THE IDENTIFICATION OF GENETIC MARKERS OF OBESITY RISK IN A SOUTH AFRICAN BLACK POPULATION

By

Venesa Pillay

Supervisors:

Dr. Zané Lombard

Associate Professor Nigel Crowther

Associate Professor Himla Soodyall

A THESIS

Submitted to the Faculty of Health Sciences, University of the Witwatersrand, in fulfillment of the requirements for the degree of

Doctor of Philosophy

JOHANNESBURG, SOUTH AFRICA

15 July, 2016

Declaration:

I Venesa Pillay, declare that this thesis is my original work. Where there has been contribution from other people, this has been duly acknowledged. This thesis is being submitted for the degree of Doctor of Philosophy at the University of the Witwatersrand, Johannesburg, South Africa. It has not been submitted before for any degree or examination at this or any other university.

Name: Venesa Pillay

Date: _____day of _____2016

Abstract

Obesity is a common risk factor for non-communicable disease and is most often described in relation to body mass index (BMI), although it is not the best predictor of body composition. Heritability estimates of obesity (predominantly based on Europeans) suggest that there is a significant genetic component. The latest genome-wide association study (GWAS) of obesity-related traits have identified over hundred loci contributing to BMI alone. These findings have yet to be robustly replicated in African populations.

The aim of this study was to assess whether risk loci previously associated with body composition in European populations showed a similar trend in a South African black population by:

i. Replicating the association of six SNPs previously linked to adult BMI, in an adolescent cohort (the Birth to twenty cohort (Bt20); N=990).
ii. Performing a replication and fine-mapping study by genotyping participants within this same cohort using the Metabochip (N=2273).
iii. Estimating the narrow-sense heritability (h^2) of body composition measures in this cohort.

In the candidate gene analysis, three of the SNPs tested were significantly associated with BMI, and showed a consistent (albeit smaller) directional effect to that observed in non-African cohorts.

Results from the replication and fine-mapping analyses reaffirmed that several loci including *SEC16B*, *NEGR1*, *FTO*, *TMEM18*, *WARS2, NRXN3, and SP110* previously found to be associated with body composition, were similarly associated in this African cohort. The associated loci were replications of previous findings but they do not involve the same SNPs observed in European, African-American and Asian populations. This suggests that GWAS-identified variants of body composition are tagged by different SNPs in an African cohort. An important finding of this study was the observation of ten cross-phenotype associations.

Heritability estimates for most of the body composition phenotypes were similar to estimates derived for European populations, albeit trending towards the upper limits of such heritability measures.

This study highlights the importance of assessing genetic factors for body composition in urban black South Africans. Results from this study suggests that more in-depth genomic studies in larger cohorts will reveal novel SNP associations for body composition and insight into the aetiology of obesity.

Acknowledgements

This body of work is dedicated to my family. My father, who gave us the gift of education and instilled in me an enquiring mind, even though you are not here with us today, your presence will always be felt. My mother, sister and grandmother who are a continuous inspiration to me, and for their unconditional love and support. Most of all, this thesis is dedicated to my "Pillars of Strength", my husband Veeran and my children Vidhar and Kairav who have encouraged and supported me in ways words could not express. Thank you for giving me the time to pursue my dreams and complete this thesis, I love you guys and thank you for all your sacrifices.

There are many others who have influenced the outcome of this thesis whom I would like to acknowledge:

To my supervisors: Dr Zané Lombard, whose encouragement and support has been both phenomenal and inspirational. Thank you for giving me the opportunity to develop and think like an independent researcher. Prof Nigel Crowther and Prof Himla Soodyall, thank you for thoughtful insights, guidance and support throughout the preparation of this manuscript.
My friends and colleagues at SBIMB and Human Genetics who have provided a lot of support and motivation when I needed you guys the most, thanks guys, I am grateful to all of you for helping me to get through this. Thank you especially to Phelelani Mpangase, Liesl Hendry, Thandiswa Ngcungcu, Ayesha Ismail, Michelle Ungerer, Andrew May, Jacqueline Frost, Angela Hobbs, Andrew Ndlovu, Richard Munthali, Drithi Sengupta, Ananyo Choudury, Cassandra Soo and Shaun Aron.
Thank you Prof. Shane Norris and colleagues at the DPPHRU for the use of the phenotype data in this study. I am also very grateful to Prof. Michele Ramsay for your support and guidance as well as access to the DNA samples and genotype data.

I am thankful to the funding bodies that have provided funding for me along this journey, especially the National Institute of Health (NIH) Non-communicable Disease Research (NCD) research programme and the National Research Foundation (NRF).

Preface

Problem identification

The completion of the human genome project (Mcpherson et al., 2001, Collins et al., 2003) has resulted in the development of new methodologies in human genomics. The development of these new technologies has created opportunities for generating huge public data sharing initiatives allowing for multiple collaborations and projects that aim to describe genomic variation globally. The International HapMap project (Gibbs et al., 2003) followed by the 1000 Genomes Project (1000 Genomes Project, 2010) aimed to elaborate on what we know about human variation. These projects allowed for the advancement of new research tools such as fine-mapping (Immunochip and Metabochip) chip array designs and invoked the follow-up of whole genome sequencing (WGS) applications. These new techniques impacted on the discovery of many genomic regions associated with disease loci or quantitative traits (Adeyemo and Rotimi, 2014). However, most of these studies have been performed in individuals of European ancestry with other ethnic populations being understudied. African-based initiatives such as Human Heredity and Health in Africa (H3Africa) and, more specifically relevant to this project, African Wits-INDEPTH Partnership for the GENomic study of body composition and cardiometabolic risk (AWI-GEN) (described below) aim to shed more light on genomic regions associated with cardiometabolic diseases in African populations, particularly because obesity rates are on the rise on the African subcontinent, especially amongst black females (Dalal et al., 2011, Ng et al., 2014).

As it is not known which genetic variants are associated with body fat mass variation in the South African black population, this research project aims to address this question by analyzing generated genotype-phenotype information and conducting a genetic association study on several body composition phenotypes. If the genetic risk factors contributing to obesity can be identified, perhaps it will give us greater insights into the underlying biological pathways associated with this phenotype, allowing the development of interventions to prevent the full-onset of the disease. It is also possible that such studies will allow us to identify subjects early in the life course who are at high risk of developing obesity and in whom early interventions can be used to prevent excess body fat accumulation. An understanding of the molecular pathways that control fat deposition may also allow the development of techniques to block sub-cellular lipid accumulation in organs such as the liver and muscle, where fat deposition can have major metabolic consequences.

The Birth to Twenty cohort (Bt20) from Soweto in Gauteng (refer to Fig. 1.) was used as a source of biological samples for this study. This cohort is a rich longitudinal resource containing data collected on several phenotypes relevant to obesity and body composition together with DNA samples. The main proposed study will generate a dataset (flagship

project) as part of a large genomics project partly funded by the H3Africa initiative. H3Africa is the culmination of a partnership among the African Society of Human Genetics (AfSHG), the Wellcome Trust (WT), and the National Institutes of Health (NIH). H3A's mandate is to support genomics research related to human diversity and disease biology that will be of scientific and clinical relevance to African populations (Consortium et al., 2014). The AWI-GEN project falls under the umbrella of H3A and focuses on elucidating the genomic and environmental risk factors for cardiometabolic disease in Africans. One of the aims of the AWI-GEN project has been to assess if known genetic risk factors for obesity are common to both European and African populations.

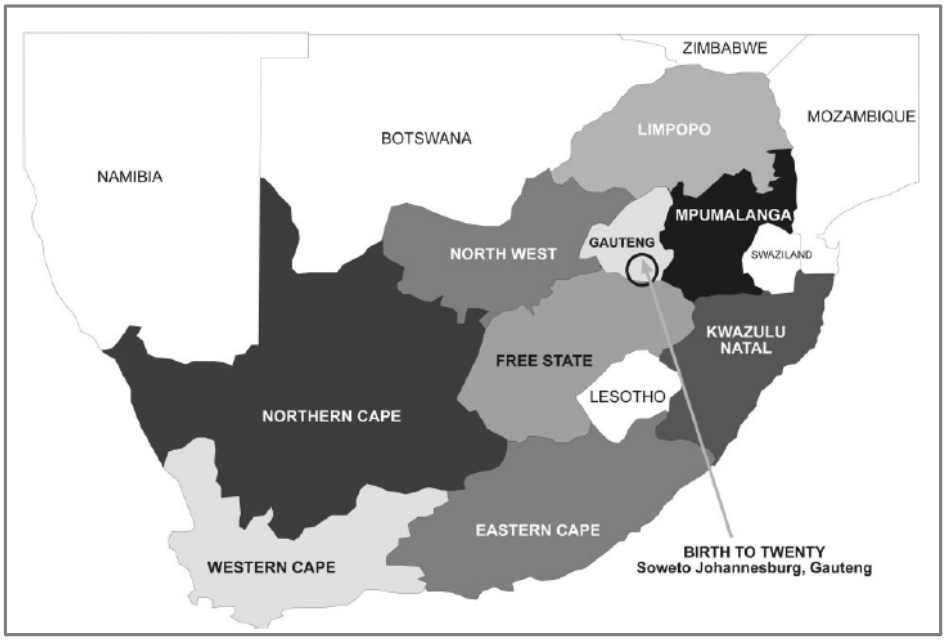


Figure 1. Map of South Africa showing the geographical location of the Birth to Twenty cohort in Gauteng province (Richter et al., 2007)

Chapter 1 of this thesis provides an overview of the literature highlighting the key aspects of the obesity phenotype including a description of the causes and contributors to obesity. It also includes an overview of the genetics of obesity, which covers monogenic forms of obesity, candidate gene studies and genome wide association studies (GWAS) plus a summary of African specific studies.

Before embarking on the flagship project, the feasibility and power to detect GWAS-derived Eurocentric associations was assessed in an African population by analyzing a previously generated dataset in a candidate-gene approach, which was termed a "pilot study". This involved the analysis of six SNPs genotyped on a BeadExpress platform on a smaller adolescent subset of the Bt20 (N=990) cohort. This was an extension of a previous study conducted by (Lombard et al., 2012) where they assessed 44 SNPs in six candidate genes

linked to the appetite regulation pathway. The published pilot study (Pillay et al., 2015) is discussed in Chapter 2 in the format of an extended publication.

Chapter 3 describes the flagship "Metabochip study" and forms the main body of this study. In this Chapter, genetic associations with body composition are reported and described. Chapter 4 focuses on the additive genetic contribution to the variance of the phenotypes used in the study by estimating narrow-sense heritability (h^2) using the genotype data generated from the Metabochip study. Heritability estimates of body composition phenotypes in African populations are limited.

Please note that the use of the term "Bantu" is in reference to various ethno-linguistic groups spoken in South Africa as a subgroup of Niger-Kordofanian languages (Schlebusch et al., 2012). Use of the term "Coloured", which may be used intermittently in this thesis, refers to a South African community of mixed-ancestry (Lombard, 2008, Chimusa et al., 2013). The term "Indian" as described in Table 1.1 refers to South Africans of Indian descent, many of whom are descendants of migrants from India and the term "White" refers to South Africans of European descent.

Publication and Presentations

Publications and published conference proceedings*:*

1. Pillay, V., Crowther, N. J., Ramsay, M., Smith, G. D., Norris, S. A., & Lombard, Z. (2015). Exploring genetic markers of adult obesity risk in black adolescent South Africans—the Birth to Twenty Cohort. *Nutrition & diabetes*, 5(6), e15.
2. Wits Research Day Abstracts, published in South African Journal of Infectious Disease (http://www.sajei.co.za/index.php/SAJEI/issue) (Vol 30, No 2 (2015).

Student's contribution to the paper (1): Design of the paper, data cleaning and analysis, writing of the manuscript (VP wrote the first complete draft).

Research Output: Conference Proceedings

Poster Presentations	
Presentation detail	Conference Details
1. Pillay V, Crowther NJ, Soodyall H & Lombard Z Genetic Risk factors associated with Obesity in the South African black population - Birth to Twenty Cohort	2nd YRF meeting, School of Public Health,WITS, (October 2013)
2. Pillay V, Crowther NJ, Soodyall H & Lombard Z. Genetic Risk factors associated with Obesity in the South African black population - Birth to Twenty Cohort	15th South African Society for Human Genetics Congress, Maslow, Johannesburg (October 2013) *Awarded the Whitehead Scientific Travel Bursary to attend conference*
3. Pillay V, Hendry LM, Choudhury A, Crowther NJ, Soodyall H, Ramsay M, Norris SA & Lombard Z. Characterizing the genetic risk for obesity/body composition in a black South African cohort.	6th Meeting of H3Africa Consortium meeting, Zambia, (May 2015). *Awarded travel scholarship to attend meeting in Zambia.* *Awarded a prize for poster presentation.*

4. Pillay V, Hendry LM, Choudhury A, Crowther NJ, Soodyall H, Ramsay M, Norris SA & Lombard Z. Characterizing the genetic risk for obesity/body composition in a black South African cohort	3rd YRF meeting, Prinsloo medical campus, University of Pretoria, (15th August, 2015)
Oral presentations	
Presentation detail	Conference Details
1. Pillay V, Crowther NJ, Soodyall H & Lombard Z. Identification of Genetic Markers of Obesity risk in a South African black population.	Faculty of Health Sciences Research Day, University of the Witwatersrand (Sep 2014) *Won best oral presentation within the Diseases of Lifestyle track*
2. Pillay V, Crowther NJ, Soodyall H & Lombard Z. The Metabochip as a tool for identifying genetic markers of obesity risk in the South African black population.	16th South African Society for Human Genetics Congress, KleinKaap Boutique Hotel, Centurion, Pretoria, (16-19th August, 2015) *Won Best Oral Presentation* *Awarded SASHG travel Bursary*

Fellowship Awards:

1. National Institute of Health (NIH), Non-communicable Disease (NCD) Training Programme PhD scholarship, 2012-2015.
2. National Research Foundation grant-holder linked bursary, 2012-2014.
3. National Research Foundation extension bursary (2015).

Travel Awards:

1. (2013) Whitehead scientific travel grant to attend SASHG 2013.
2. (2013) Wellcome Trust travel scholarship to attend workshop entitled “Genomic Epidemiology in Africa.”
3. (2014) GEoCoDE Exchange programme travel award to University of Bristol, United Kingdom.

Table of Contents

List of Tables

Chapter 1:

Chapter 2:

Chapter 3:

Chapter 4:

List of Figures

Chapter 1:

Chapter 2:

Chapter 3:

Abbreviations and term definitions

1000G	1000 genomes project
A1	tested allele (minor allele by default)
AA	African ancestry
ACE	*angiotensin converting enzyme*
ADIPOQ	*adiponectin C1Q*
ADRB2	*beta(2)-adrenergic receptor*
ADRB3	*beta(3)-adrenergic receptor*
AfSHG	African society of human genetics
AGRP	*agouti-related peptide*
APOH	*apolipoprotein H*
AR	*androgen receptor*
ASW	African ancestry in Southwest USA (HapMap)
AWIGEN	Africa, Wits-INDEPTH Partnership for GENomic studies
BDNF	*brain-derived neurotrophic factor*
BDNFOS	*brain-derived neurotrophic factor opposite strand*
BMC	bone mineral content
BMI	body mass index
BP	base-pair position
BRE	*brain and reproductive organ expressed (TNFRSF1AModulator)*
Bt20	birth to twenty cohort
4C	circular chromosome conformation capture
C7orf31	*chromosome 7 open reading frame 31*
C12orf5	*chromosome 12 open reading frame 5*
CAPN10	*calpain-10*
CAD	coronary artery disease
CART	*cocaine- and amphetamine-regulated transcript*
CD36	cluster of differentiation 36
CDCV	common disease common variant hypothesis
CDKAL1	*CDK5 regulatory subunit associated protein 1-like 1*
CELSR2	*cadherin, EGF LAG seven-pass G-type receptor 2*
CEPT	*cholesteryl ester transfer*
CEU	Utah residents with Northern and Western European ancestry
CHARGE	Cohorts for Heart and Aging Research in Genome Epidemiology
CHD	coronary heart disease
CHR	chromosome
CI	confidence intervals
cm	centimetre
CNTNAP5	*contactin associated protein-like 5*
CNV	copy number variant
COBLL1	*Cordon-Bleu WH2 repeat protein-like1*

CT	computerised tomography
CTNNBL1	*catenin beta-like 1*
DIAGRAM	DIAbetes Genetics Replication And Meta-analysis
DOH	Department of Health
DOHAD	Developmental Origins of Health and Disease
DPHRU	Developmental Pathways for Health Research Unit, Wits University
DNA	deoxyribonucleic acid
DRD2	*dopamine receptor D2*
DRD4	*dopamine receptor D4*
DXA	dual-energy x-ray absorptiometry scans
DXH34	*DEAH (Asp-Glu-Ala-His) box polypeptide 34*
ENPP1	*ecto-nucleotide pyrophosphatase/phosphodiesterase 1*
EMP1	point-wise estimate of significance
EMP2	family-wise estimate of significance
ER	endoplasmic reticulum
ETV5	*Ets variant 5*
FDR	false discovery rate
FFM	fat free mass
FID	family identifier (family name)
FM	fat mass
FTO	*fat mass and obesity-associated locus*
GALNT10	*polypeptide N-acetylgalactosaminyltransferase 10*
GCTA	genome-wide complex trait analysis
GIANT	**G**enetic **I**nvestigation of **AN**thropometric **T**raits
GBD	global burden of disease
GC	genomic control
GCTA	genome-wide complex trait analysis
GI	genomic inflation factor
GNPDA2	*glucosamine-6-phosphate deaminase 2*
GP2	*glycoprotein 2*
GRM	genetic relationship matrix
GW	genome-wide
GWAS	genome wide association study
h^2	narrow-sense heritability
h^2_g	narrow-sense heritability determined from all GWAS SNPs
h^2_{gwas}	narrow-sense heritability determined from GW-signficant SNPs
H^2	broad-sense heritability
H3A	human heredity health in Africa
HC	hip circumference
HDL	high density lipoprotein
HIC	high-income countries
HREC	human research ethics committee

HWE	Hardy–Weinberg equilibrium
HXE	estimates of homozygosity on the X-chromosome
IBD	identity by descent
IBS	identity by state
ICBP	International Consortium for Blood Pressure
IID	individual identifier (sample name)
IQR	interquartile range
IRS1	*insulin receptor substrate 1*
IRX1	*iroquois-class homeodomain protein 1*
IRX3	*iroquois-class homeodomain protein 3*
KCTD15	*potassium channel tetramerization domain containing 15*
KLHL32	*kelch-like family member 32*
KLF9	*Kruppel-like factor 9*
KNCMA1	*potassium-channel gene homologue to Slo1*
L95	lower confidence interval
LCT	lactase
LD	linkage disequilibrium
LDL	low density lipoprotein
LEP	*leptin*
LEPR	*leptin receptor*
LHX2	*lim homeobox 2*
LM	lean mass
LMIC	low and middle-income countries
LMM	linear mixed models
LWK	Luhya, Kenya
LYPAL2	*lipoprotein(a)-like 2*
LYPLAL1	*lysophospholipase-Like 1*
MAF	minor allele frequency
MAF	*v-maf avian musculoaponeurotic fibrosarcoma oncogene homolog*
MAGIC	Meta-Analyses of Glucose and Insulin-related traits Consortium
Mb	megabase(s)
MC3R	*melanocortin 3 receptor*
MC4R	*melanocortin 4 receptor*
MHC	major histocompatibility complex
MI	myocardial infarction
MIR148-A	*microRNA 148A*
MKK	Maasai, in Kinyawa, Kenya
MLMA	mixed linear model association
MRP522	*mitochondrial ribosomal protein 522*
MSRA	*methionine sulfocide reductase A*
MTA	material transfer agreement
MTCH2	*mitochondrial carrier 2*

MtDNA	mitochondrial DNA
MTNR1B	*melatonin receptor type 1 B*
MRI	magnetic resonance imaging
N	sample size
NAN	SNPs not called in Genome Studio
NBEAL1	*neurobeachin like 1*
NCD	noncommunicable disease
NEGR1	*neuronal growth factor 1*
NFE2L3	*nuclear factor erythroid 2-related factor 3*
NHLS	National Health Laboratory Service
NCD	Non-Communicable Disease research program
NIH	National Institutes of Health
NMISS	number of non-missing individuals in PLINK analysis
NRF	National Research Foundation
NRXN3	*neurexin 3*
P-value	asymptotic P value for test statistic
P_{adj}	P- value adjusted for covariates
PAGE	Population Architecture using Genomics and Epidemiology
PAX5	*paired box 5*
PC	principle components
PCA	principle component analysis
PCSK1	*pro-hormone convertase subtilisin/kexin type 1 (PCSK1)*
PFM	percentage fat mass
POMC	*pro-opiomelanocortin*
PPP1R3B	*protein phosphatase 1 regulatory subunit 3B*
PPARG	*peroxisome proliferator activated receptor gamma*
PSRC1	*proline/serine-rich coiled-coil 1*
QC	quality control
QQ	quantile-quantile
QT	start of the Q wave and end of the T wave
QT-IGC	QT-interval international GWAS consortium
RNA	ribonucleic acid
RETN	*resistin*
RREB1	*ras responsive element binding protein 1*
SA	South Africa/n
SANHANES-1	South African National Health and Nutrition Examination Survey
SADHS	South African Demographic and Health Survey
SAT	subcutaneous adipose tissue
SD	standard deviation
SDCCAG8	*serologically defined colon cancer antigen 8*
SEB	southeastern Bantu speakers
SEC16B	*SEC16 homolog B*

