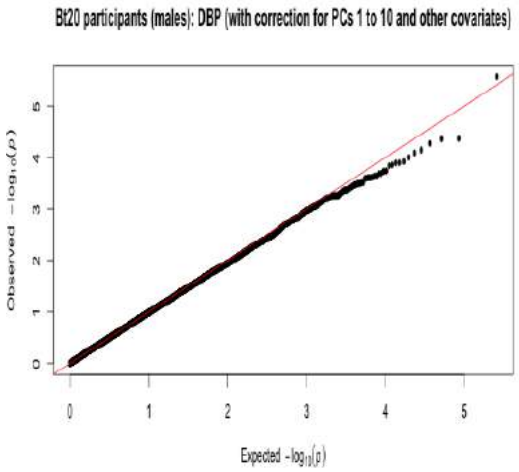
Bt20 participant dataset (males only):

Phenotype	Before correction for PCs	After correction for PCs
SBP	λ = 1.9769	NA
DBP	λ = 1.0009	λ = 1.0009
High versus normal/low	λ = 0.9261	λ = 0.9718

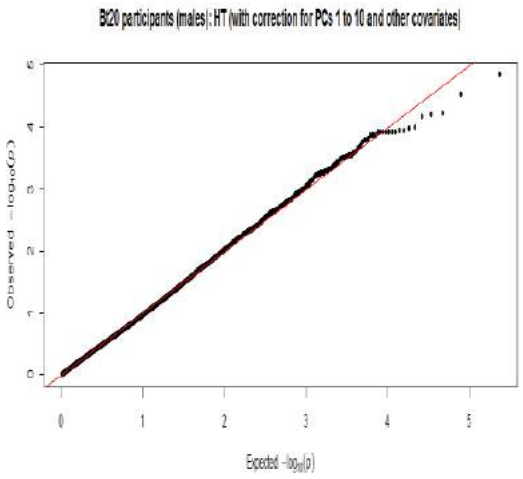
SBP (no correction for PCs):



DBP (correction for first 10 PCs):



High versus normal/low BP (correction for first 10 PCs):

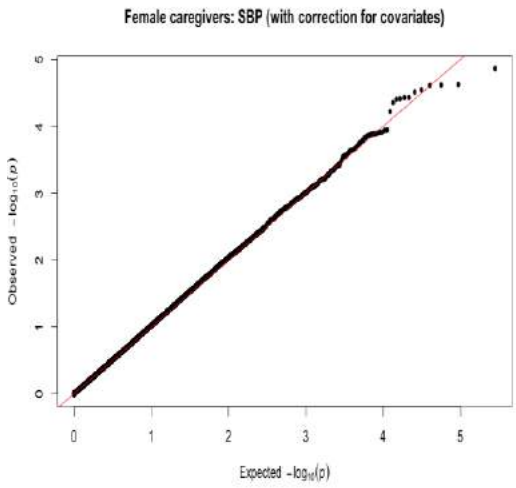


AGE-SPECIFIC ANALYSIS:

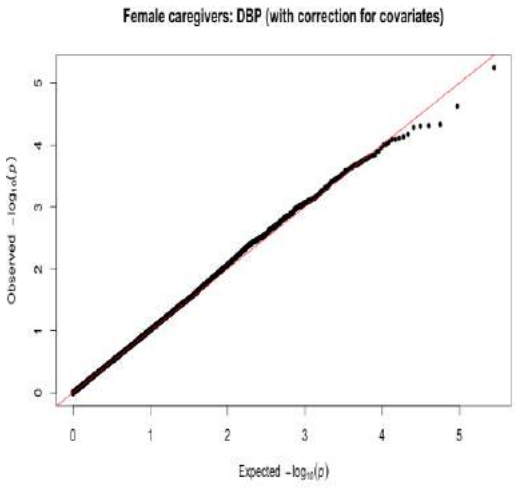
Female caregiver dataset:

Phenotype	Before correction for PCs	After correction for PCs
SBP	λ = 1.0249	NA
DBP	λ = 1.0297	NA
High versus normal/low	λ = 1.0469	NA

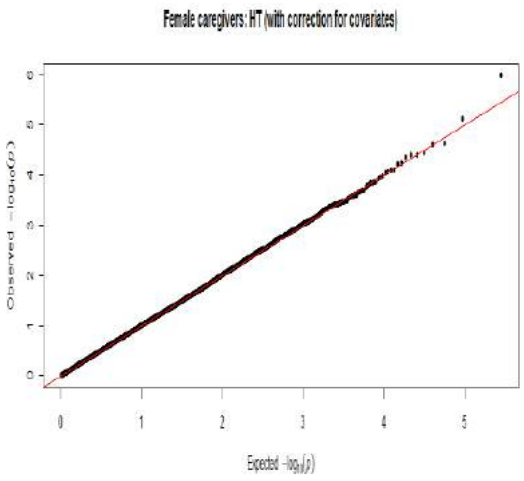
SBP (no correction for PCs):



DBP (no correction for PCs):



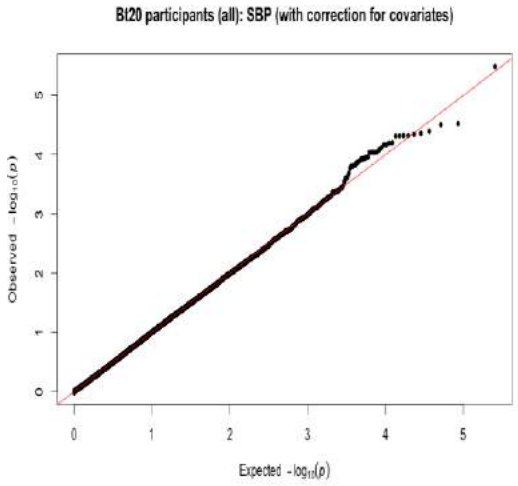
High versus normal/low BP (no correction for PCs):



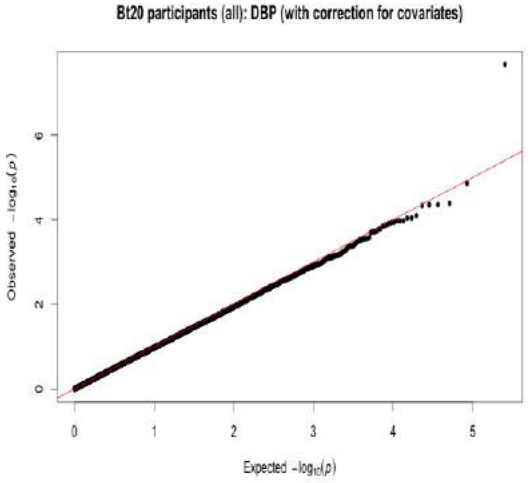
Bt20 participant dataset:

Phenotype	Before correction for PCs	After correction for PCs
SBP	λ = 0.9591	NA
DBP	λ = 0.9623	NA
High versus normal/low	λ = 0.9345	NA

SBP (no correction for PCs):



DBP (no correction for PCs):



High versus normal/low BP (no correction for PCs):