SGIP1	*SH3-containing GRB2-likeprotein 3-interacting protein 1*
SH2B1	*SH2B adaptor protein 1*
SLC17A4	*solute carrier family 17*
SNP	single nucleotide polymorphism
SORT1	*sortilin-1*
SP110	*SP110 nuclear body*
SPRY2	*sprouty homolog 2 (Drosophila)*
SSA	sub-Saharan African
STK33	*serine/threonine kinase 33*
SWB	south-western Bantu speakers
T2D	type 2 diabetes
TBW	total body water
TFA2B	*transcription factor AP-2 beta (Activating Enhancer Binding protein 2 Beta)*
TMEM18	*transmembrane 18*
TNF	*tumour necrosis factor*
TFAM	transposed fam format
TPED	transposed ped format
TRPM7	*transient receptor potential melastatin*
U95	lower confidence intervals
US	United States of America
VAT	visceral adipose tissue
WARS2	*mitochondrial form of tryptophanyl-tRNA synthetase*
WC	waist circumference
WGS	whole genome sequencing
WHO	World Health Organisation
WHR	waist-to-hip-ratio
WT	Wellcome Trust
YRI	Yoruba, Nigeria
ZFYVE9	*Zinc finger FYVE-type containing 9*

Chapter 1

Literature Review

1. Overview of the chapter

Obesity has many negative effects on health (Guh et al., 2009) and is on the increase in low and middle-income countries (LMIC) such as South Africa (SA), particularly in the South African black population. A recent South African National Health and Nutrition Examination Survey (SANHANES-1) (Shisana et al., 2013) has shown that the prevalence of overweight and obesity was significantly higher in females than males (24.8% and 39.2% compared to 20.1% and 10.6% for females and males, respectively), with the highest figures for obesity being recorded in the SA black female population. Obesity is a complex disease in that both genetic and environmental factors play a role. The heritability of body mass index (BMI) an indicator of of overweight and obesity has been shown from studies in European populations to range between 40-70% (Bodurtha et al., 1990, Walley et al., 2009). Most genome-wide association studies (GWAS) conducted on obesity have been carried out in mainly European study cohorts. Figure 1.1 shows that almost 90% of GWAS have been conducted in European/Asian populations with a paucity of African-embedded research (~7%). This has been mainly localised to African populations thare are admixed and have arisen from migration events such as the African diaspora, thus reducing their genetic variability. African populations, in particular sub-Saharan African populations, remain understudied for many of these traits. More importantly, given that sub-Saharan African populations harbour the greatest amount of genetic diversity (Ramsay et al., 2011, Schlebusch et al., 2012, Mboowa, 2014, 1000 Genomes Project, 2015) and have lower levels of linkage disequilibrium (Teo et al., 2010), exploring this variation would be a very useful contribution towards defining the causative variants for obesity and related phenotypes.

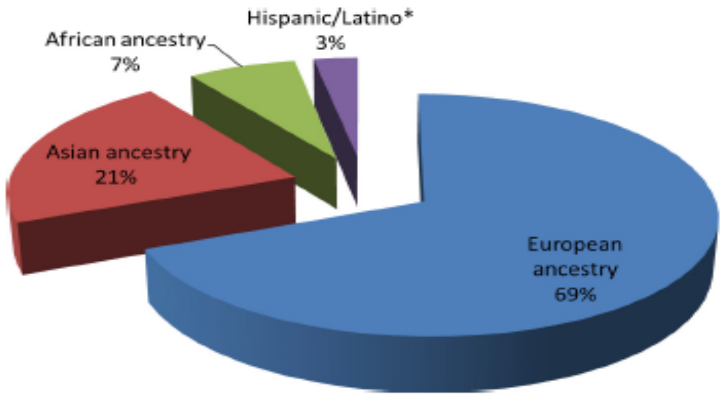


Figure 1.1 **Breakdown of GWAS recorded in GWAS catalogue**
Breakdown of GWAS recorded in GWAS catalogue (accessed- 11th December 2013 http://www.genome.gov/26525384) according to ancestry of populations from whom DNA was collected (Adeyemo and Rotimi, 2014). This data illustrates the scarcity of African-focused GWAS (n=1744 studies)

2. Obesity

Obesity is described as an imbalance between energy intake and output. Consequently, excess energy ends up being stored in fat cells, which eventually grow bigger and/or increase in number (Goedecke et al., 2005). Increased intake of energy-dense foods and a decrease in physical inactivity (Speakman and O'Rahilly, 2012) contribute to the high obesity rates reported globally. Some have argued that obesity is a consequence of our changing environment i.e. humans have become obese over time due to changes in eating habits and modes of transportation (influencing physical activity) and increased sedentary behaviour (Speakman and O'Rahilly, 2012). However, identifying the causal relationship of environmental factors with obesity is not an easy process (Hebebrand et al., 2013), with many such factors other than food intake and level of exercise having been linked to obesity. These include gut microbiota, stress, socio-economic status (Speakman and O'Rahilly, 2012), duration of sleep, education level and other cultural factors (Feeley et al., 2012).

The genetic input into obesity is also complex. Several studies have reported on the high heritability of BMI in European populations (Bodurtha et al., 1990, Walley et al., 2009), suggesting that genetic factors play a role in driving body composition. Moreover, gene-gene and gene-environment interactions may also play a role in the variability observed in BMI among human populations. This thesis will concentrate on the genetic aetiology of obesity in sub-Saharan African study participants and the present chapter will give an overview of the epidemiology, consequences, measurement and causes of obesity with a particular emphasis on the genetic determinants of body fat mass.

2.1 Epidemiology of obesity

2.1.1 Globally

Obesity has become a global epidemic with an estimated 700 million obese and about 2 billion individuals being classified as overweight according to WHO estimates in 2010 (Speakman and O'Rahilly, 2012). The most recent Global Burden of Disease (GBD) study has confirmed that these combined overweight and obesity figures have increased from 857 million in 1980, to 2.1 billion in 2013 (Ng et al., 2014). If the obesity trend continues at this rapid rate, it is estimated that over 50% of adults in the United States (US) will be clinically obese by 2030. The GBD study (Ng et al., 2014) found that countries like Tonga and Kuwait have extremely high prevalences of obesity and that greater than half of the ~671 million obese individuals in the world reside in just ten countries which collectively accounted for 13% of obese individuals worldwide. These researchers also observed other interesting

trends spanning the period between 1980 - 2013 (Ng et al., 2014):

- Globally, the combined prevalence of overweight and obesity increased by 27,5% for adults and 47,1% for children.
- An increase in overweight in both women and men in both high-income countries (HIC) and LMIC but with different sex patterns; there were more *overweight and obese females in LMIC* while HIC showing the converse.

2.1.2 South Africa

It has been suggested that the increase in overweight and obesity in SA is associated with both adopting a "westernised" lifestyle (nutrition transition) together with urbanisation (geographical transition) (Armstrong et al., 2006). Also, it has been shown that South African urban black females are at greatest risk of becoming obese (Janssen et al., 2004, Goedecke et al., 2005, Obesity Task Force, 2005). Findings from the first South African Demographic and Health Survey (SADHS) conducted in 1998 on 13 089 individuals showed that the overall prevalence of overweight (BMI >25) and obesity (BMI >30) in South Africa was high, with 29% of men and 56% of women being classified as overweight or obese (Puoane et al., 2002). The most recent South African National Health and Nutrition Examination Survey (SANHANES-1) conducted by (Shisana et al., 2013) reaffirmed that obesity levels are on the rise in SA, and also highlighted that black females have the highest prevalence of obesity amongst South Africans (summarized in Table 1.1). This combined prevalence of overweight and obesity according to SANHANES in females and males (≥15yrs) is ~64% and 30.7%, respectively. The GBD study estimates of the combined prevalence for South Africa are slightly higher than the local estimates as indicated in Table 1.1 with figures of overweight and obesity at 69.3% and 38.8% in females and males, respectively. The inconsistency in GBD and local estimates is associated with a difference in collection periods as well as the inclusion of adolescent data in the local estimates (≥15yrs) whilst GBD (≥20 years) estimates are based on an entirely adult dataset. Ultimately both figures highlight the significant burden of obesity in countries undergoing epidemiological transition, such as South Africa and the need for urgent intervention, especially in black females.

2.1.3 Obesity in children and adolescents

Globally the prevalence of overweight and obesity has increased (1980-2013) by almost 47% in children (2-18 years) in both HIC and LMIC with little difference between sexes (Ng et al., 2014). It seems that factors that predispose children to an overweight or obese trajectory may persist to adulthood (Singh et al., 2008, Lundeen et al., 2015). A recent study based on the Birth to Twenty Cohort (Soweto, Johannesburg) has shown that longitudinal data is a valuable resource for determining periods of "greater risk" for assessing childhood obesity (Lundeen et al., 2015). This study shows that for the females in the cohort, overweight and

obesity prevalence gains in momentum throughout childhood and adolescence, and the years just succeeding puberty seem critical for the development of obesity (Lundeen et al., 2015). In boys, however, even by late adolescence, overweight and obesity are not as prevalent as for girls. This highlights the importance of planning interventions at critical time points especially with younger females.

Table 1.1 Prevalence of obesity and overweight in South Africa
Prevalence of overweight and obesity in South Africa stratified by race and sex. Summarised from (Shisana et al., 2013, Ng et al., 2014)

Population Groups	**Overweight (%)**		**Obese (%)**		**Combined overweight and obese**	
	Males	Females	Males	Females	Males	Females
Black	19.1	24.9	9.4	39.9		
Indian	32.2	22.8	7.6	32.4		
Coloured	22.1	24.4	15.1	34.9		
White	36.3*	18.2*	26.5*	22.7*		
SANHANES Overall (≥15yrs)	20.1	24.8	10.6	39.2	**30.7**	**64.0**
GBD Overall (≥20yrs)			13.5	42.0	**38.8**	**69.3**

* Data for the white population were included from Puoane et al. (2002) since the study by Shisana et al. (2013) lacked available data

The heritability of obesity susceptibility seems to increase throughout childhood and adolescence until adulthood, where it then decreases, as suggested by evidence from longitudinal twin studies (Den Hoed et al., 2010, Loos, 2013). This would suggest that the environmental exposure time in childhood is far less than in adults and that genetics might play more of a critical role in influencing the obesity phenotype. A better understanding of the influence of genetic factors on obesity may lead to better management and strategies to introduce timely interventions that could prevent and/or reduce the number of obese children from becoming obese adults.

2.2 Consequences of obesity

2.2.1 Global

The latest GBD study reported that overweight and obesity were estimated to cause about 3.4 million deaths worldwide in 2010, which equates to 4% of all lives lost (Ng et al., 2014).

Overweight and obesity are the fifth leading risk factor for deaths, globally (Obesity Task Force, 2005). High BMI is the seventh leading risk factor for males, the third leading risk factor in females and overall sixth leading risk factor contributing to the overall global burden of disease (Lim et al., 2013). Specifically overweight and obesity are accountable for 44% of diabetes cases, 23% of ischemic heart disease cases and a range of certain cancer burdens ranging from 7%-41% (Lau et al., 2015).

2.2.2. Local (South Africa)

Obesity is a considerable risk factor for the development of several chronic, non-communicable diseases (NCDs) in LMIC such as South Africa (Tollman et al., 2008, Mayosi et al., 2009), as it is globally. It has been outlined by Goedecke et al. (2005) and others that being obese increases the risk for a myriad of conditions including type 2 diabetes, coronary heart disease (CHD), hypertension, certain forms of cancers, psychological conditions, osteoarthritis as well as morbidities in children. It has been reported that obesity itself can be classified as a chronic disease and can be separated into two categories, those diseases associated with *increased fat mass* like sleep apnoea, osteoarthritis and psychological problems and those associated with the *metabolic effects* of adiposity like CHD, hypertension, type 2 diabetes and certain cancers (Joubert et al., 2007). A study by Joubert et al. (2007) based in a South African setting, reported that excess body weight was a major risk to health, particularly among South African females where the burden of disease was almost double that of males. They also reported that overall almost 87% of type 2 diabetes, 68% of hypertensive disease, 61% of endometrial cancer, and 45% of ischaemic stroke was as a result of BMI ≥ 21 $kg.m^{-2}$. Variations in body composition measures, including BMI, waist circumferences, fat mass and others described below have been used to assess the level of obesity in individuals.

2.3. Estimating body composition

The *whole body* (Fig. 1.2- Level V) level of body composition describes body size and configuration which is often explained by the anthropometric measures described below. Body composition can also be described on a *molecular level* (Fig. 1.2- Level II), which takes into account six main components including water, lipid, protein, carbohydrates, bone minerals and soft tissue minerals (Santos et al., 2014). These measures can be estimated through a variety of methods such as *direct* body composition methods (estimates of levels I-III described in Fig. 1.2) including isotope dilution, total body counting and neutron activation (Duren et al., 2008). *Criterion methods* measure a property of the body e.g. density or describe quantities and distribution of adipose, skeletal or muscle, and include Dual-Energy

X-Ray Absorptiometry (DXA) scans, Computerized Tomography (CT) and Magnetic Resonance Imaging (MRI) scans. *Indirect methods* provide estimates of body composition based on results from direct or criterion methods and include anthropometry and bioelectrical impedance (Duren et al., 2008). Indirect methods rely on the biological interplay between body components and tissue types that are measured by direct or criterion methods and their distribution among normal individuals (Duren et al., 2008). Therefore these methods are subject to large predictive errors and the extrapolation of data may be limited especially when obtaining molecular body composition measures (Duren et al., 2008, Santos et al., 2010).

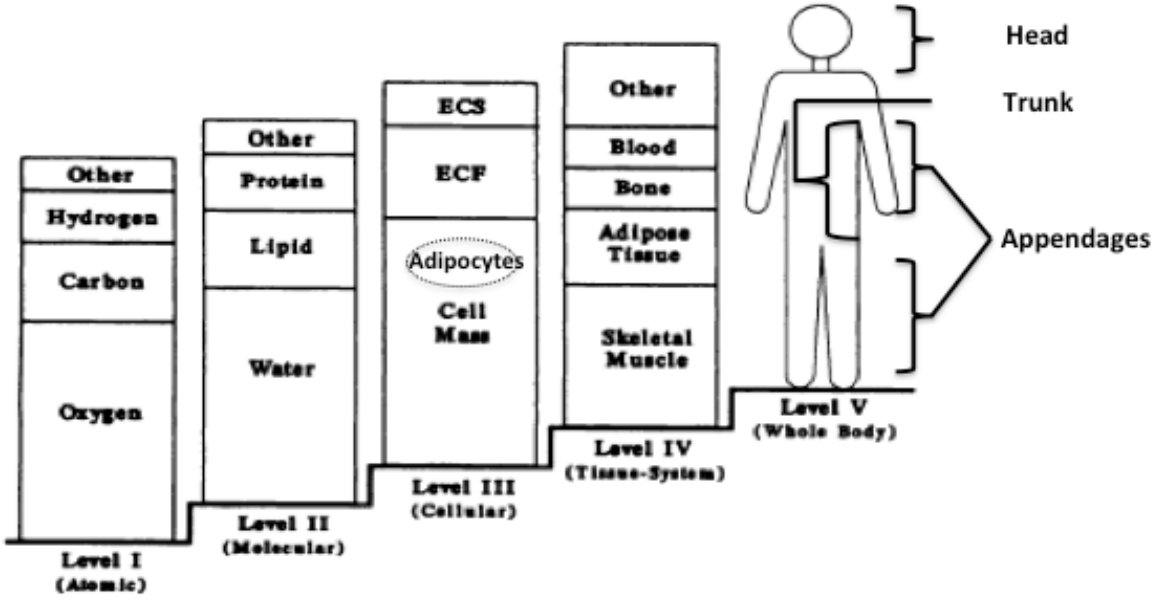


Figure 1.2 The five levels of human body composition
[Adapted from (Wang et al., 1992)] according to (Heymsfield, 2005). ECF - refers to extracellular fluid, ECS- extracellular solids.

2.3.1 Indirect measures

Anthropometric measurements are the simplest way of determining body composition and are used to describe body mass, shape, degree of fatness, and size of individuals. A basic assessment of adiposity can be determined from body size which changes with a gain in weight (Duren et al., 2008). For adequate assessments of overall adiposity is it important that standardized anthropometric techniques are used for comparison as they are very much influenced by sex, age and population from which references are derived.

BMI

Typically individuals with higher body weight have higher amounts of fat and body weight is

highly related to stature (Duren et al., 2008). BMI is a readily measurable proxy for body fat on a population-level based on an individual's weight in kilograms divided by the square of height in meters (kg/m^2) (WHO, 1995). Obesity and overweight are defined according to criteria outlined by the World Health Organisation (WHO, 1995). Being overweight is described as having a BMI of ≥ 25.0 kg/m^2 and obese as having a BMI of ≥ 30 kg/m^2 (WHO, 1995). There are further grades within the 'obese class' that are defined by BMI ranges (class I: 30.0-34.9 kg/m^2, class II: 35.0-39.9 kg/m^2, class III: >40.0 kg/m^2) (Nishida et al., 2010). BMI is often thought to be a crude indicator of obesity because it doesn't take into account body composition or weight distribution, however it has been traditionally used, as it is an easy and inexpensive measure to obtain, requiring only a scale and stadiometer. Also, because of the availability of comprehensive reference data and its established relationships with fatness, morbidity and mortality in adults (Duren et al., 2008), BMI has conventionally been used as a predictor of increased risk for non-communicable diseases (NCDs) (Janssen et al., 2004, Obesity Task Force, 2005, Menke et al., 2007).

WAIST CIRCUMFERENCE, HIP CIRCUMFERENCE AND WAIST-TO-HIP RATIOS

Waist circumference (WC), and waist-to-hip-ratios (WHR) are measures of central adiposity (abdominal obesity) and have been shown to be good predictors of health risk, metabolic syndrome or cardiometabolic diseases (Janssen et al., 2004, Menke et al., 2007, Crowther and Norris, 2012, Murphy et al., 2014), even after adjusting for BMI. The reason for this is that abdominal obesity reflects the amount of visceral adipose tissue (VAT) and the level of VAT has been speculated to be strongly correlated with cardiovascular disease and related pathologies (Lear et al., 2010). According to WHO (2011), WC measurements should be made at the midpoint between the lower margin of the last palpable rib and the top of the iliac crest in centimetres (cm). It is strongly noted by WHO (2008b) that sex, age, reproductive status (including parity and menopause) and ethnicity are key regulators for both WC and WHRs and inform the cut-offs used to predict increased metabolic disease risk.

All studies utilising hip circumference (HC) measurements show that the measurement should be taken around the widest portion of the buttocks in centimetres (WHO, 2008 (b)). Recently more studies have included HC to examine whether the components of WHR exert their effects primarily through waist circumference or hip circumference. The WHR is commonly defined as WC divided by HC and more correctly reflects the ratio of abdominal (includes both subcutaneous and VAT) circumference to hip circumference. The ratio of the two measurements gives a basic index describing the distribution of abdominal adipose tissue (Duren et al., 2008) and seems to be more accurately measured than skin fold estimates (WHO, 2011). This index provides a description of both subcutaneous and visceral

abdominal adipose tissue and is therefore much more useful as an indication of both fat distribution and adiposity.

African women have been shown to have a lower levels of visceral but higher subcutaneous fat mass than BMI-matched European women (Crowther and Ferris, 2010). Fat distribution is strongly associated with sex and ethnicity, and many of the genetic signals related to the distribution of body fat seem to be more strongly associated in females than males (Visscher, et al., 2012). It has also been suggested that the current anthropometric cut-points used to define adiposity (particularly waist circumference), certainly as indicators of cardiometabolic risk, may not be suitable for African populations (Crowther and Norris, 2012, Murphy et al., 2014), as African populations have been shown to be different not only in terms of body fat distribution but also in terms of height, and fat mass when compared to their European counterparts (Kruger et al., 2015). Both of these studies have recommended changes in the currently recommended cut-point for waist circumference, which would have an impact on the prevalence of abdominal obesity and of metabolic syndrome in these populations.

SKINFOLD MEASUREMENTS

These measurements are used to describe the thickness of subcutaneous adipose tissue (SAT) at various regions of the body including trunk skinfolds, arm skinfolds, and leg skinfolds (Lukaski, 1987, Duren et al., 2008), of which the tricep and subscapular skinfold measurements seem to be the most well documented. Skinfold measurements are made on two assumptions: (i) that the sites chosen for measurements represent the average thickness of the SAT and (ii) the thickness of the SAT is a consistent proportion of the total body fat (Lukaski, 1987). Skinfold measurements are made by grasping the skin and neighbouring subcutaneous tissue (trying to exclude the underlying muscle) whilst pulling away from the body to allow the caliper jaws to grasp onto the skinfold (Lukaski, 1987). This technique is subject to operator accuracy and therefore duplicate readings are encouraged to improve the precision involved with this technique. The utility of this techniqe seems to be more limited for use in overweight or obese adults because of upper measurement constraints of the caliper (Duren et al., 2008) and difficulty in grasping the skinfold. Body fat estimation in children is more effective using this technique due to mainly subcutaneous localisation of fat deposits and smaller body size (Duren et al., 2008).

BIOELECTRICAL IMPEDANCE

This method is reliant on the principle that electrical conductivity is greater in fat free tissue

mass than it is in fat mass (Lukaski et al., 1985). With this technique, estimates of total body water (TBW) and fat free mass (FFM) are generated by measuring the body's resistance as a conductor to small electrical impulses (Duren et al., 2008). Here, an impedance index (stature^2/resistance) is proportional to the volume of TBW and is used to predict body composition by acting as an independent factor in regression equations (Duren et al., 2008). The validity of this technique and its value in evaluating body composition seems to be more useful at a group level than for individuals due to large errors (Duren et al., 2008), and has limited use for obese individuals. Also, this method is only accurate when the population in whom the measurements are being made closely match the reference population usually Caucasian, from whom the equations used to predict body composition were derived, in body size and shape (Duren et al., 2008) and are therefore not good measures to employ in Sub-Saharan African populations.

2.3.2 Direct measures

TOTAL BODY WATER (TBW) estimates are obtained by isotope dilution. Water/isotope dilution volumes allow for the prediction of FFM, and fat (calculated by body weight – FFM) because of the stable relationship of water to FFM (Duren et al., 2008). This is measured in normal weight individuals and TBW has limited use in obese individuals. This is because the density of excess adipose tissue in obese individuals results in higher and water and protein content of FFM. An increased water content of FFM then results in a greater estimation bias for percentage fat with increasing levels of obesity (Lohman and Milliken, 2003).

TOTAL BODY COUNTING measures the amount of naturally produced radioactive potassium 40, which inherently provides an estimate of body cell mass, since most cell bodies contain potassium (Ellis, 2005, Duren et al., 2008). FFM can then be calculated once an idea of total body potassium is established on the premise that there is a constant concentration of potassium in FFM (Ellis, 2005, Duren et al., 2008).

NEUTRON ACTIVATION involves the exposure of a subject to a neutron field, where the generated gamma output can be measured as well as many elements in the body like carbon, nitrogen, sodium and calcium (Ellis, 2005, Duren et al., 2008). Measures of nitrogen can be quantified to predict the amount of protein in the body, which can then be used to establish components of FFM. A limitation of this technique is the long exposure to radioactive elements.

2.3.3 Criterion methods (DXA, CT and MRI)

DUAL-ENERGY X-RAY ABSORPTIOMETRY (DXA) SCANS

Dual energy X-ray absorptiometry (DXA) is the preferred method of inferring body composition by quantifying fat, lean and bone tissues. This method allows for the resolution of total adipose tissue and soft tissue, bone mineral content and bone mineral density (Duren et al., 2008). The advantages of DXA is its ease of use and the speed of conducting a whole body scan, but estimates are limited by differences in models and the software between manufacturers as well as between machines. There are also physical limitations in terms of excessive body length, body weight, thickness and width, which is pronounced with obese subjects (Duren et al., 2008, Santos et al., 2010).

Estimates of body composition can be estimated regionally (head, trunk or appendages) or for the whole body to estimate bone mineral content (BMC) and soft tissue (fat mass and fat-free mass), respectively. These estimates can be used to get an idea of body composition as well as body fat distribution. This in indicated in the DXA image (Fig. 1.3) where the different colours correspond to different tissue types that can be differentiated. Measurements of total fat mass, fat-free and percentage fat mass (PFM) can be used as indicators of adiposity. The main limitations with the use of DXA scans is perhaps its impractical use in the field and caution should be exercised with regard to the error in measurement when detecting small changes in body composition (Santos et al., 2010).

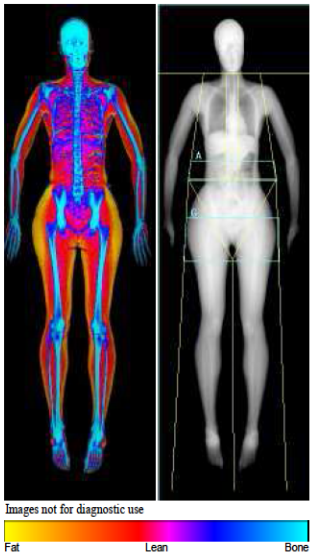


Figure 1.3 Modified DXA scan
Modified DXA scan showing the three tissues scanned; fat (yellow), lean (red) and bone (blue) (https://www.measureup.com.au/media/docs/female_body_composition_example)

COMPUTERISED TOMOGRAPHY AND MAGNETIC RESONANCE IMAGING

Newer techniques such as computerised tomography and magnetic resonance imaging are becoming more popular for use in estimating body composition. Again their utility for use on obese subjects is limited. Although CT is able to accommodate larger body sizes, whole body assessments are not possible because this would involve long exposures to harmful radiation (Duren et al., 2008). However it is possible to concentrate on a specific area e.g. intra-abdominally (Duren et al., 2008). MRI does not allow for larger-bodied individuals and is more suited for normal-slightly overweight people. Accurate whole body estimates from CT and MRI depend on extra time and software (Duren et al., 2008).

Indirect measures such as anthropometry may provide an easy and inexpensive way to determine body composition, however, the inconsistencies that may arise make it less reliable than others. Criterion methods, in particular DXA, are precise, have low-radiation doses, and are easier to access than other ways to examine body composition, and therefore seem to be the most appropriate and valuable tool to describe body composition (Santos et al., 2014).

2.4 Causes of obesity

It is well accepted that many factors may contribute to an obesity phenotype, given its highly complex nature including the influence of genetics, the environment and more recently epigenetics and the microbiome.

2.4.1 Non-Genetic

NUTRITION AND PHYSICAL EXERCISE

Most agree that the recent increase in obesity prevalence is orchestrated by increases in the consumption of cheap, tasty foods and sedentary lifestyle (Goedecke et al., 2005, Farooqi, 2012). The epidemiological transition, which describes a change in environment from one in which the prevalence of infectious diseases is high (with associated periodic famine, malnutrition, poor sanitation) to one in which chronic diseases are more common, is mainly as a result of moving to more industrialised areas (urbanisation) and the resulting changes in lifestyle as previously mentioned (Omran, 2001, Popkin, 2006, Feeley et al., 2012, Popkin et al., 2012). The nutrition transition, which emerged in response to the epidemiological transition, is as a result of alterations to dietary patterns with concomitant increases in the intake of fats, sugars and refined foods (Popkin, 2006, Feeley et al., 2009, Feeley et al., 2012, Popkin et al., 2012) and increased occurrences of non-communicable diseases, of which increases in obesity rates drive these occurrences. This is particularly relevant in the South African context where the ‘rapid transformation’ from traditional diets (lower in fats,

higher in carbohydrates) to more “Westernised” diets (higher in calories) has resulted in increased obesity rates (Feeley et al., 2012). This nutrition transition is influenced by a number of factors including social, economic factors and cultural practices, that all work together to influence what we consume. The pattern emerging over time, is a substantial shift in the foods we consume, our level of physical activity (more sedentary behaviours), and what we drink (increased consumption of sugary drinks) and these changes are influencing body composition (Popkin et al., 2012).

Epidemiological studies have noted the influence of environmental factors in early life together with changes in gene expression (epigenetic and genetic mechanisms) that result in changes in disease patterns (Silveira et al., 2007). These observations are supported by similar experimental observations that underpin a strong relationship between unsuitable environmental exposures during the early stages of life and the resulting chronic diseases that prevail throughout life (Silveira et al., 2007), the culmination of which has led to the concept of the Developmental Origins Of Health And Disease (DOHaD) hypothesis (Burgio et al., 2015).

DEVELOPMENTAL ORIGINS OF HEALTH AND DISEASE

The idea of a ‘thrifty genotype” was first proposed by James Neel in 1962 (Neel, 1962), where he suggested that random genetic mutations that were responsible for efficient processing and deposition of fat when food was plentiful, were selected for during periods of famine, over the course of human evolution (Gillman, 2005, Silveira et al., 2007, Burgio et al., 2015). When populations were exposed to caloric-rich environments, this “thrifty genotype” would then be detrimental to those individuals carrying it and eventually lead to obesity and co-morbid diseases. This theory has been criticised over the years where researchers have shown that populations with high prevalences of obesity or diabetes, have not historically been exposed to famines (adverse conditions) yet still have high levels of insulin resistance (Burgio et al., 2015).

Lucas (1991) first introduced the concept of *foetal programming* where he postulated that “a stimulus or insult at a critical period of development has lasting or lifelong significance.” Later Hales and Barker (1992) proposed the idea of a “thrifty phenotype” with regards to the origins of type 2 diabetes, which proposed that inadequate nutrition in early life, impaired the development of the pancreas leading to the development of type 2 diabetes in late life, particularly in those subjects who were exposed to an obesogenic environment in adult life. This concept then led Barker and colleagues to propose that adverse intrauterine, childhood conditions and as a result, low birth weight would result in the onset of many chronic diseases later in life including cardiovascular and type 2 diabetes disease phenotypes

(Barker et al., 2002). These observations together with other studies conducted around the same time regarding low birth weight, refined the “thrifty phenotype” hypothesis, which postulates that the foetus is able to adapt to adverse intrauterine conditions by reducing its energy needs to guarantee survival. This adaptation can be costly, affecting certain organs over others, like the heart and pancreas and producing long-standing effects later in life (Silveira et al., 2007). It has been hypothesised that these foetal adaptions are the result of epigenetic modifications of the human genome.

Important research by Ravelli et al. (1976) conducted on male offspring born to mothers who were exposed to food rationing during different periods of gestation (Dutch Famine 1944-1945), showed different body composition patterns in the offspring later in life (Silveira et al., 2007). The incidence of obesity was increased in the offspring of mothers exposed to famine earlier in their pregnancy (first six months), than if the insult occurred later, specifically in the last three months (Ravelli et al., 1976, Silveira et al., 2007).

EPIGENETIC CONTRIBUTIONS TO OBESITY

In order for a foetus to survive during adverse conditions that may occur in the intrauterine environment such as exposure to maternal malnutrition, certain adaptations or changes have to be made during critical developmental periods (Smith and Ryckman, 2015). These adaptations can also have a lingering metabolic effect, increasing one’s risk of developing chronic diseases such as obesity later on in life. These “programmed” adaptations are termed ‘developmental programming’ and may be the result of epigenetic changes to the way genes are expressed, or permanent changes in the organisation of organs and tissues. Epigenetic changes refers to changes in gene expression that are heritable and that do not involve changes to the existing DNA sequence but rather changes to the biochemical structure of DNA (Smith and Ryckman, 2015). These epigenetic changes can occur through various modifications including, DNA methylation, histone modification, and non-coding RNA processes (Li, 2002, Goldberg et al., 2007, Mercer et al., 2009). Epigenetic modifications alter the way the transcriptional units are accessed, and in doing so determine whether the gene is expressed or not (Van Dijk et al., 2015).

Maternal over-nutrition has also been implicated in increasing risk to the offspring for developing obesity (Kaar et al., 2014, Smith and Ryckman, 2015). Most of the evidence that portray developmental programming as critical to the outcome of chronic NCD’s is observed through epidemiological and now recently well-established animal model studies. There are many studies that point to body composition being in part regulated by epigenetic mechanisms, in particular DNA methylation studies where the pattern of methylation of

several genes have been implicated in obesity (Burgio et al., 2015). Prader-Willi syndrome (PWS) for example results in developmental delays, impaired cognition, uncontrolled eating and life-threatening obesity (Xia and Grant, 2013), caused epigenetic errors. In the majortiy of cases, the paternal copy of the region is deleted, and the maternal copy, while present, is silenced and non-transcribed due to the normal imprinting pattern of this locus. Anglemann syndrome, the reciprocal disorder to PWS, shows growth retardation and together with Beckwith-Wiedermann and Russel-Silver syndromes, serve as examples to illustrate that imprinted loci are often involved in early development, and aberrant imprints lead to either undergrowth or overgrowth syndromes. Also imprinting may be cell or locus specific at the human leptin (explained below) promoter region because of differing methylation pattern between alleles and cells (Xia and Grant, 2013); thus regulating BMI variance. More investigations have to be conducted to fully understand the epigenetic mechanisms by which body weight is regulated.

MICROBIOME

Increasing evidence from studies have shown that some interaction exists between host, environment, diet, microbiota and obesity and these have mostly been investigated through metagenomic and metabolomic approaches (Burgio et al., 2015, Graham et al., 2015). The human microbiome is said to encompass 100 trillion microbial cells containing almost 100-fold more unique genes than the human genome (Burgio et al., 2015, Graham et al., 2015). Most microbiota belong to two main classes, *Firmicutes* and *Bacteroidetes*, with the rest of the organisms belonging to both gram-positive and gram–negative bacteria (Graham et al., 2015). Although there is a proposed core set of microorganisms that dominate the microbiome, a large proportion of the variation in the microbiome between individuals depends on the environment and diet to which the host is exposed, during various stages of the life cycle (Graham et al., 2015).

Initiatives such as the Human Microbiome Project aim to assess the relationship between the human microbiome and particular health outcomes (Ding and Schloss, 2014). Metagenomic studies have shown that some combinations of gut microbia either predispose or protect individuals from weight gain (Burgio et al., 2015). In some instances obese individuals have been shown to have increased colonisations of *Firmicutes* and decreased *Bacteroidetes* species in the gut compared to non-obese subjects (Tsai and Coyle, 2009, Ley, 2010). The mechanisms through which gut microbiota may influence body fat mass is complex but it has been proposed (Tsai and Coyle, 2009, Ley, 2010, Parekh et al., 2014, Burgio et al., 2015, Graham et al., 2015) that:

- The permeability of the gut is modified by the microbiome causing the absorption of lipopolysaccharides from the gut bacteria which leads to an inflammatory response thus increasing insulin resistance,
- Gut microbiota affects the hosts metabolism by increasing the energy harvest from the ingested food, or increasing fat storage in adipocytes, altering endocrine function or modulating the breakdown of lipids,
- The microbiome interferes with the regulation of epigenetic signals in obese and type 2 diabetic individuals.

Although gut microbiota have been implicated to influence obesity, the extent to which they bring about weight gain or loss is yet to be completely elucidated because of the strong interplay between diet and the microbiome. The current obesity epidemic has been mainly attributed to adopting a "Westernized" lifestyle as stated earlier, however not everyone exposed to this obesogenic environment –(intrauterine, epigenetic or microbiome), becomes obese.

2.4.2 Obesity Genetics (Syndromic/monogenic, common)

Genetics has an important part to play in programming our physiological state and therefore has a large influence on body composition (Loos, 2013) with strong evidence supplied by twin-, family- and adoption studies supporting the heritability of body fat mass and distribution (Bodurtha et al., 1990, Walley et al., 2009). In order to foster viable treatment and prevention strategies, we need to search for 'at-risk' genes with a view to understand pathways and networks that mediate body weight regulation. The discovery that the disruption of certain biological pathways, due to monogenenic mutations, could result in obese phenotypes was perhaps the first step in trying to understand the influence that genetics has on the regulation of body weight.

MONOGENIC AND SYNDROMIC OBESITY

When obesity is a result of a chromosomal aberration or a single gene effect, the pattern of inheritance is often Mendelian and normally results in an extreme form of obesity with an early-onset (Loos, 2013). These forms of obesity are mostly rare and are often classified into either monogenic or syndromic obesity. More than 200 instances of these monogenic (single-gene) disorders have been reported, where they seem to predominantly affect the leptin-melanocortin pathways, which are critical in body weight regulation by controlling food intake through the brain-satiety pathway located in the hypothalamus (Loos, 2012, Zegers et al., 2012, El-Sayed Moustafa and Froguel, 2013). The main genes implicated in these pathways

are *leptin* (*LEP*) and its receptor (*LEPR*), *pro-opiomelanocortin* (*POMC*) and its derived hormones, the *melanocortin-4 receptor* (*MC4R*) as well as *pro-hormone convertase subtilisin/kexin type 1* (*PCSK1*) (Loos, 2012, Zegers et al., 2012, El-Sayed Moustafa and Froguel, 2013). Syndromic obesity is associated with an extreme obesity phenotype that is often accompanied by other abnormalities including developmental delay and brain dysfunction. The most common forms include Prader-Willi syndrome (mentioned previously) and Bardet-Biedl syndrome where the genetic causes include single-gene defects, chromosomal and imprinting abnormalities as well as X-linked disorders (Hebebrand et al., 2013, Loos, 2013).

COMMON POLYGENIC OBESITY

Heritability estimates indicate that genetic factors contribute to almost 40-70 % of the inter-individual variation in measures of body fat mass, particularly BMI (Elks et al., 2012, Nan et al., 2012). Common obesity is a complex disorder (Walley et al., 2009) of which the interplay between a large number of genetic variants and gene-environment interactions (Loos, 2012) complicate the understanding of its genetic aetiology. Family-based linkage studies alluded to the initial genetic link to common obesity together with studies on specific ethnic groups to estimate heritability of obesity risk. Heritability studies aim to quantify how much of variation in obesity risk is explained by a heritable contribution. The heritability of BMI in indigenous African populations is largely understudied, barring a single publication on a Nigerian cohort (Luke et al., 2001) where heritability estimates were shown to be 50% for weight, BMI, fat mass and percentage fat mass.

The main approaches for investigating complex diseases such as polygenic obesity have involved genetic association studies or recently the use of whole genome sequencing (WGS) efforts. When using genetic association studies to identify loci for disease-related traits, one could employ two strategies, the candidate-gene approach, which is hypothesis-driven, or the genome-wide approach. Direct candidate gene analysis relies on an *a priori* hypothesis to identify a disease-causing variant by direct genotyping/sequencing. The rationale behind focusing on allelic variation in specific, biologically relevant regions of the genome is that certain mutations will directly impact the function of the gene in question, and lead to the phenotype or disease state being investigated. This approach usually uses case-control study designs to try and answer questions “Is one allele of a candidate gene more frequently seen in subjects with the disease than in non-disease controls. GWAS relies on an indirect association to locate a disease-causing variant. GWASs typically focus on associations with SNPs and traits that contribute to disease. If one type of the variant is more frequent in

people with a particular phenotype, the variant is said to be associated with that phenotype. The associated SNPs are then considered to mark a region of the human genome that may influence the risk of disease.

3. Investigating genetics related to obesity phenotypes

Public databases such as dbSNP (Sherry et al., 2001), HapMap and recently the 1000 genomes project (1000G) have given us more insight into the variation harboured within the human genome, which has led to the development of tools to help in the study of polygenic diseases such as obesity. The HapMap project in particular enabled the use of tag-single nucleotide polymorphisms (SNPs). Tag SNPs allow for the genotyping of representative SNPs in a block of the genome that has high linkage disequilibrium (LD) and is sufficient to identify haplotypes (Gibbs et al., 2003), thus reducing the cost of the genotyping exercise. A haplotype is made up of a group of specific alleles located near each other on the same chromosome that tend to be inherited together (Pearson and Manolio, 2008). The use of tag SNPs is particularly relevant in candidate gene studies and GWAS carried out in populations of varying ethnicities, where tagSNPs can be selected against the LD background for that particular population (or closest to it). When designing approaches to investigate genetic associations, there are two possible interpretations of a positive association. The first one (direct) being that the association arises due to the identification of the disease causing allele or secondly (indirect) that the disease- associated allele is in LD with the true causal variant (Collins et al., 1997). The direct association methods forms the basis for candidate gene studies in that they are hypothesis driven.

Both these interpretations of a positive association (direct or indirect) rely on the premise of common variants underlying susceptibility to common diseases. This common disease-common variant (CDCV) hypothesis (Lander, 1996, Cargill et al., 1999, Chakravarti, 1999, Reich and Lander, 2001) predicts that there will be one or few predominating disease-causing alleles at each of the major underlying disease loci (Reich and Lander, 2001).

3.1 Candidate gene studies of obesity

Candidate gene studies are hypothesis driven, relying on what we understand about the current biology and pathophysiology of the disease that is being investigated (Loos, 2009). Genes that are implicated in energy homeostasis in animal models and extreme monogenic models of obesity are tested for correlation with obesity-associated phenotypes (variances in

body composition) at the population or case control level. Historically candidate gene studies have not been as successful as GWAS in reinforcing obesity-related associations. The main limitations being that *sample sizes* are often too small to detect small or moderate effects, that the *tag-SNPs* did not efficiently capture the variation within the gene or that the biological pathway was not fully understood so incorrect candidate genes were selected (Loos, 2013). Large-scale candidate gene studies of BMI conducted recently have the necessary power to detect smaller effects and also to refute false-positive associations. Most large-scale candidate gene studies have identified common variants (most of them resulting in changes to the amino acid sequence) in about six genes, *MC4R*, *β-adrenergic receptor 3* (*ADRB3*), *PCSK1, brain-derived neurotrophic factor* (*BDNF*)*, melatonin receptor type 1 B* (*MTNR1B*) *and lactase* (*LCT*) (Loos, 2013).

3.2 Genome wide association studies (GWAS)

GWAS are designed to identify genetic associations with observed phenotype(s)/trait(s) that occur across the genome and exploit the principle of LD at the population level (N.I.H., 2007, Pearson and Manolio, 2008). LD is the non-random (more likely to be inherited together) association between alleles at different loci that are in close proximity to each other (Pearson and Manolio, 2008). The LD is measured using Pearsons coefficient of correlation- r^2 or D´ (Ardlie et al., 2002). LD arises due to evolutionary forces such as selection, drift, mutation and is broken down by recombination events (Hartl and Clark, 1997). The genomic distance at which LD decays determines how many genetic markers are used to tag/represent a haplotype (Visscher et al., 2012). The HapMap Project in particular (Gibbs et al., 2003), resulted in a list of tag SNPs that captured most of the common variation in currently 11 populations of various ethnic backgrounds included in HapMap-Phase 3. Subsequent to, and in response to the strides made by the HapMap project, commercial companies such as Illumina and Affymetrix have produced dense SNP chips that allowed hundreds of thousand of SNP markers to be combined in a single experiment (Visscher et al., 2012). GWAS studies have limitations in that they rely on the LD between the SNPs being genotyped and the causal variants and thus the causal variant may be quite distant from the identified tagSNP (Visscher et al., 2012). Also statistical association (assesses the relationship between two alleles) is based on the frequency of the alleles in the population. So the strength of the association decreases as the level of LD decreases. Most GWAS are designed to detect associations with causal variants that are common in the population (minor allele frequency- MAF ≥ 0.05). Also to minimise the number of false positive associations that may arise because of the large numbers of SNPs on the array, multiple testing corrections and being able to replicate associations in an independent study are vital

(Fall and Ingelsson, 2014). The accepted 'genome-wide' (GW) significance levels are *P*-values ≤ 5 X 10^{-8} (Hoggart et al., 2008, Panagiotou and Ioannidis, 2012). This limit is based on a significant *P*-value of α=0.05/number of independent SNPs. Despite these limitations, GWAS have contributed to our understanding of the genetic aetiology of polygenic diseases and have provided us with more information on the allele frequencies and effect sizes of variants involved in these diseases (Andersen and Sandholt, 2015).

3.2.1 GWAS associated with Body Mass Index (BMI)

The distribution of risk alleles associated with variations in BMI at a population level would be represented by only very few individuals having the least risk alleles and at the other extreme by very few having the most risk alleles, if the data was normally distributed as shown in Fig. 1.4. Most individuals in the population would have an intermediate number of risk alleles. Being able to genotype a large enough number of individuals at the population-level is necessary for discovery and validation of these at risk variants, which has led to more and more collaborative efforts of GWAS using a meta-analysis approach (Hebebrand et al., 2013).

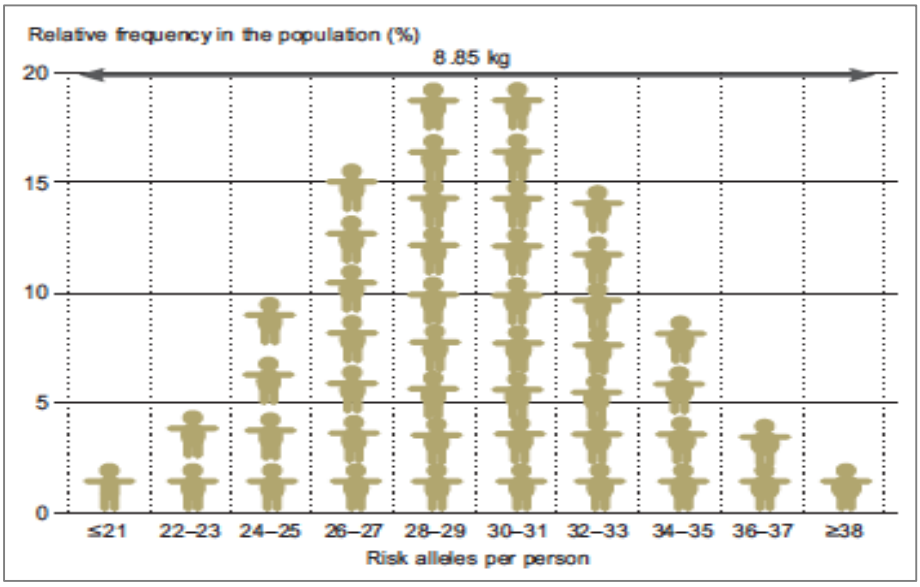


Figure 1.4 Effect estimators of genetic variants for body weight
Effect estimators of genetic variants for body weight identified by meta-analysis of GWAS (Hebebrand et al., 2013)

EUROPEAN COHORTS

The first gene variant associated with increased BMI identified by GWAS using high-throughput methodologies was the *FTO* (fat mass and obesity associated) locus, which was

found to be strongly associated with type 2 diabetes (Frayling et al., 2007). However when the investigators adjusted the association for BMI, the association fell away, thus implying that *FTO* was associated with BMI and not type 2 diabetes. They later replicated their finding in a larger cohort (both adult and childhood) and concurrently other researchers have replicated this locus as an obesity risk locus in a childhood obesity study (Dina et al., 2007). This was then followed by a meta-anaysis (Loos et al., 2008) where the *FTO* association was replicated and a new locus near *MC4R* was both discovered and further replicated. It became evident that there was a need for an increase in sample size through collaboration. The first major collaborative effort was from the Genetic Investigation of ANthropometic Traits (GIANT) consortium whose mandate is to gain insight into genetic loci that influence body size and shape, including height and measures of obesity.

The first publication by the GIANT consortium (Willer et al., 2009) together with a study by Thorleifsson et al. (2009) confirmed the previous findings with regard to *FTO* and *MC4R* and further identified and replicated eight additional BMI risk loci including those in/near *TMEM18*, *KCTD15*, *GNPDA2*, *SH2B1*, *MTCH2*, *NEGR1*, *ETV5* and *BDNF*. Following that, a landmark paper (Speliotes et al., 2010) increased the total of BMI-associated loci to 32 via a large increase in sample size (~124 000, replication in ~126 000 individuals). However the latest GWAS and Metabochip meta-analyses by GIANT has shown **97 loci** (refer to Fig.1.5 for an overview of the BMI-associated loci discovered to-date) to be robustly associated with BMI at genome-wide significance (Speliotes et al., 2010, Locke et al., 2015). This collaborative effort encompassed almost 340 000 individuals of mostly European (~322 000 individuals) descent and about ~17 000 individuals that were non-Europeans (mainly African American but also included Hispanic, Filipino, Jamaican and South Asian individuals) (Locke et al., 2015) (supplementary data 2). Of the 97 loci found to be significantly associated with BMI, 56 were novel to the study (Locke et al., 2015). Cumulatively it has been suggested that only ~3% of the total variation in BMI can be accounted for by the common genetic variants uncovered by GWAS (Andersen and Sandholt, 2015). Also a recent study has reported that when BMI was examined for age-specific effects, 15 BMI associated loci were discovered (four novel), with 11 of the loci showing larger effects in the younger cohort (Winkler et al., 2015). Even though missing heritability remains a problem (discussed further on), further investigations in populations with increased genetic diversity such as the current fine-mapping study may offer new avenues in terms of addressing the variation in adiposity traits accounted for by genetic variants.

NON-EUROPEAN POPULATIONS

Prior to 2012 most BMI-associated GWAS performed in non-Europeans did not report any GW significant findings, even though the variants were directionally consistent with those identified in European GWAS (Fall and Ingelsson, 2014). These studies included GWAS in a large mixed Asian cohort (Chinese, Malay and Indians from Singapore) (Dorajoo et al., 2012), a Filipino cohort (Croteau‐Chonka et al., 2011), a Native American cohort (Malhotra et al., 2011), a Japanese cohort investigating hypertension traits, including BMI (Hiura et al., 2010) and two GWAS including populations of African Ancestry (AA) (Kang et al., 2010, Monda et al., 2013). In 2012 two GWAS focusing on Asian populations, more specifically mixed Chinese, Korean and Indonesian cohorts (Wen et al., 2012) and a Japanese cohort (Okada et al., 2012) reported replication of signals from GIANT at seven and five obesity risk loci (including *FTO*, *BDNF* and *SEC16B*), respectively. These two studies also reported four novel non-European signals (near *CDKAL1*, *PCSK1*, *GP2* and *KLF9* genes) with *CDKAL1* seen in both Asian cohorts (Okada et al., 2012, Wen et al., 2012). Also, a GWAS meta-analysis of individuals of African ancestry (Monda et al., 2013) reported the discovery of three other novel BMI-associated loci near *GALNT10*, *MIR148-A-NFE2L3* and *KLHL32*, as well as replicating 32 out of the then 36 signals from GIANT (Speliotes et al., 2010), five of which were at GW-significance and the others at $P = 9.7 \times 10^{-7}$ (Monda et al., 2013). Investigation of variants in other populations has shown an overlap of variants associated with increased obesity risk in European populations, but also the discovery of 11 novel obesity risk loci (Andersen and Sandholt, 2015). Further investigation is warranted in individuals of non-European ancestry, which as a result of their greater genetic diversity, can better inform genetic studies. This is relevant for risk factors that vary in prevalence among different ancestral populations, and might impact the distribution of rare variants in WGS studies and population structure in GWAS.

3.2.2 GWAS Associated With Fat Distribution (waist circumference, hip circumference and waist-to-hip ratio)

EUROPEAN COHORTS

Visceral and abdominal fat distribution has been shown to be relevant in the development of metabolic disease, so determining the genetic factors that underpins this process, would be very relevant to the obesity phenotype. WC and WHR give an indication as to the distribution of body fat and in most studies are adjusted for BMI. This is because BMI is highly correlated with WC and WHR (Fall and Ingelsson, 2014). The effect of WC and WHR to fat distribution should be examined, taking BMI into consideration. To date there have been 69 loci associated with body fat distribution from GWAS (Andersen and Sandholt, 2015).

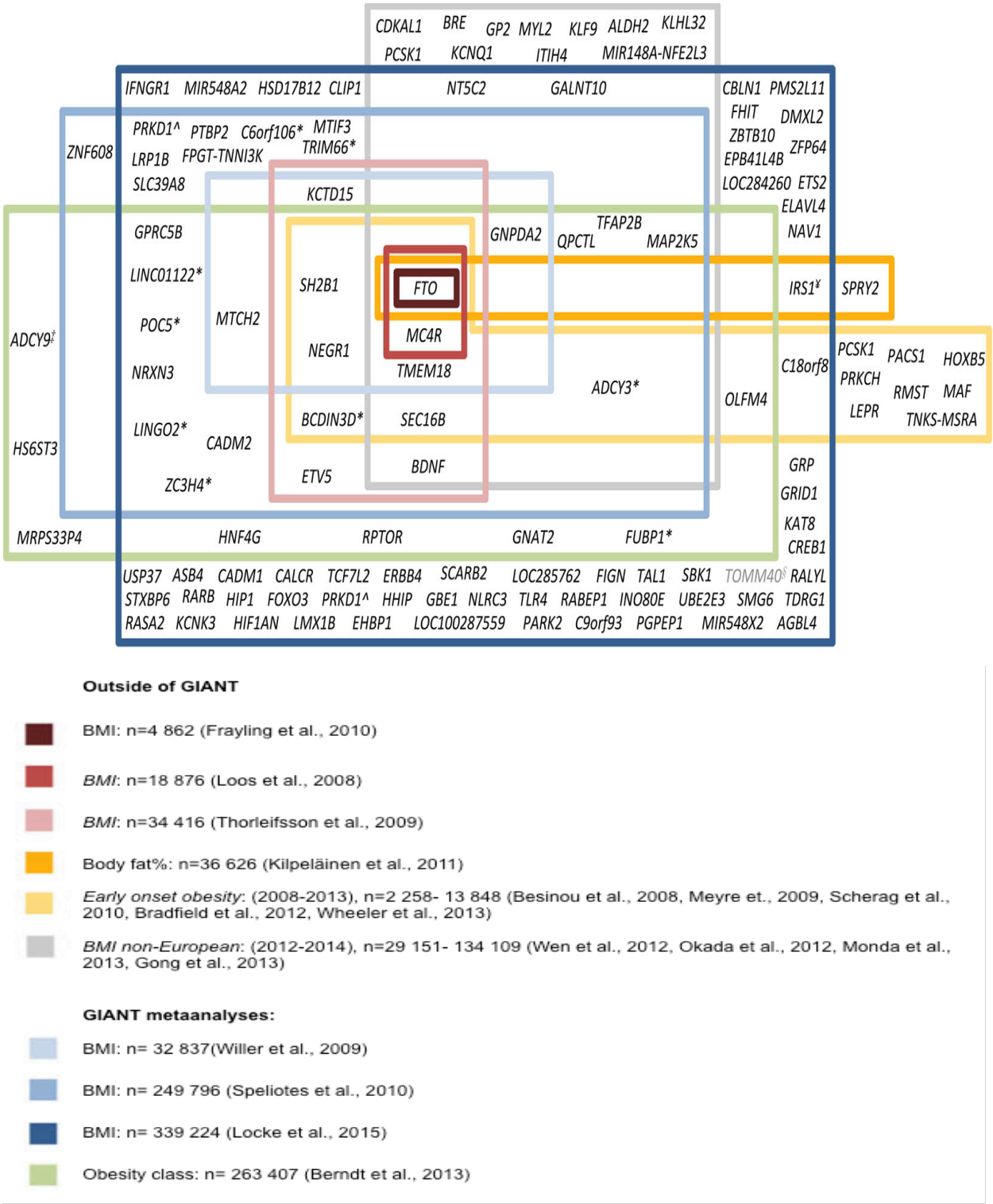


Figure 1.5 View of the current loci associated with increased obesity risk
Adapted from (Andersen and Sandholt, 2015)

The first study to show an association with WC (Frayling et al., 2007) showed *FTO* to be associated with both BMI and WC. This was followed by two meta-analyses spearheaded by the GIANT and CHARGE (Cohorts for Heart and Aging Research in Genome Epidemiology)

consortia where they examined both WC and WHR and WC only, respectively. The CHARGE study confirmed the *FTO* association with WC (Heard-Costa et al., 2009). The GIANT consortium paper identified two loci near *TFAP2B* and *MSRA* associated with WC and *LYPLAL1* to be associated with WHR in females only (Lindgren et al., 2009). Further, a meta-analysis by GIANT focussing on $WHR_{adjustBMI}$ reported the discovery of 13 new loci (highlighted in Fig.1.6) at GW-significance and replicated the known effect observed at *LYPLAL1* (Heid et al., 2010). They also reported that seven out of the 13 new loci showed sex-specific effects with a majority of the loci have a stronger effect on WHR in females over males.

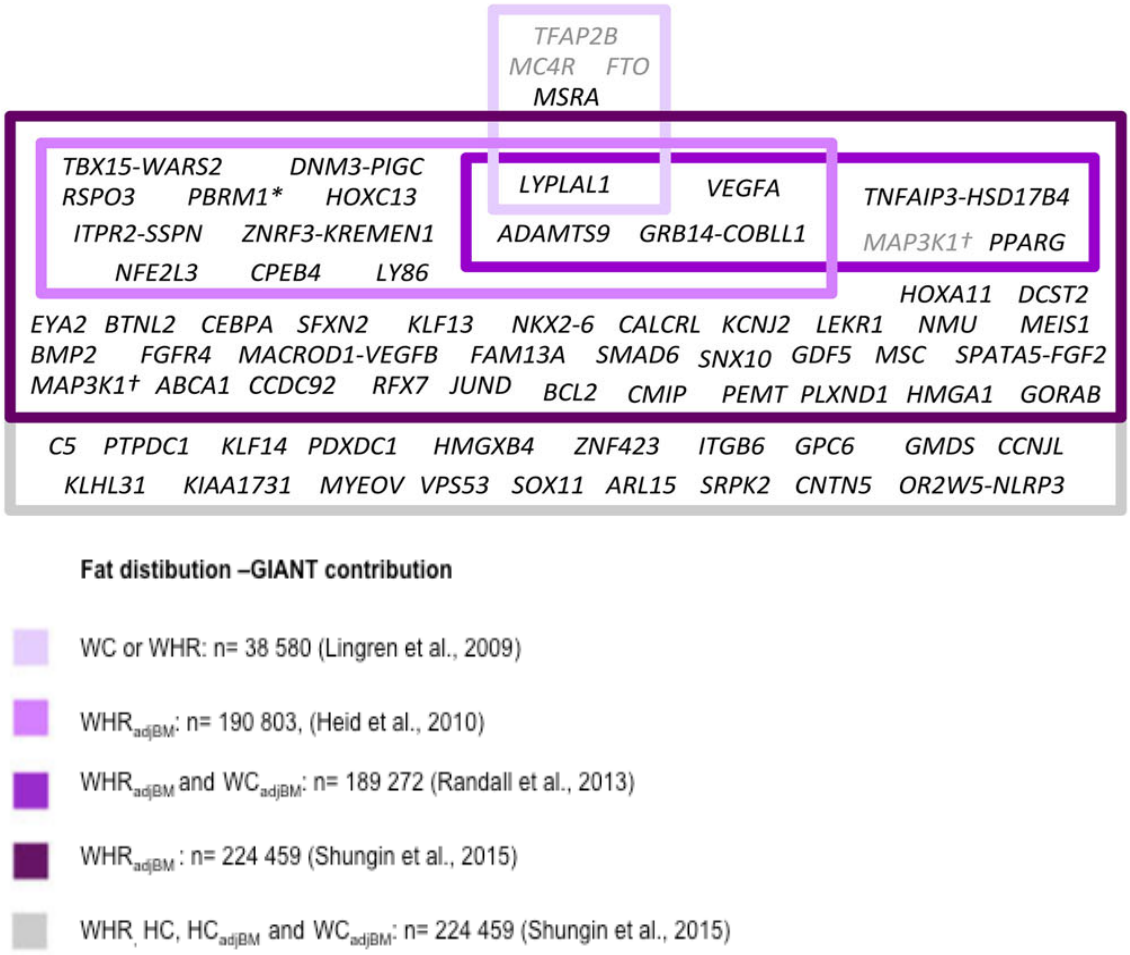


Figure 1.6 View of the current loci associated with body fat distribution
Adapted from (Andersen and Sandholt, 2015)

The most current GIANT study has robustly confirmed 49 loci (33 novel to the study) associated with $WHR_{adjustBMI,}$ 19 loci associated with WHR, $WC_{adjustBMI,}$ $HC_{adjustBMI}$ and HC (Shungin et al., 2015). HC and WC with and without adjustment for BMI were included as phenotypes to clarify the distribution of fat deposits independent of adiposity, which allowed the discovery of the additional 19 loci (Sandholt et al., 2015). There seems to be a strong

influence of sex in the regulation of fat distribution and most of the loci discovered regarding the distribution of body fat account for 1.36% of the phenotypic variance observed, more so in females (Andreasen et al., 2009, Randall et al., 2013). The sexual dimorphism effects on body shape ($WHR_{-adjust\ BMI}$) was recently confirmed by a meta-analysis examining age- and sex-specific effects where 44 loci were reported at GW-significance (17 novel), 28 of which had larger effects in females, five had larger effects in males and 11 of the loci had opposite effects in each sex (Winkler et al., 2015). This study also reported that there were no age-dependent effects with regard to $WHR_{adjustBMI}$ between older and younger adults.

NON EUROPEAN POPULATIONS

Few studies have been reported on the genetic influence on body fat distribution in non-European populations. A study carried out in Indian Asians showed a GW significant association between WC (and independently with insulin resistance) and *MC4R* (Chambers et al., 2008) while another study in Asian individuals reported a novel locus near *C12orf51* was associated with WHR (Cho et al., 2009). The novel discovery in the latter study is yet to be replicated and confirmed in other populations. The only GWAS study of body fat distribution conducted in African Americans, reported 2 novel loci near *LHX2* ($WC_{adjustBMI}$), *RREB1* ($WHR_{adjustBMI}$) (Liu et al., 2013) and overlap with six Euro-centric loci (Heid et al., 2010). There was nominal evidence for sexual dimorphism shown in this study.

3.2.3 Other GWAS relevant to obesity phenotype (Morbid Obesity, Early-Onset Obesity, Adiposity)

GWAS have been carried out in individuals with extreme phenotypes including morbid/extreme obesity in the expectation of discovering more penetrant variants than those found in the general population (Sandholt et al., 2015). These analyses were carried out predominantly in a case-control (obese vs. normal weight) fashion, however there was no consensus regarding inclusion criteria for obesity classes amongst the various studies. Criteria included established cut-offs for obesity or cut-offs based on percentile distribution of the trait (Fall and Ingelsson, 2014). A majority of the studies were conducted on extremely obese adult cohorts although two studies included child cohorts (Meyre et al., 2009, Scherag et al., 2010). All of the studies found overlap with SNPs previously reported for BMI or body fat distribution and only three studies reported finding novel variants near *MAF* (Meyre et al., 2009), *SDCCAG8* (Scherag et al., 2010) and *KNCMA1* (Jiao et al., 2011) genes. A study focussing on both extreme phenotypes associated with adiposity as well as stratifying the data according to clinical obesity classes, found no GW significant associations for the

former but found seven novel loci for the latter outlined in Fig. 1.5 (Berndt et al., 2013). Five of these associations were also found in the largest GIANT BMI meta-analysis (Locke et al., 2015), hinting at the notion that these loci might not really be associated specifically with extreme obesity classes, but rather correlated with the increase in sample size (Sandholt et al., 2015).

There are very few studies published on direct measures of adiposity derived from DXA and bioelectrical impedance. Two studies assessed derived measures of fat mass and obesity, one utilizing DXA scans (Liu et al., 2008b), the other bioelectrical impedance (Melka et al., 2011). This resulted in the discovery of loci associated with fat mass, *CTNNBL1* and *PAX5* and *MRP522*, respectively. One study (Kilpeläinen et al., 2011) used percentage fat mass and confirmed the association of *FTO* and adiposity as well as finding two novel loci, *IRS1* (associated with a decrease in % fat mass) and *SPRY2* at GW significance.

3.3 Fine mapping studies

GWAS studies help us better understand the biological mechanisms that underline the obesity phenotype however they do not necessarily pinpoint the causal and functional variant. Most of the variants that have been implicated in fine mapping studies may provide a solution for this by trying to identify all known variants within a specific region identified by GWAS and then honing in on the strongest signals in large validation cohorts or by combining data on the same regions from different populations with different LD backgrounds (Fall and Ingelsson, 2014). Populations with lower LD structure such as African or Asian populations are useful in unmasking causal variants that are hidden by higher levels of LD in European populations (Buyske et al., 2012, Crawford et al., 2013). Fine mapping studies in populations of varying ethnicities are also useful in the identification of additional causal variants that may not be present or occur at extremely low frequencies in other discovery populations (Buyske et al., 2012).

The most current fine mapping approaches have shown that non-coding variants are likely to be responsible for associations at established loci and that newly discovered loci had a large number of enhancer elements (Horikoshi et al., 2015). Recent studies have fine-mapped genome-wide obesity-related loci (Gong et al., 2013, Liu et al., 2014, Horikoshi et al., 2015) and some have chosen to focus on single regions (candidate fine-mapping) such as *FTO* or *SH2B1* (Hassanein et al., 2010, Peters et al., 2013, Volckmar et al., 2015). Most of the cohorts examined have been European populations but the use of African ancestry populations (Gong et al., 2013, Peters et al., 2013, Liu et al., 2014), has shown that fine-

mapping across populations of differing LD backgrounds reduces the number of common alleles that are likely to be functional candidates and is therefore a powerful tool in gaining insight into the causal variants for polygenic obesity.

To drive the discovery of additional loci, and to hone in on the actual causal SNPs, higher resolution SNP arrays have been developed to examine trait/s of particular interest; these include the dense genotyping/fine mapping arrays such as the Metabochip (Voight et al., 2012).

3.3.1 The Metabochip

The Metabochip is a custom genotyping array from Illumina that allows for the genotyping of almost 200 000 SNPs known to influence cardiometabolic and atherosclerotic traits (Voight et al., 2012). The chip contains 11 main traits captured in Tier 1, of which BMI and WHR are included (Table 1.2). The SNPs included on the chip are the top (highest association signals) ~5000 (Tier 1) and ~1000 (Tier 2) independent association signals obtained from GWAS meta-analyses for the 23 traits captured. These include SNPs from the International HapMap and 1000G project [refer to (Voight et al., 2012) for breakdown of SNPs]. Together with each GWAS-identified SNP, additional proxy SNPs (that are in high LD with the index SNP; r2> 0.9) together with four other supplementary SNPs (r2> 0.5) were selected from the CEU (Utah residents with Northern and Western European ancestry) and YRI (Youruba, Nigeria) HapMap Phase II datasets, respectively, to be included on the array. Also SNPs were included for fine-mapping purposes to elucidate regions related to metabolic traits, as well as Major Histocompatibility Complex (MHC) SNPs, copy number variant (CNV)-tagging SNPs, mitochondrial DNA SNPs, X and Y chromosome SNPs as well as “wildcard” SNPs (Voight et al., 2012, Spencer et al., 2013). “Wildcard” SNPs – refers to additional SNPs of particular interest to each consortia present on the chip.

The Metabochip differs from GWAS chips in its SNP coverage of rare and low frequency variants (Buyske et al., 2012). This describes variants with a MAF of < 1%. For example there is a notable difference in MAF captured by Metabochip SNPs in African Americans versus ASW (African ancestry in Southwest USA) compared to Affymetrix 6.0 and Illumina 1M arrays. A total of 21.6% of the polymorphic Metabochip SNPs have MAF less than 0.025, compared to 5.8% and 6.8% for the Affymetrix 6.0 and Illumina1 M arrays, respectively (Buyske et al., 2012).

Studies assessing the generalisation of GWAS signals in European discovery populations in non-Europeans populations using the Metabochip have shown directionally consistent

effects. However, the differential effect sizes of these variants vary, more so in populations of African Ancestry (Carlson et al., 2013) and this is due mostly to differing LD backgrounds rather than true differences in effect sizes.

Table 1.2 Breakdown of SNPs of the Metabochip by trait and tier. Tier 1 contains 11 traits of primary interest, Tier 2 contains 12 traits of secondary interest. Adapted from Voight et al., (2012)

Consortium	Trait Name	Fine Mapping			Replication SNPs
		# Loci	Size (Mb)	# SNPs	
Tier 1					
DIAGRAM	Type 2 Diabetes	34	6.56	16 717	5 067
CARDIoGRAM	MI and CAD	30	9.60	19 558	6 485
Lipids	HDL Cholesterol	23	4.62	12 150	5 024
	LDL Cholesterol	21	4.06	9 981	5 060
GIANT	Triglyceride	20	4.68	9 784	5 057
	Body Mass Index	24	7.48	18 211	5 055
MAGIC	Waist -to-hip ratio*	15	2.25	5 464	5 056
ICBP	Fasting Glucose	19	5.05	13 644	5 058
	Diastolic Blood Pressure	20	8.34	13 239	5 060
QT-IGC	Systolic Blood Pressure	21	6.01	10 641	5 059
	QT Interval	18	4.08	10 910	5 041
Tier 2					
DIAGRAM	T2D Age of Diagnosis	0	0.00	0	1 039
	T2D Early Onset	0	0.00	0	1 040
HaemGen	Mean Platelet Volume	0	0.00	0	657
	Platelet Count	0	0.00	0	577
	White Blood Cell	0	0.00	0	598
Lipids	Total Cholesterol	0	0.00	0	941
Body Fat	Body Fat Percentage	0	0.00	0	1 035
GIANT	Height	0	0.00	0	1 050
	Waist Circumference*	2	0.50	1 374	1 048
MAGIC	24-Hour Glucose	3	0.61	1 249	1 038
	Glycated Haemoglobin	5	0.46	2 181	1 045
	Fasting Insulin	2	0.67	1 309	1 046
TOTAL	with Redundancy	257	64.97	146 453	68 126
	Unique Regions/SNPs	257	45.52	122 241	63 450

*Waist -to-hip ratio and waist circumference were adjusted for body mass index

Many studies have demonstrated the usefulness of the Metabochip in fine mapping signals at a particular locus as well as population-specific signals for example, at the *CELSR2/PSRC1/SORT1* locus, a study reported the strongest associated SNP for LDL cholesterol to be rs12740374 (P= 3.6 x 10^{-11}). This particular variant could not be discriminated from other SNPs in high LD in European cohorts (Buyske et al., 2012). Further they also reported a SNP rs17231520 at *CEPT* to be associated with HDL-cholesterol in African Americans, which occurs at a very low frequency in European populations and would not have been tagged on normal GWAS chips.

More specifically with regard to obesity, one study has demonstrated the effective use of the Metabochip to densely genotype and evaluate 21 BMI loci identified in European GWAS in ~29 000 African Americans from the Population Architecture using Genomics and Epidemiology (PAGE) study (Gong et al., 2013). They showed that eight of the 21 loci were associated with BMI at P= 5.8 x 10^{-5} and that for most of the loci, fewer variants were in LD (r^2>0.5) with the most significant SNP in African American populations than European populations. Further, they also reported two new variants associated with BMI, *BRE* and *DHX34* as well as a putative independent signal near *GNPDA2* (Gong et al., 2013). It is unknown how this chip will perform in a sub-Saharan African population but it has potential to narrow down 'risk intervals' in susceptibility loci revealed by GWAS and also potentially will aid in the discovery of population-specific loci, as will be examined in this study.

4. African Data from Candidate Gene Studies and GWAS

Reports on the prevalence of obesity in Africa are few with approximately 20 population-based studies published throughout Africa (Kengne et al., 2013). The prevalence of obesity varies throughout Africa, with rates as low as 5,3% in Uganda (rural setting) and as high as 45,7% in rural South Africa (Kengne et al., 2013, Yako et al., 2015). The reason for these varying rates may not only be a reflection of the differing dietary environments or socio-economic statuses but also the frequency of obesity-associated genetic variants. Generally it is understood that obesity is a complex disease resulting from the cumulative effect of "at-risk" variants described previously in the chapter when "at-risk" carriers are exposed to an obesogenic environment (Yako et al., 2015). According to a recent review paper (Yako et al., 2015), over 300 studies have been conducted in Africa regarding genetic variants associated with obesity risk with variants in genes such as *ACE, ADIPOQ, ADRB2, AGRP, AR, CAPN10, CD36, C7orf31, DRD4, FTO, MC3R, MC4R, SGIP1, LEP* and *LEPR* having

been reported.

Prior to the work outlined in this thesis, only the *FTO* and *MC4R* GWAS-validated loci have been investigated for correlations with BMI in a South African (Lombard et al., 2012), Gambian (Hennig et al., 2009) and Nigerian cohort (Adeyemo et al., 2010). Both the South African and the Nigerian study found associations with variants in *FTO*. The Gambian study investigating correlations in *FTO* and BMI, did not report any significant associations, which may be due to a number of reasons, including a very lean study cohort (Hennig et al., 2009). Most studies investigating obesity-related loci in African populations adopted a candidate-gene approach and there is not a lot of overlap in loci examined, barring *FTO, LEP* and *LEPR* (>three studies) (Yako et al., 2015). This lack of overlap between studies of various African populations prevents the comparison of effects of obesity-linked variants. Also most studies on African populations have been conducted in South Africa, Tunisia, Nigeria or the Gambia with single studies conducted on obesity loci in other African populations. The studies are not consistent with regard to *sample size* (ranging from N= 85 – 2 332; with five studies N>1000), *age* (ranging from 2-93yrs) or *phenotype* (all including BMI, although WC, WHR, percentage fat mass and BMI as a dichotomous trait have been reported) (Yako et al., 2015).

Only some studies accounted for confounding factors in the association analysis including population stratification (Yako et al., 2015). There is a growing body of evidence to suggest that variations in ethnicity associate with genetic variation and that population structure is greater within Africa than elsewhere (Teo et al., 2010) and therefore must be considered. Also African populations exhibit the highest levels of genetic variation and the lowest levels of LD among human populations (Tekola-Ayele et al., 2013). A meta-analysis including individuals of AA (Monda et al., 2013) found that 32 of the loci associated with BMI (Speliotes et al., 2010) were directionally consistent with European populations albeit not GW significant. A similar Asian GWAS (Wen et al., 2012) also confirmed the lack of reproducibility of GW significant signals in non-European populations. To date there has been one GWAS study conducted in an African (Nigerian) cohort (Kang et al., 2010) on anthropometric traits. There could be a number of reasons responsible for the lack of reproducibility of these signals including inconsistent genotyping strategies across the studies, smaller sample sizes, or the difference in the genetic architecture and LD patterns (Teo et al., 2010, Yako et al., 2015).

In summary, **188 loci** have been reported to associate with obesity-related phenotypes with major contributions from the GIANT BMI (97 loci) and WHR_{adjBMI} (49 loci) meta-analyses,

GIANT analyses of obesity classes (three loci) and other fat distribution measures (19 loci), as well as independent GWAS of early-onset obesity (eight loci), non-European populations (11 loci) and phenotypes (one locus) (Sandholt et al., 2015). The small amount of variation in BMI explained by data from GWAS in populations of European descent suggests that studies must be undertaken in other ethnicities that have lower LD levels and are more polymorphic. African populations meet both these requirements and the detailed analysis of relevant obesity phenotypes within such populations could lead to the identification of additional genetic risk markers for obesity, including those associated with fat distribution and whole body adiposity.

5. Missing heritability

A common thread mentioned throughout this literature review speaks to the high heritability of body composition traits that contribute to obesity. The fact that the level of heritability of height is 80-90% (Visscher, 2008) was aptly described by Maher (2008) "… if 29 centimetres separate the tallest 5% of a population from the shortest, then genetics would account for as many as 27 of them." Despite the high heritability estimates (40-70%) for obesity related traits, GWAS-identified variants are only able to explain a very small percentage of this heritability (Maher, 2008, Bogardus, 2009, Manolio et al., 2009). Plausible explanations for this "missing heritability" could lie with variants that are too rare to be identified by GWAS that have large effects on body fat mass, and the inability of GWAS to detect other variant types e.g. copy-number variations (Maher, 2008, Manolio et al., 2009, Marian, 2012). Also that there are a lot of frequent variants that are not penetrant (Maher, 2008), so GWAS cannot statistically detect those associations. It has also been suggested that gene-gene, gene-environment interactions as well as the contribution of microRNAs in bringing about epigenetic changes might explain some aspects of hidden heritability (Marian, 2012).

However a majority of what we currently know about genetic factors linked to obesity is focussed on European and some Asian studies, with African populations remaining understudied. There are significant advantages in studying African populations, in that the high genetic diversity and low linkage disequilibrium allows for narrowing down of loci where previous signals have been identified (Teo et al., 2010). This could potentially result in unveiling and discovery of causal variants. The main motivation for this study is the lack of genetic studies, either association or heritability estimates, of traits contributing to obesity in diverse sub-Saharan African populations.

6. Study Aim and Objectives

Differences have been observed in obesity risk factors, distribution of body fat and environmental and nutritional exposure when comparing European and African populations. Given these observations, this study focuses on exploring whether genetic factors contribute to measures of body fat in African subjects. The aims of this study are therefore to: (1) identify genetic markers associated with body fat and lean measures in a South African black population and (2) establish if the heritability estimates of these body composition measures derived from European populations are similar in African populations.

To achieve these overall aims, the following objectives were explored:

i. Conduct a pilot study to assess the feasibility and power to detect GWAS-derived Eurocentric SNP associations of obesity in an African population via a candidate gene association study performed in an adolescent subset of the Birth to Twenty Cohort.
ii. Perform a genome-wide replication and fine-mapping study by genotyping Birth to Twenty cohort participants (young adults and older caregivers) using the Illumina Metabochip array to assess if the same variants associated with body fat and lean measures in European populations can be observed in an African population.
iii. Calculate the heritability estimates of each of the anthropometric measures described above using the data from the Metabochip array, and compare to data from heritability studies conducted in European populations. The heritability will be estimated using software called “genome wide complex trait analysis (GCTA)”.

Chapter 2

Pilot study: Exploring genetic markers of adult obesity risk in black adolescent South Africans - the Birth to Twenty Cohort

The aim of this pilot study was to assess the association of six genetic variants known to be associated with BMI in Europeans, in a subset of the Birth to Twenty cohort. This would provide preliminary evidence that common variants influence this complex trait in an indigenous African population. Adolescent data were used to facilitate the identification of genetic loci that predispose to obesity early in life, as it is known that overweight/obese children have an elevated risk of becoming obese adults (Singh et al., 2008, Lundeen et al., 2015). This study focused on genetic variants shown to be associated with increased BMI in both adult and childhood cohorts as undertaken by the GIANT consortium (Thorleifsson et al., 2009, Willer et al., 2009, Speliotes et al., 2010). A previous study (Lombard et al., 2012) showed that BMI was associated with genetic variants known to influence appetite control in adolescent black South Africans. Some aspects of the work presented in this chapter were not included in the manuscript (Pillay et al., 2015), due to editorial limitations of the journal (refer to Appendix G for a copy of the manuscript).

1. Introduction

Recent studies show that the mean prevalence of overweight and obesity (combined) in South African children and adolescents is approximately 15% (Shisana et al., 2013). The prevention of childhood obesity is a key global health priority as obesity is a major contributor to increased mortality in adulthood (Reilly and Kelly, 2010), and is increasing in prevalence in both HIC and LMIC, such as South Africa.

The risk of developing obesity is modulated by both heritable and environmental factors (Wardle et al., 2008). Heritability studies of body mass index (BMI) demonstrated that a significant proportion of the variance in BMI (40-70%) is due to heritable factors e.g. genetics, epigenetics, intrauterine environment (Bodurtha et al., 1990) with genome-wide association studies (GWAS) of obesity-related traits identifying more than 97 risk loci thus far (Loos, 2012, Fall and Ingelsson, 2014, Locke et al., 2015). An explicit deficit of African-centric GWAS data for body composition and obesity exist, especially for sub-Saharan African populations. In addition, most published genetic association studies of BMI have primarily focused on the association with adult BMI. Identifying loci that predispose to obesity early in life could provide a better understanding of the early determinants of adult obesity and may also uncover potential new targets for the therapeutic prevention of obesity.

Previously, Lombard et al., (2012) investigated the role of gene variants in appetite regulating genes with BMI in an adolescent cohort, replicating SNP associations in *FTO* and *MC4R*, as well as establishing a novel association with variants in the *LEP* gene. In this paper, the

replication within the same cohort, of six variants are described, based on previous evidence of robust association with BMI in non-African populations in large meta-analyses undertaken by the GIANT consortium (Thorleifsson et al., 2009, Willer et al., 2009, Speliotes et al., 2010). These include SNPs in or near *GNPDA2*, *MTCH2*, *NEGR1*, *SH2B1*, *STK33* and *TMEM18*.

2. Materials and Methods

2.1 Subjects

Bt20 is Africa's largest and longest running birth cohort study, focused on the health and development of children born in Soweto, South Africa (Richter et al., 2007). Initially in 1990, the Bt20 cohort began to track the development of 3,273 newborn infants and their caregivers. The participants in the Bt20 cohort are South African blacks who self-identified as Sotho speakers (speaking southeastern Bantu languages), thus belonging to the Niger-Kordofanian ethno-linguistic group. The inclusion criteria included all births in a seven week window (April 23 to June 8, 1990) and if the mothers enrolled in the study resided in Soweto-Johannesburg for six months following delivery (Richter et al., 2007).

Various biological data have been collected from the Bt20 cohort and caregivers for use in several studies over the years following informed consent. A subset of individuals (43%, n = 990) from the Bt20 cohort were randomly selected for this study, and consisted of 524 (53%) female and 466 (47%) male adolescents, with a mean (± SD) age of 13.7 ± 0.2 years. Phenotype data including age, sex, BMI and pubertal stage collected from these individuals at year 13 were used in the study. Anthropometric measurements were obtained using standard methods (Cameron et al., 1998) where BMI was computed as weight (measured in kg) divided by the square of the height (measured in meters) of an individual. The pubertal stage was assessed using a validated self-assessment method (Norris and Richter, 2005). BMI is the most commonly used measure of obesity and for comparison purposes we utilised BMI in adolescents. Due to the lack of published data in African populations, standardized child BMI categories (BMI-specific age and sex cut-offs) according to Cole et al. (2007) were used. Written assent was obtained from all adolescents in conjunction with written consent from parents/legal guardians, prior to a blood sample collection. The most current consent forms are included in Appendix H. It is also important to note that at every data collection wave, participants reconsent for the data collected. With regard to genomic studies there is a

form of blanket consent (refer to reconsent forms in Appendix H), however each respective study (including analyses) have to be approved by the Wits Human Research Ethics committee (medical) (HREC).

Ethics approval from the HREC (Certificate nr. M120647- Appendix H) was obtained to use the DNA samples in the study. The Wits HREC previously permitted the collection of DNA samples and phenotype data from the adolescent cohort (Certificate nr. M010556- Appendix H).

2.2. SNP selection, DNA extraction and genotyping platform

SNPs previously associated with BMI (Thorleifsson et al., 2009, Willer et al., 2009, Speliotes et al., 2010) from GWAS were selected for analysis, and included rs2568958 near *NEGR1*, rs6548238 near *TMEM18*, rs10938397 near *GNPDA2*, rs10769908 in *STK33*, rs10838738 in *MTCH2* and rs7498665 in *SH2B1*. DNA was extracted using a salting-out technique (Miller, et al., 1988). Participant DNA samples are housed within the National Health Laboratory Service (NHLS) at the University of the Witwatersrand, Division of Human Genetics bio-repository. DNA was quantified using absorbance spectroscopy (Tecan Infinite® 200 PRO NanoQuant), and normalized to a concentration (50ng/ul) required for genotyping. Genotyping was performed using the GoldenGate™ VeraCode assay (Illumina, San Diego, CA, USA) (steps outlined in Fig. 2.1). Internal quality control was performed on all raw genotype data according to the supplier's specifications using the genotyping module of BeadStudio (Framework version 3.1.3.0; Illumina, San Diego, CA). Further quality control filters based on minor allele frequency (MAF<0.01) and Hardy–Weinberg equilibrium (HWE<0.05) were used as exclusion criteria. Ancestry information markers (used to assess population substructure) previously genotyped in this cohort show no significant global population structure within the cohort (Lombard et al., 2012). Ancestry infornative markers are a subset of genetic markers that differ in frequencies across different continental populations (Kosoy et al., 2009).

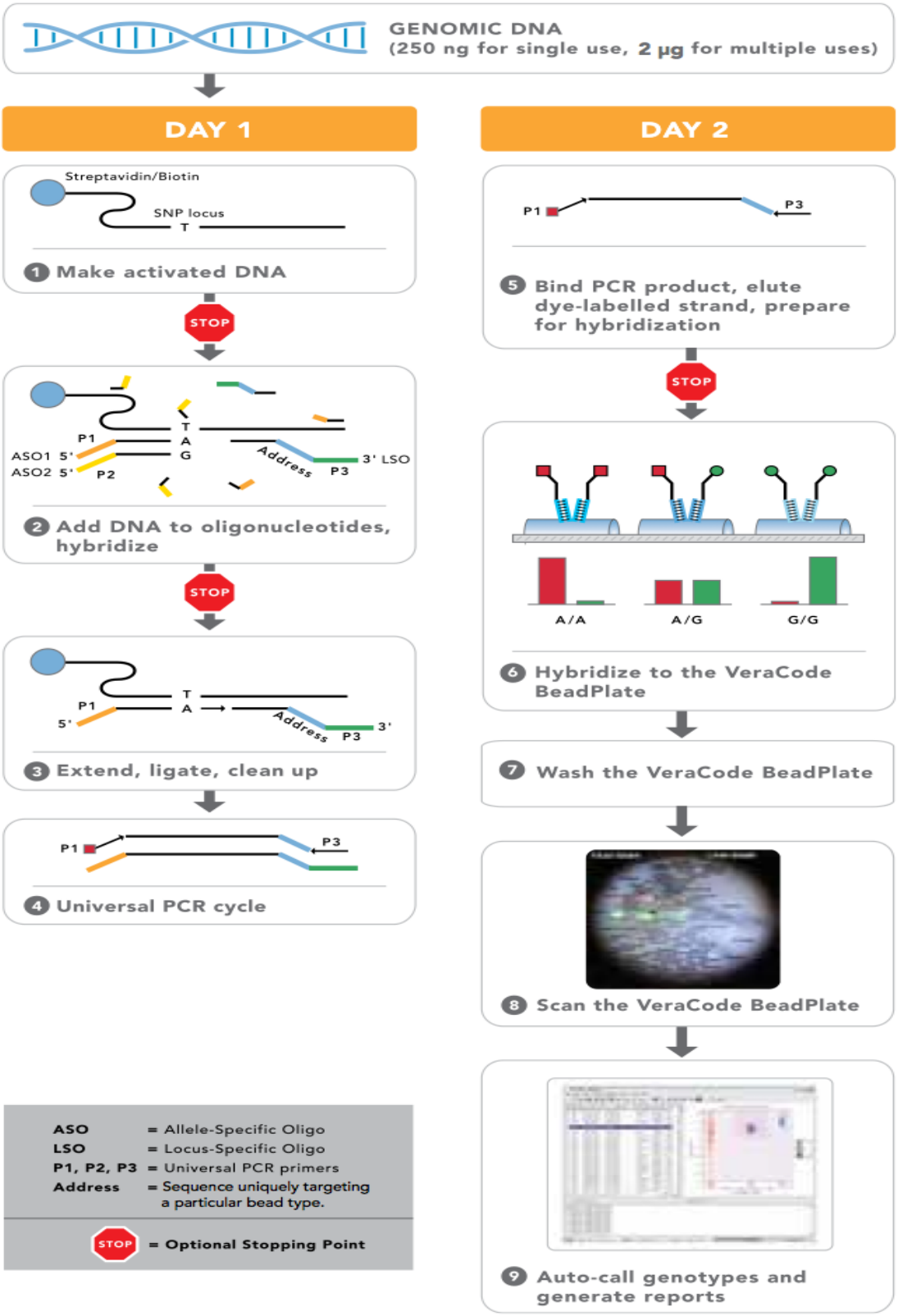


Figure 2.1 Outline of the workflow used in VeraCode GoldenGate
(Obtained from Technical Note: SNP Genotyping: GoldenGate™ VeraCode assay, Illumina)

2.3 Statistical Analysis

PLINK v.1.9 was used for all statistical analyses, unless stated otherwise (Purcell et al., 2007). The distribution of BMI was skewed and therefore was log-transformed to approximate normality for all analyses. Linear regression was used to assess the association of the selected SNPs with BMI. As BMI correlated significantly with sex, pubertal stage and age, analyses were adjusted for these variables by including them in the linear model as covariates (P_{adj}). Given the strong prior information about the correlation of the SNPs tested here with BMI, this was considered this a replication study, and therefore, *P* values below 0.05 were considered significant. Given the minor allele frequencies of the SNPs tested, the study achieves 80% power to detect differences among the means of BMI, with a standard F-test for linear regression, as previously calculated (Lombard et al., 2012). To assess the combined impact of risk alleles on BMI, a risk allele score was calculated by summing the number of BMI-increasing alleles per individual. This score was calculated including the risk alleles described here, as well as the four previously identified BMI-associated variants in this cohort: *FTO* (rs17817449), *LEP* (rs10954174 and rs6966536) and *MC4R* (rs17782313) (Lombard et al., 2012).

The BMI across genotypes for each SNP was also assessed using ANCOVA in Statistica v11.0 (StatSoft, Tulsa, OK, USA), after adjusting for age, sex and pubertal stage. The Tukeys *post hoc* test was used to examine the difference between genotype group medians after adjusting for the effect the covariates mentioned above. To visualise the relationships between BMI and the genotypes box and whisker plots were constructed in Microsoft Excel for only the significant SNPs observed in the association testing.

2.4 Computational predictions

RegulomeDB is a database that contains manually curated regions from the ENCODE project, CHIP-seq information, chromatin state information, expression quantitative trait loci (eQTL) data together with computational predictions and predicting DNAase footprinting (Boyle et al., 2012). This database was used to elucidate the functionality of the SNP variants found to be signficantly associated with BMI. Lower scores in this database indicate increasing evidence for a variant to be located in a functional region as outlined in Table 2.1. There are sub-categories within category 1 that indicate additional information about the confidence of the annotation with 1a being the most confident (containing evidence for transcription factor (TF) binding, a motif for that TF, and a DNase footprint) to the lowest sub category 1f (containing only TF binding or a DNase peak) (Boyle et al., 2012).

Table 2.1 The classification of variants using RegulomeDB
Classification of variants according to RegulomeDB using a large body of regulatory information as reported by (Boyle et al., 2012)

Category	Description
	Likely to affect binding and linked to expression of a gene target
1a	eQTL + TF binding + matched TF motif + matched DNase footprint + DNase peak
1b	eQTL + TF binding + any motif + matched DNase footprint + DNase peak
1c	eQTL + TF binding + matched TF motif + DNase peak
1d	eQTL + TF binding + any motif + DNase peak
1e	eQTL + TF binding + matched TF motif
1f	eQTL + TF binding/DNase peak
	Likely to affect binding
2a	TF binding + matched TF motif + matched DNase footprint + DNase peak
2b	TF binding + any motif + matched DNase footprint + DNase peak
2c	TF binding + matched TF motif + DNase peak
	Less likely to affect binding
3a	TF binding + any motif + DNase peak
3b	TF binding + matched TF motif
	Minimal binding evidence
4	TF binding + DNase peak
5	TF binding or DNase peak
6	Motif hit

eQTL-refers to expression quantitative trait loci, TF- refers to transcription factor binding sites; DNase- refers to DNase footprinting to demonstrate protein binding sites.

3. Results

The study group consisted of 524 (53%) female and 466 (47%) male adolescents, with a mean (± SD) age of 13.7 ± 0.2 years. Summary statistics and trends related to BMI in this subset were described previously (Lombard et al., 2012), and based on these data all analyses were adjusted for sex, sex-specific pubertal stage and age.

Table 2.2 reports the association of the SNP variants with (log)BMI. Of the six SNPs investigated, three were replicated in this African cohort, and showed a similar (albeit smaller) directional effect to that observed in the discovery studies. Significant correlations were identified between BMI and rs10938397 (effect allele-G) near *GNPDA2* (P_{adj} = 0.003),

rs7498665 (effect allele-G) in *SH2B1* (P_{adj} = 0.014), and with rs6548238 (effect allele-C) near *TMEM18* (P_{adj} = 0.030).

Table 2.2 Results of the six SNPs tested for association with log(BMI) in the Bt20

				Alleles[1]				*P* values		
Nearest gene	**SNP**	**Chr**	**N**	**Effect[2]**	**Other**	**Frequency effect allele**	**Frequency effect allele***	**P_{unadj}**	**P_{adj}[3]**	**Effect size[4] Beta (s.e.m)**
GNPDA2	rs10938397	4	961	G	A	0.21	0.45	0.001	**0.003**	0.013(0.004)
MTCH2	rs10838738	11	982	G	A	0.07	0.36	0.680	0.739	-0.003(0.007)
NEGR1	rs2568958	1	976	A	G	0.44	0.67	0.713	0.837	0.001(0.004)
SH2B1	rs7498665	16	963	G	A	0.29	0.38	0.086	**0.014**	0.007(0.004)
STK33	rs10769908	11	985	C	T	0.25	0.57	0.663	0.701	-0.002(0.004)
TMEM18	rs6548238	2	982	C	T	0.92	0.15	0.086	**0.029**	0.011(0.006)

[1]Allele coding according to the forward strand (NCBI dbSNP Build 134). [2]Effect allele associated with increased BMI in the original study. [3]P-values are adjusted for age, sex and sex-specific pubertal stage. [4]Effect sizes in log (kg/m^2). Significant P-values are shown in bold. *- refers to frequency of effect allele in Caucasian populations.

To assess the combined impact of risk alleles on BMI, a risk allele score (Fig. 2.2) was calculated, in which the three risk alleles currently described were included, as well as four previously identified risk alleles in the same cohort (rs17817449 in *FTO*, rs10954174, rs6966536 in *LEP* and rs17782313 in *MC4R*) (Lombard et al., 2012). The difference in average BMI between individuals with a high genetic-susceptibility score (defined as having ≥10 BMI-increasing alleles) and those with a low genetic-susceptibility score (≤4 BMI-increasing alleles) was 3.90 kg/m^2, signifying a 21.7% increase in average BMI between these two groups. In comparison, if only the three SNPs from this study are used, the increase in average BMI (comparing the lowest to highest risk score) is 2.06 kg/m^2, signifying a 10.5% increase in average BMI between these two groups. This implies that both set of risk SNPs drive the effect observed when all SNPs are used.

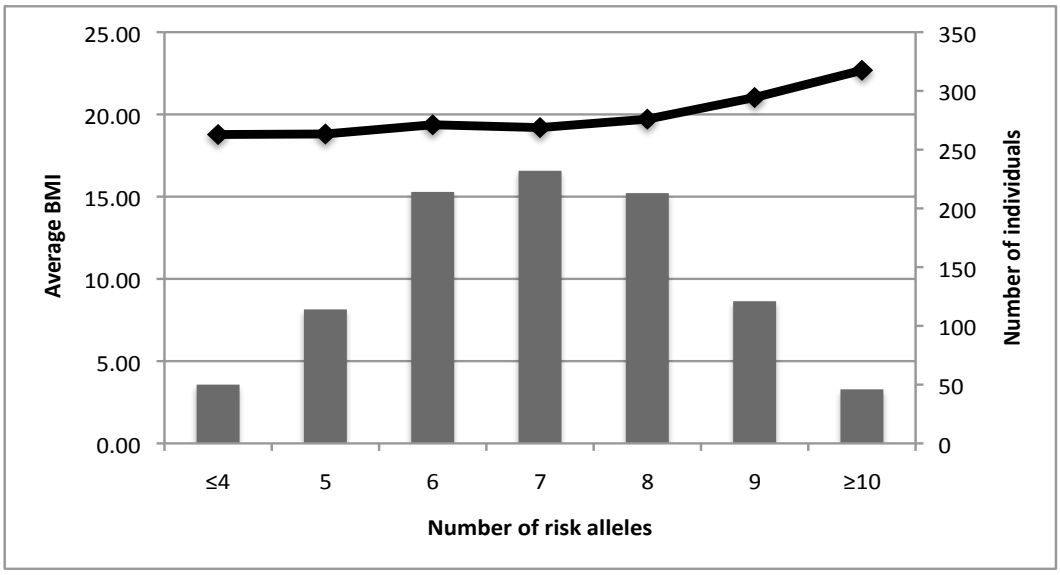


Figure 2.2 Combined impact of risk alleles on average BMI in the Bt20 cohort.
Risk alleles were summed for each individual. The number of individuals in each risk allele category is shown along the x-axis

BMI was assessed across the genotypes for each SNP in Table 2.3. The CC genotype (*TMEM18*) and the AA genotype (*GNPDA2*) were significantly associated with an increase and a decrease in BMI, respectively. Figures 3.2 a, b and c refer to box and whisker plots illustrating the relationship between BMI and genotypes for significant SNPs, *TMEM18* (a), *GNDPA2* (b) and *SH2B1* (c).

Table 2.3 Showing genotypes associated with each SNP tested and BMI

Gene		Data		
NEGR1	***Genotype***	**AA**	**AG**	**GG**
	N	180	492	304
	BMI#	18.52 (17.18, 20.43)	18.93 (17.17, 21.49)	18.51 (16.97, 20.87)
TMEM18	***Genotype***	**CC**	**CT**	**TT**
	N	827	150	5
	BMI	18.77 (17.15, 21.34)*	18.45 (16.82, 19.96)	18.57 (18.19, 19.41)
GNPDA2	***Genotype***	**AA**	**AG**	**GG**
	N	597	316	48
	BMI	18.53 (16.92, 20.65)**	19.06 (17.34, 21.90)	18.77 (16.97, 21.29)
STK33	***Genotype***	**CC**	**CT**	**TT**
	N	60	368	557
	BMI	18.34 (17.11, 21.25)	18.75 (17.05, 20.97)	18.65 (17.15, 21.09)
MTCH2	***Genotype***	**AA**	**AG**	**GG**
	N	851	124	7
	BMI	18.69 (17.06, 21.21)	18.94 (17.45, 20.79)	18.44 (14.63, 19.13)
SH2B1	***Genotype***	**AA**	**AG**	**GG**
	N	481	413	69
	BMI	18.75 (17.22, 20.81)	18.53 (16.95, 20.82)	19.08 (17.26, 22.49)

BMI data# expressed as median (interquartile range); **P*<0.05, ***P*<0.005 vs. heterozygote in ANCOVA analysis adjusted for age, sex and pubertal stage. NC ran the data analysis in Statistica for this table.

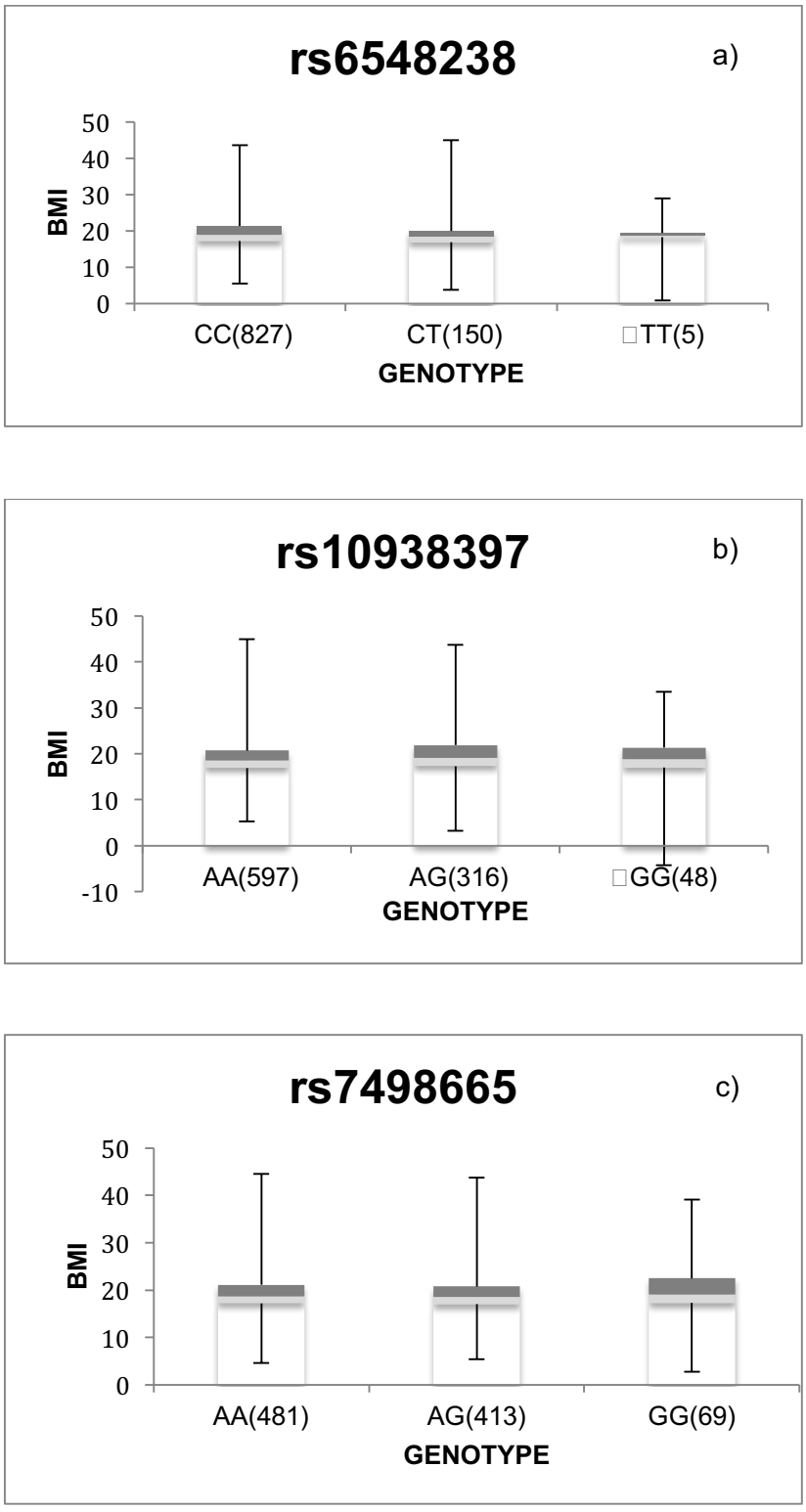


Figure 2.3 Box and whisker plots showing the association of genotype with BMI - for the significantly associated SNPs: a) *TMEM18*, b) *GNPDA2*, c) *SH2B1*, where the number of samples with a particular genotypes are shown in brackets

Results tabulated in Table 2.4 show that the variant observed in our study, rs7498665 in/near *SH2B1,* had a score (1f) that is suggestive of being "likely to affect binding and linked to expression of a gene target" (Boyle et al., 2012). rs10938397 in/near *GNPDA2* has no data available on the RegulomeDB database as to its proposed functionality and rs6548238 in/near *TMEM18* has a score of (4) which shows minimal evidence of binding any targets.

Table 2.4 Results of RegulomeDB functional analysis

Gene	Chr	BP start position[1]	SNP	Regulome DB Score	Proposed Action
SH2B1	16	28883240	rs7498665	1f	Likely to affect binding and linked to expression of a gene target
TMEM18	2	634904	rs6548238	4	Minimal binding evidence
GNPDA2	4	45182526	rs10938397	-	No Data

Chr, chromosome number[1], Chromosomal base pair position is based on NCBI dbSNP Build 134.

4. Discussion and Conclusion

Several common genetic variants have been robustly associated with adult obesity risk. This study provides confirmation that three of these variants are also associated with BMI in a South African cohort of adolescents. The G-allele of both rs7498665 (*SH2B1)* and rs10938397 (*GNPDA2)*, and the C-allele of rs6548238 (*TMEM18)* were shown to be associated with an increase in BMI. A similar, albeit smaller, directional effect was observed to that seen in the discovery studies in non-African cohorts. Further investigation of the genotypes associated with significant SNPs and the resulting effect on BMI showed that the CC genotype associated with *TMEM18* (Table 2.3 and Fig. 2.3a) and the AA genotype associated with *GNPDA2* (Table 2.3 and Fig. 2.3b) were significantly associated with an increase (P<0.05) and a decrease in BMI (P<0.005), respectively. These results correlate with the allelic association analysis conducted in PLINK. The GG genotype associated with *SH2B1* also was associated with an increase in BMI, but was not statistically significant. This was probably due to the small number of individuals (N=66) with that genotype. These results demonstrate that genetic variants for adult BMI are also associated with BMI earlier in life, which may provide insights into the genetic aetiology of obesity within an indigenous African population.

GNPDA2, *TMEM18* and *SH2B1* are all highly plausible biological candidates for adiposity – all three are expressed in the brain, with evidence of a role in appetite regulation or affecting adipose tissue biology (Speakman, 2013). *GNPDA2* is expressed in the hypothalamus, alluding to a neuronal influence on energy balance, and has been associated with BMI in both paediatric (Zhao et al., 2009) and adult cohorts, including a replication in an African-American cohort (Gong et al., 2013). *SH2B1*'s link to metabolic function is well established (Ren et al., 2007) and deletions in this gene are associated with severe early-onset obesity (Bochukova et al., 2010). RegulomeDB aims to annotate the functional importance of non-coding variants, since most inter-individual changes (captured by GWAS studies) occur outside of coding variants (Boyle et al., 2012). It also aids in the interpretation of what role these potentially regulatory variants will play biologically. The database was queried for the three variants found to be significantly associated with BMI in our study. Data on two of the three variants associated with increased BMI in this study was uninformative with no information available for rs10938397, whilst rs6548238 showed minimal binding evidence. The accuracy of RegulomeDB is very much limited by the availability of experimental data however, the database still remains highly informative when information on variants is available as in the case of rs7498665. The score generated by the database predicted that this variant will likely affect binding and expression of a target gene. Furthermore, rs7498665 in *SH2B1* is a missense-coding variant and results in a substitution of alanine with threonine, which likely affects protein activity and expression. TMEM18 is ubiquitously expressed, and although a direct link to obesity is still elusive, early evidence suggests a likely role through transcriptional regulation of critical targets (Speakman, 2013).

A number of limitations are acknowledged in our study. The sample size is moderate and therefore not powered to detect small effects on BMI, suggesting that the potential contribution of *MTCH2* and *STK33* warrants follow-up investigation in a larger cohort. The significant differences in the genomic structure between African and non-African genomes, could lead to a situation where SNPs shown to be associated with a trait in European populations may be weak predictors for causal variants in African populations, due to differences in linkage disequilibrium (Teo et al., 2010). Another consideration is that, although BMI is an established obesity index, it is not the best indicator of adiposity (Wells, 2014) and the use of more suitable measures may help to elucidate the role of genetics in adiposity.

Finally, the data in this study are derived from a cohort of adolescents in the midst of puberty. It is therefore possible that the effects on weight of some polymorphisms may have been masked by puberty-associated changes in body fat mass. Furthermore, the effects of some

polymorphisms on BMI are conceivably only observed later in life. Elucidating the genetic component of obesity in children is important because it may uncover factors that have a stronger phenotypic effect than those gene variants that only become apparent after years of exposure to an obesogenic environment. Also, gene variants that give rise to childhood obesity may provide important information on important metabolic or neurological pathways that could be therapeutically manipulated to reduce adipose tissue accumulation. The discovery of such polymorphisms may also help identify individuals with a high risk of obesity and hence allow early lifestyle interventions.

In conclusion, our study has replicated associations for increased BMI with SNPs present in or near *TMEM18*, *SH2B1* and *GNPDA2* in an African adolescent population. These observations suggest that variants in these genes or neighbouring loci may be important in body weight regulation in divergent populations.

Chapter 3

The Metabochip as a tool to identify genetic variants associated with obesity and body composition in African subjects

1. Introduction

Following the outcomes of the exploratory study outlined in Chapter 2 where three obesity risk variants in or near *GNPDA2*, *SH2B1* and *TMEM18* were found to be significantly correlated with changes in BMI in adolescents, it encouraged us to further pursue the investigation of obesity risk variants in a larger adult cohort (within Bt20) of black South Africans. The purpose of this part of the study was to replicate and fine map in an African population signals captured from GWAS that were associated with obesity in European subjects.

GWAS (described in the literature review) are based on LD and incorporate mostly common genetic variation in populations that are in LD or tag other variants that are not directly typed on an array (Crawford et al., 2013). This is in keeping with the common-disease-common-variant hypothesis. Due to the fact that very few GWAS variants have been identified as true 'risk' variants, it is difficult to elucidate their potential contributions to the underlying biology of the particular disease under investigation. Fine-mapping, which is the dense interrogation of GWAS-identified loci using many more SNP markers (Crawford et al., 2013), may be useful in identifying the true causal variants.

Given that African populations have lower LDs compared to European and Asian populations, fine-mapping African populations may offer greater value in identifying potential risk variants, masked by the high LD in European populations. Fine-mapping studies in African populations are limited to mainly African American populations. The Metabochip (described earlier in Chapter 1) was used in this study to assess both fine-mapping and replication of about ~200 000 SNP markers. These markers encompass the top-signals from GWAS together with dense coverage around the GWAS-associated loci, related to metabolic diseases and traits including obesity. This would provide the most comprehensive and cost-effective approach toward replicating GWAS-discovered signals in a novel population as well as provide the opportunity to discover new population-specific obesity-related signals. The transferability of obesity-associated risk variants across different populations are not well understood, more so for African populations, and this study provides the opportunity to assess this.

2. Materials and Methods

2.1 Participants, study design and sample selection

In this part of the study, both phenotype data and DNA samples were used from the participants (young adults) and the caregivers of the Birth to Twenty (Bt20) cohort that has been described in detail elsewhere (Richter et al., 2009) and briefly in Chapter 2. Following informed consent, blood samples were collected and DNA extracted using a salting-out method (Miller et al., 1988) (highlighted earlier in Chapter 2). The DNA is currently stored at the Division of Human Genetics, National Health Laboratory Service (NHLS) and the University of the Witwatersrand, Johannesburg, South Africa. This study was approved by the Human Research Ethics Committee (Medical) of the University of the Witwatersrand (Ethics clearance certificates number M010556 and M120647, Appendix H).

Participants (N=1240) and female caregivers (N=1033) from the Bt20 cohort, in whom phenotype data collected was at age 17/18 (young adults) and age 40 (middle-age; caregivers), respectively were selected for this study. This included 973 caregiver-young adult pairs, with 61 unrelated caregivers and 275 unrelated young adults. The term “caregiver” describes the person or relative who accompanied the participant during assessment and in most cases is the mother of the participant. If the mother was unavailable or deceased then the term refers to a female relative (either maternal grandmother, sister or aunt), who accompanied the participant (personal communication, SA Norris, 15 July 2014).

2.2 Sample preparation and genotyping

Prior to genotyping, the DNA samples (isolated using salting-out) were normalized to a concentration of 50ng/µl using the Tecan Freedom EVO® (Tecan Trading AG, Switzerland). This was done following quantification using either the Tecan Infinite® 200 PRO NanoQuant (Tecan Trading AG, Switzerland) or PicoGreen® dsDNA Quantitation Reagent (Thermo Fisher Scientific, Wilmington, Delaware USA). Genotyping was performed using the Illumina Metabochip (Illumina, San Diego, CA, USA), which is an Illumina iSelect assay (Illumina, San Diego, CA, USA). Genotyping was performed at the DNA Technologies Core of the University of California Davis, California, USA by the genotyping services team. Genotyping was performed in two batches – first for the caregiver samples, and finally for the participant sample. A set of duplicate samples from each batch was sent together with the unique samples to rule out batch effects on the chip and also to ensure that samples were genotyped consistently on the same chip. Necessary documentation (materials transfer

agreement and export permit) was obtained prior to shipment (Appendix H). We obtained permission to access the phenotype data together with stored DNA samples in accordance with Developmental Pathways and Health Research Unit (DPHRU's) policy for data requests.

2.3 Power analysis calculation

The power to detect associations was assessed using the software package Power GWAS/QT vs.1.0 that uses the F-test (Feng et al., 2011). In this test an *a priori* power analysis option was assumed, which allowed the calculation of the statistical power of this study to detect an association if SNPs explained a certain percentage of the heritability of a particular trait with a given sample size N. This programme is designed for quantitative traits. The heritability estimates as well as the effect sizes for the power calculations were based on European populations from Speliotes et al. (2010) because there are no other data available from African ancestry populations therefore we applied the next-best proxy from Caucasian populations.

Results from power analyses (Power plots are shown in Appendix A) showed that with PowerGWAS/QT using an F-test with a 0.05 significance level (standard for linear regression models) we have 80% power to detect associations with effect sizes of 0.035 in 1000 individuals.

2.4 Phenotype data

Weight, height, age, waist circumference (WC), hip circumference (HC), subtotal fat mass (grams) and subtotal fat-free mass (grams) (excluding bone mineral content) were measured and recorded in all young adults and caregivers. Subsequently, BMI, percentage fat mass (PFM) and waist-to-hip ratio (WHR) were calculated from these measures. Anthropometric measurements were obtained using standard methods (Cameron et al., 1998). Weight was measured with a digital scale (Dismed, Halfway House, South Africa) and height with a stadiometer (Holtain, Crosswell, UK), with light clothing and shoes removed. BMI was computed as weight (measured in kg) divided by the square of height (measured in meters) of an individual. The methods for obtaining WC, HC and body composition measurements were according to (Feeley, 2012). Briefly, WC was measured with subjects standing, with a soft measuring tape, at the level of the smallest girth above the umbilicus and HC was measured at the widest part of the buttocks (Feeley, 2012). Both WC and HC were measured in centimetres. Waist to hip ratio was computed as WC divided by HC. Body

composition readings were obtained using dual energy X-ray Absorptiometry (DXA, Hologic, Bedford, USA) as per guidelines recommended by the International Society of Clinical Densitometry (Feeley, 2012). Subjects were asked to remove all items of jewellery and other metal objects, and wore light clothing. Total body fat and lean mass (in grams) were measured. PFM was calculated as fat mass (in grams) divided by total fat mass (in grams). DXA scans for the whole body did not include measurements for the head as many subjects wore wigs and hair attachments (weaves) that would result in inaccurate measurements because of similar densities to that of soft tissue (Feeley et al., 2012).

Pubertal stage was not included in the young adults in the subsequent analyses, as it was expected that the majority of participants, would have finished puberty already at this age. Another study assessing fracture patterns and bone mass density in the same cohort excluded the effects of puberty in the young adults in the analysis due to a lack of correlation (Thandrayen et al., 2014).

2.5 Quality Control (QC)

The quality of both genotype and phenotype data was assessed in the dataset to try to remove potential confounders that could affect the downstream association analysis.

2.5.1 Phenotype Data QC

Following the receipt of the phenotype data, some inconsistencies were noted between sample information recorded in DNA biobank database and the phenotype database. An extensive quality control exercise was conducted by retrieving original patient files, where available, to remove ambiguous data, and to confirm that DNA profiles matched the phenotype dataset. In other words, it was confirmed that DNA samples and phenotype data collected on the day, was from the same person. It was also confirmed that the phenotype data that were recorded in the original patient files, matched the information captured in the phenotype database.

2.5.2 Genotype Data QC

Genotypes were called using GenomeStudio (vs. 2011.1) (Illumina, San Diego, CA, USA), with calls based on a modified clustering manifest trained on the sample data. Final data reports were produced in the forward strand orientation. The raw genotyping data was converted into a PLINK-compatible format that was used in all downstream applications (Appendix B). A rigorous quality control assessment was performed on genotype data according to published quality control filters (Anderson et al., 2010, Clarke et al., 2011). The QC process outlined in Fig. 3.1 was divided into three main steps:

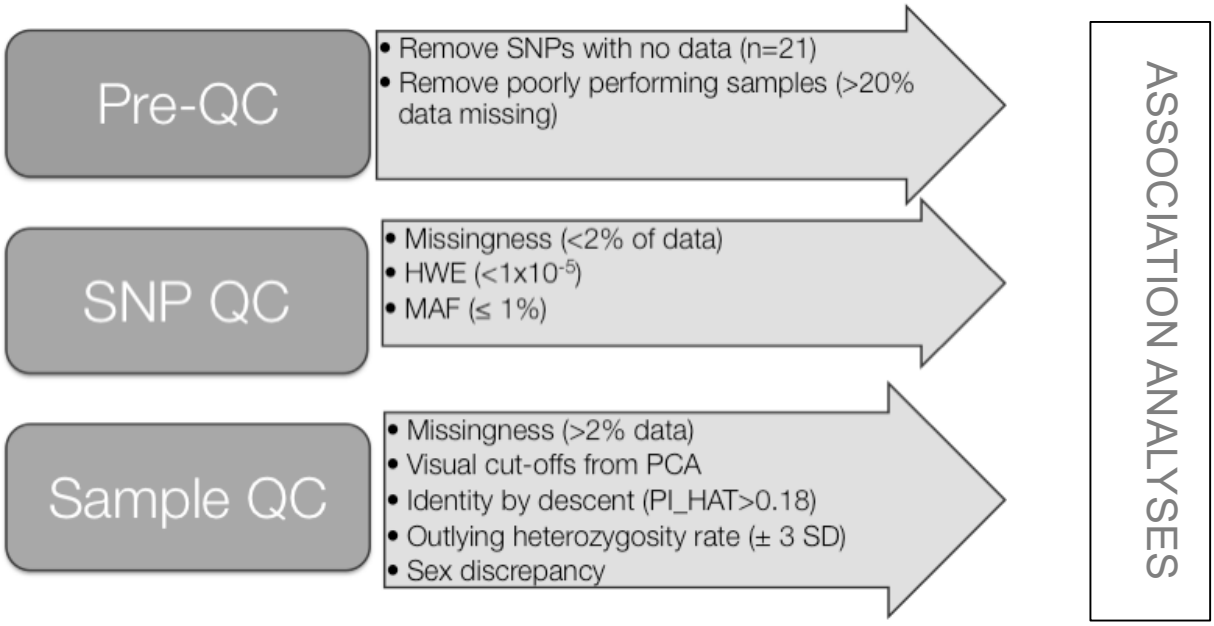


Figure 3.1 Overview of the quality control process, outlining the key steps and criteria selected for the analysis. HWE- refers to Hardy-Weinberg equilibrium; MAF- minor allele frequency; PCA- principal component analysis, SD – standard deviations; n= sample size

PRE-QC STEP

This step involved the conversion of the final reports in the forward orientation into tped/tfam (transposed) format using the script *convert2tped* (Appendix B). Subsequently, data was converted into a binary format (bed files) using PLINK v.1.9 (Purcell et al., 2007)[1]. '.bed' files are a primary representation of genotype calls as biallelic variants and are usually accompanied by .bim and .fam files (Purcell et al., 2007). During genotype calling, SNPs that cannot be called by the software for various reasons are termed "NaN" SNPs. All NaN SNPs were removed at this stage together with samples that had greater than 20% of the genotypes missing. At this stage of the analysis it was decided to QC batches separately, due genotyping performance, before merging the final datasets.

All of steps below were performed in PLINK unless otherwise stated using default cut-offs (Anderson et al., 2010) or cut-offs determined using this data.

[1] Any further mention of PLINK will refer to PLINK v. 1.9 Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., M.A.R, F., Bender, D., Maller, J., Sklar, P., De Bakker, P., Daly, M., et al. (2007). PLINK: a toolset for whole-genome association and population-based linkage analysis. *Am J Hum Genet,* 81.

SNP QC

1. **SNP missingness** distribution was assessed based on plots of maximum missing rate vs. number of SNPs remaining in the study where thresholds (0.02) were used for both batches. If there was high SNP missingness (values >2%) then the SNPs were discarded from the analysis. High SNP missingness implies that there is an inadequate separation of that SNP into a particular cluster and is inaccurately called and therefore has to be removed.
2. The distribution of **minor allele frequency (**MAF) was assessed and a default MAF > 0.01 was set for both batches according to (Anderson et al., 2010). Genotype calling algorithms tend to perform inefficiently for SNPs with a low MAF. Therefore SNPs that had a MAF of <1% were discarded.
3. A range of *P* value thresholds were tested for deviations from **Hardy-Weinberg equilibrium** (HWE) and a default value of $P < 1x10^{-5}$ was selected as a cut-off for both the batches according to Anderson et al., (2010). Extreme deviations from HWE can be due to genotyping error or any violations of HWE assumptions. SNPs were then filtered out if they showed extreme HWE deviation based on the above cut-off.

SAMPLE QC

1. A **sample missingness** threshold was determined by plotting the maximum missing rate vs. number of samples remaining in the study. The threshold for sample missingness was set to 2% and 3% for Batches 1 and 2, respectively which means that samples were excluded if the genotype information for that samples was less than 98% or 97%, respectively.
2. A **sex check** was conducted on the raw genotype data. Here, homozygosity on the X-chromosome was estimated (HXE) to determine 'genetic sex'. Males have a single copy of the X- chromosome therefore males are assumed to be hemizygous for all X chromosome SNPs (excluding the pseudo autosomal region). Males have HXE >0.80, females have HXE: <0.20-0.35 and those coded as 'ambiguous sex' have estimates in the range of 0.35>HXE<0.80. The 'genetic sex' was estimated using the above estimates and compared to the sex information provided with the phenotype data and samples were removed from both batches for discordant sex information. Inconsistencies may be due to mis-labeling of samples which may have arisen during sample collection or sex being incorrectly reported during recruitment. All inconsistencies were reported to the biobank where the DNA samples are stored and to the project manager of the cohort.

3. To ensure that the individuals within each batch were unrelated (i.e. the maximum relatedness between any pair of individuals is less than a second degree relative) (Anderson et al., 2010, Laurie et al., 2010) a criteria called the 'Identity by state' (IBS) was calculated. This score is calculated for each pair of samples and is based on the average proportion of alleles shared in common for genotyped SNPs excluding the sex chromosomes. The IBS scores rely on the SNPs being unlinked; therefore the data was pruned (to remove regions of extended LD) using a 50kb window. **Identity by descent** (IBD) scores were estimated from IBS data in PLINK. IBD, pi_hat =1 for duplicates/monozygotic twins, pi_hat =0.5 for first-degree relatives, pi_hat =0.25 for second-degree relatives and pi_hat =0.125 for third degree relatives. We removed all samples in the dataset with IBD pi_hat scores > 0.1875 (halfway between 2nd and 3rd degree relatives).
4. Duplicates (included in the study for QC purposes) were also removed using IBD scores where pi_hat =1.
5. Individuals were then removed based on outlying **heterozygosity rates**. Heterozygosity rates were calculated by dividing the number of total non-missing genotypes (N) - homozygous genotypes (0) by the total non-missing genotypes (N). Excess heterozygosity gives an indication of possible sample contamination whilst less than expected heterozygosity rates indicate possible inbreeding. The threshold for inclusion of samples is within ±3 standard deviations (SD). Samples were removed outside of the set cut-offs.
6. Principal component analysis (PCA) plots were constructed (*smartpca*) in EIGENSTRAT-vs.3.0 (HelixSystems, Maryland, USA) and Genesis (http://www.bioinf.wits.ac.za/software/genesis/) was used to visualize PC plots. Association analyses can be confounded by population structure. Population structure may be present when an allele is more prevalent in one population over another, resulting in a spurious association between the trait being tested for and any genetic characteristics, which vary between the two different groups of people (Price et al., 2006). PCA plots allow us to examine if population substructure exists in dataset by contextualising genetic variation using various population groups. Genome-wide data previously generated for the Bt20 cohort (May et al., 2013) were pruned and combined with this data to reflect only SNPs in common. The same was done for other African 1000 genomes (1000G) datasets (YRI-Yoruba from Ibadan in Nigeria, LWK-Luhya from Webuye, Kenya, MKK-Maasai from Kinyawa in Kenya) and data from southeastern Bantu-speakers (SEB) and southwestern Bantu-speakers (SWB) (Schlebusch et al., 2012) were included in the analysis. Outliers were removed manually from both batches using visual cut-offs.

Following the removal of both SNPs and samples that performed below the quality parameters described above, the resultant dataset was used for the association analysis.

2.6 Association analysis

A higher frequency of a genotype in individuals within a cohort that have a particular phenotype (e.g. higher BMIs) could mean that genotype is correlated with an increased risk for a particular disease (Lewis and Knight, 2012). An estimation of these associations can be gained through a variety of tests of which the 'allelic test' is the simplest. Along with the allelic and genotypic tests, additive, dominant and recessive models can be employed. However additive genetic tests are most commonly used with complex traits because of the possible contribution of several genetic variants to that trait. Within these tests significant associations can be determined by employing a variety of statistical methods depending on the sample composition as well as traits/phenotypes being tested. All statistical analyses were performed with PLINK for unmerged data or GCTA (genome-wide complex trait analysis) vs.1.24 (Yang et al., 2011a)[2] for merged data. In analysis scenarios where related individuals were combined it was necessary to use GCTA as PLINK cannot factor in relatedness. In this study the association between each SNP and BMI (natural log-transformed), WC, HC, WHR, FM, LM and PFM were estimated with both a basic association test and linear regression analysis.

2.6.1 Basic Association Testing (without covariates)

The appropriate genotype files (.bed, .bim and .fam) together with a phenotype file (.txt file where phenotypes are listed in columns) are required as input files for PLINK. The statistical determination of genotype-phenotype associations were assessed with Wald's t-test within PLINK by initiating the --assoc flag. This test generates a 't'-statistic based on a distribution of 't' as well as an asymptotic *P*-value. Within PLINK the flag --assoc-name allows more than one phenotype to be called and tested using one text file. The name of the phenotype should always correspond to the column identifier within the phenotype file. The --ci 0.95 generates 95% confidence intervals for the estimated parameters, in addition to the L95 (lower confidence interval) and U95 (upper confidence interval) fields in the output files.

Different strategies were employed for association testing, to exploit the disparate age and sex dynamic in this dataset. The choice of strategy is later addressed in Chapter 5, under

[2] Any further mention of GCTA will refer to GCTA vs.1.24 Yang, J., Lee, S. H., Goddard, M. E.&Visscher, P. M. (2011a). GCTA: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics,* 88**,** 76-82.

the limitations, as to how the composition in terms of age and sex might influence the outcomes of the analysis. The strategies are summarised below:

- Females only (combining females from the young adult dataset with female caregivers),
- All samples merged,
- Young males and females merged,
- Young females only,
- Older females only and
- Young males only.

Data will be presented on results of different association strategies outlined in the results section.

To ensure that population structure was controlled for by removal of the outliers, the genomic control method (GC) was flagged during the association analysis using PLINK v.1.9 to calculate the genomic inflation (GI) scores for each phenotype by including the --adjust flag. According to Devlin and colleagues (2001), this method uses inherent characteristics of the genome to correct for stratification. This is performed by estimating the degree of "overdispersion" of the statistics generated and is used to assess association (Devlin et al., 2001).

Results from genomic control (GC) indicated genomic inflation (GI) scores of 1.01 for BMI, 1.00 for WC, 1.00 for HC, 1.00 for WHR, 1.00 for fat mass, 1.01 lean mass and 1.00 for PFM.

2.6.2 Linear Regression (with covariates)

Each phenotype was analysed using linear regression under an additive model with adjustments for covariates (listed in Table 3.1) for each of the strategies outlined above. Within PLINK the flags --covar (txt. file) together with --covar-name allows more than one covariate to be called and included in the analysis. The covariates included in the analysis are *age*, *height* (was included in the analysis where necessary, outlined in Table 3.1) *and sex* (sex was only included, when young males and females are combined). The --linear, --ci 0.95, --pheno, --pheno-name, --covar and --covar-name flags were called in PLINK to generate a plink.assoc.linear output file which generates *P* values together with effect sizes (beta) estimated from the regression analysis for each analysis scenario.

Table 3.1 Linear regression models used for analysis of anthropometric variables

Phenotypes	Model 1	Model 2
	Adjusted for AGE only	Adjusted for AGE and HEIGHT
BMI	✓	
Waist circumference		✓
Hip circumference		✓
Waist-to-hip ratio	✓	
Fat mass		✓
Fat-free mass (lean)		✓
Percentage fat mass		✓

✓–Data are presented for phenotypes adjusted for these models in the results section

2.6.3 Mixed Linear Model Association (MLMA)

Due to the fact that individuals from the datasets were related when we merged datasets, mixed linear model association analyses (MLMA) were employed using the programme GCTA where the effect of genetic relatedness was taken into account. Both datasets were merged to increase the sample size and therefore power to detect genetic associations (De Bakker et al., 2005, Spencer et al., 2009). However, because these subjects are related, this decreases the power of the test to detect genetic associations, because the estimate of variances contributed by the SNPs are biased because of the shared environment of related individuals (Yang et al., 2011a).

Linear mixed models (LMMs) have been used increasingly in GWAS as a statistical tool for identifying genetic associations and trying to evade confounding factors like relatedness and population substructure (Zhou and Stephens, 2012, Eu-Ahsunthornwattana et al., 2014). LMMs integrate both “fixed effects” and “random effects” (i.e., “mixed effects”). The independent variables in a linear regression may be thought of as fixed effects. In order to find the random effects in a mixed model, something should be known about the variance and covariance of these random effects. The covariance structure of these random effects follows a polygenic (multifactorial) model and integrates the genetic relationship (kinship) between each pair of individuals (Eu-Ahsunthornwattana et al., 2014). GWAS MLMA uses this kinship matrix to correct for cryptic relatedness as a random effect and can include any additional fixed effects in the model.

Briefly, GCTA calculated a relatedness/kinship matrix based on the pairwise covariance between genotypes and then estimated the effect of each SNP on the phenotype while controlling for the relatedness matrix with a linear mixed model. An additive model of inheritance was assumed and all phenotypes were treated as continuous variables. Also the linear effects of age, sex, principal components and relatedness were controlled for during the analysis. The method is explained below in greater detail. The scripts used in the analysis can be found in Appendix C.

Merging the two data sets:

- Following QC both datasets (caregivers and young adults) were merged using the --bmerge flag in PLINK. Sample identifiers were updated for the individual identifier (IID) column in the .fam file using the (--update-id) flag in PLINK. This was done to distinguish which batch they were from as related samples share the same family identifier (FID).
- Mitochondrial DNA (MtDNA) SNPs were removed together with non-overlapping SNPs to create a prunedbmerge.* file.
- Following the merging of the datasets we re-ran basic QC on the dataset and removed any additional SNPs and samples accordingly. A new prunedbmerge2* was created for further downstream analysis.
- This file (prunedbmerge2.*) was then used to create the combined_females.* file by excluding all the male samples in the analysis (--exclude in PLINK) .
- The phenotype files were also merged and sorted according to the FID column to create a sorted_combined_batches.phen file. (GCTA requires the phenotype file to be a text file with.phen appended).
- The first step for MLMA in GCTA is to produce a genetic relationship matrix (GRM) between pairs of individual from the prunedmerge2.* or combined_female.* datasets. By incorporating fixed effects such as MAF (0.01) and PCA's (10), principal components are created (.eigenvec and eigenval) which can be used as covariates in the downstream analysis (make_grm2.sh in Appendix C).
- The appropriate covariate models were then created (using Linux commands), by combining principal components (eigenvectors generated by --pca command above when the --make-grm flag is called in GCTA) and the various covariate columns from sorted_combined_batches.phen for both the combined_all as well as the

*Refers to the input files required by PLINK (.bed, .bim, .fam)

combined_female analysis. The model names are the same for linear regression as outlined in Table 3.1.

- Once the GRM and covariate models were created (GRM_combinedall.**/ GRM_combinedfemales.**) they were inputted in the script explained below for association analysis.
- A script mlma_2.sh (Appendix C) was written to include all the phenotypes (eight) as well as taking into account all the various linear regression models, (this is explained in Appendix C) to create output files for each phenotype run with each respective model.

2.7 Adjusting for multiple testing

Multiple testing concerns come about when many hypotheses are tested at the same time, with some test statistics showing significance even if there are no real associations (Dudbridge and Gusnanto, 2008, Han et al., 2009). It is important to reduce type 2 errors (false-negative associations) or the inability to detect an effect that is present. According to Mayr et al. (2007) a statistical test can result in non-significance for two reasons; the null is accepted and is retained correctly or the alternative hypothesis holds but the test has been inefficient at detecting deviations from the null.

In this study the Bonferroni correction was applied to adjust for multiple testing. The Bonferroni genome wide (GW) significance level for Metabochip data was calculated by using only unlinked loci on the Metabochip. This was achieved by running the command: --indep-pairwise 50 5 x --out ldpruned, where x = a range of LD (proxy is r^2) values from 0.1 - 0.9 that were tested. This command considers windows of 50 SNPs at a time and calculates LD between each pair of SNPs in that window and removes one pair of SNPs if LD is greater than the value stipulated by x. It then shifts the window five SNPs forward and reiterates the process for the next 50 SNPs. Various r^2 values were tested to establish a cut-off value before the greatest number of SNPs were lost.

An r^2 of 0.5 resulted in 82231 SNPs remaining following LD pruning. Due to the fact that the Bonferroni test assumes independent tests and therefore only unlinked markers, the "Bonferroni genome-wide significance level for Metabochip" SNPs was calculated by dividing α=0.05 by the number of independent tests (unlinked Metabochip SNPs) i.e. (0.05/82231)

** --make-grm creates output files .grm.bin, .grm.N.bin and .grm.id.

which gives $P \leq 6.1 \times 10^{-7}$. Given the prior information about the association of SNPs captured on the Metabochip with obesity risk as well as the small sample size, P values $\leq 5 \times 10^{-5}$ were considered as suggestive of association and potentially interesting leads. To address the possible introduction of Type II errors through the application of this rigorous correction, we chose to also present a second category of results where a cut-off of $P \leq 5 \times 10^{-5}$ were met.

2.8 Visualisation of results following association analysis

Quantile-quantile (QQ) plots were drawn in R vs.3.2.2 (Development Core Team, 2008) using the package "qqman" to visualise the distribution of the test-statistic for each of the phenotypes. These plots are shown in Appendix D and showed no evidence of population stratification in the combined data sets.

Results from the association analyses described above were visualised using both Manhattan as well as LocusZoom plots. Manhattan plots allow the visualization of PLINK association results across the genome whilst LocusZoom is a tool used to plot regional association results. Manhattan plots were drawn using Haploview vs. 4.2 (Barrett et al., 2005) while LocusZoom plots were drawn using LocusZoom vs.1.1 (Pruim et al., 2010). Manhattan plots can be described as representing GWAS on a genomic scale where P values ($-\log_{10}$) are represented on the Y-axis and chromosomal position is represented on the X-axis of the plot (Ehret, 2010). SNPs with significant P values appear higher up on the plot thus resembling a Manhattan skyline. LocusZoom plots have the additional advantage of being able to display the association signal together with recombination information, LD and the closest genes in the region (Pruim et al., 2010). LocusZoom plots also display P values ($-\log_{10}$) as represented on the vertical axis and chromosomal position on the horizontal axis. All LocusZoom plots were drawn using hg18/1000 Genome June 2010 builds for LD background.

3. Results

3.1 Results from QC

3.1.1 Population Structure and PCA Analysis

PCA was used to identify outliers during the sample quality control process. Figures 3.2 and 3.3 show PC plots following the removal of outliers (details shown in Fig 3.4 and 3.5). The Bt20 samples represent a fairly homogenous group. PC plots (based on PC1 and PC2) were drawn using 13500 and 12100 SNPs for caregivers and young adults, respectively. Both groups (red triangles) form a close cluster and show strong overlap with previously studied Bt20 participants (black Sowetans-BSO) (green squares) and southeastern Bantu-speakers (blue circles). Both groups cluster distinctly from the African 1000G samples but seem to share more ancestry with the Luhya (blue triangles) and Yoruba (purple triangles). The Herero (SWB) from Botswana and Namibia (yellow squares) also share some ancestry with Bt20, illustrated by the close clustering in Figures 3.2 and 3.3.

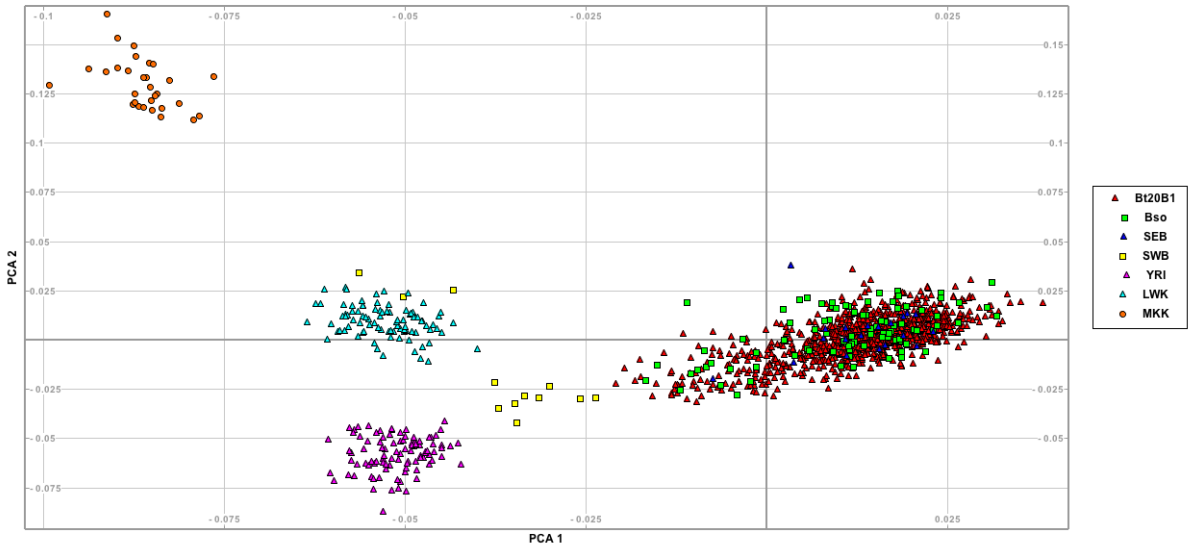


Figure 3.2 Principal component analysis plot comparing Bt20 (caregivers) genetic variation to various African populations following quality control, using PC 1 and 2.
PC 1 captures 60% whilst PC 2 captures 22% of the variation.

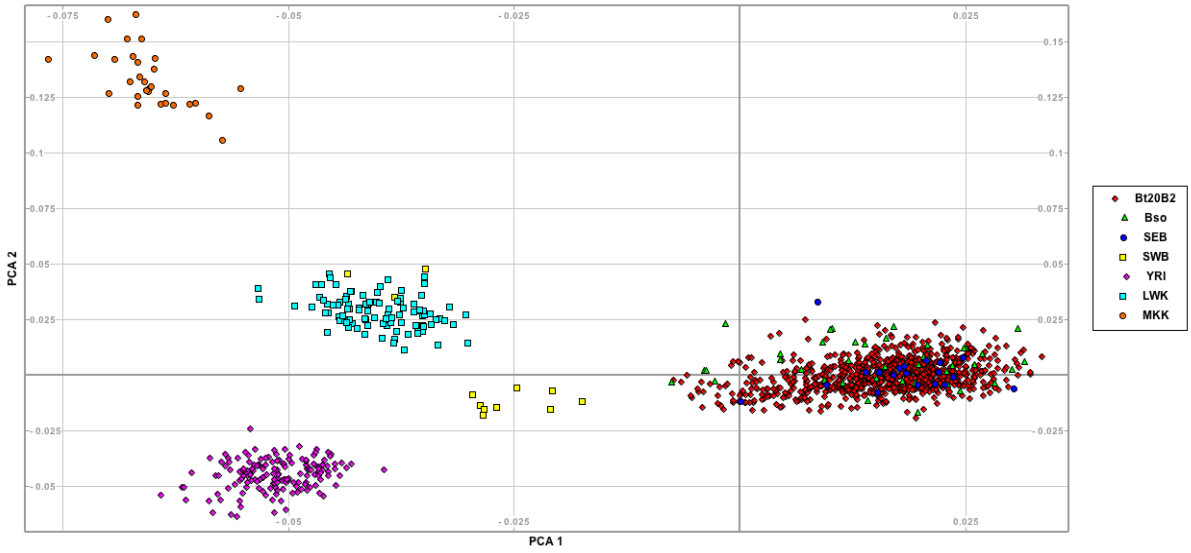


Figure 3.3 Principal component analysis plot comparing Bt20 (young adults) genetic variation to various African populations following quality control.
PC 1 captures 60,5% whilst PC 2 captures, 22,5% of the variation.

3.1.2 SNP and Sample QC

Following both SNP and sample QC measures - focusing on data missingness, Hardy-Weinberg equilibrium, relatedness, population stratification and phenotype QC, the final dataset comprised 972 samples containing 140649 SNPs and 954 individuals containing 127764 SNPs in caregivers and young adults, respectively.

Figure 3.4 illustrates that the greatest number of SNPs was lost in both datasets due to monomorphic SNPs and SNPs having a MAF < 1%. In the young adults most SNPs were removed due to SNP missingness and failing HWE criteria as shown in the figure. In terms of sample QC, Fig. 3.5 illustrates that more samples were removed in young adults due to poor genotyping and sample missingness than the female caregivers. There was also a higher degree of cryptic relatedness in the young adult as illustrated by the amount of samples removed for IBD, with both data sets having the same number of samples that were removed due to inconsistencies with sex between genotype and phenotype.

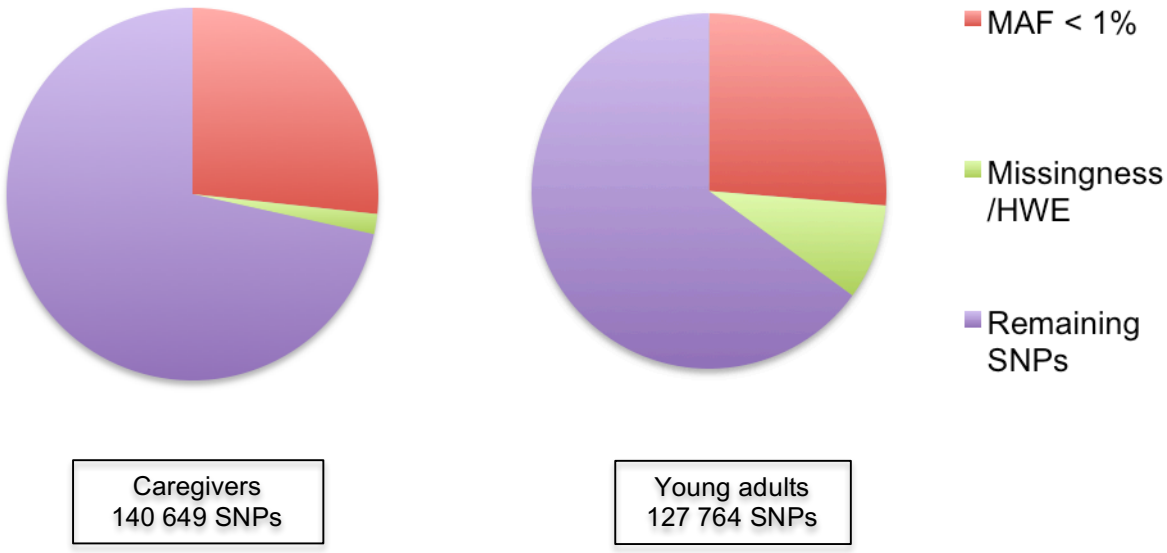


Figure 3.4 Results of SNP QC, showing the number of SNPs removed for each QC step. 52 593 SNPs and 51 819 SNPs were removed for MAF <1%, 6644 SNPs and 22 286 SNPs for combined missingness and HWE in the caregivers and young adults, respectively

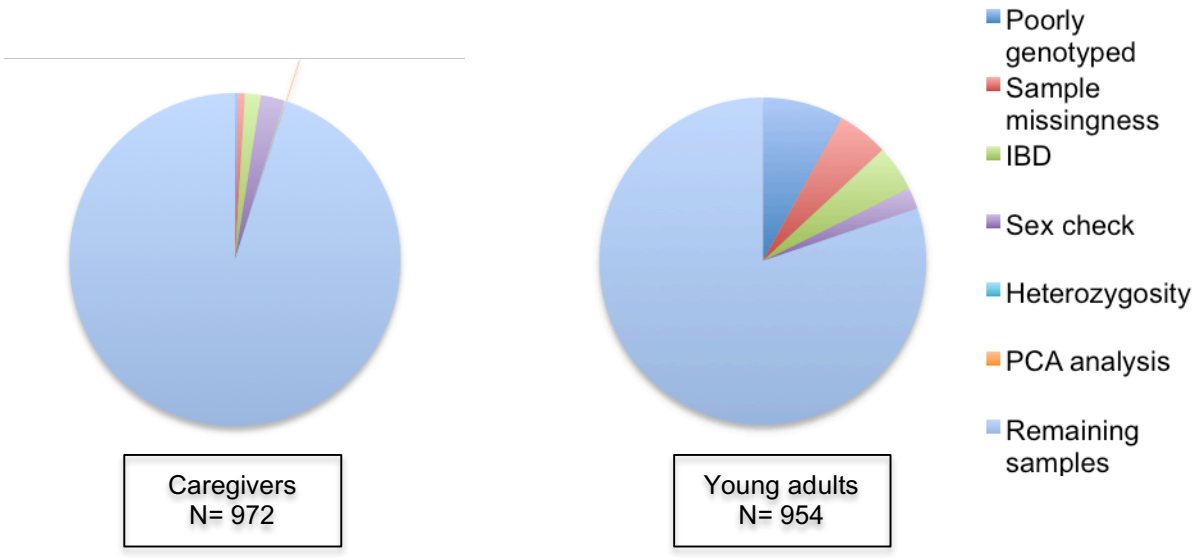


Figure 3.5 Results of sample QC showing proportion of samples removed for each QC step. The most samples, N=24 and N=100 were removed due to discordant sex information and poor genotyping for caregivers and young adults, respectively

3.2 Summary Statistics

The young adults were a mixed sex sample with N= 505 males, and N= 449 females. The group had a mean age of 17.9 years. The caregiver group was an all female group (N= 972) with a mean age of 41.8 years. The summary statistics are presented on the dataset following QC (Table 3.2), which includes data from 954 participants in the young adult group and 972 individuals in the caregiver group.

Table 3.2 Description of the Birth to Twenty cohort

The summary statistics for the young adults are further subdivided into males and females

Subjects	**Caregivers**		**Young Adults**					
			All		Females		Males	
N	972		954		449		505	
Females (%)	100		47		100		0	
	Mean* (SD)	*IQR*	**Mean* (SD)**	*IQR*	**Mean* (SD)**	*IQR*	**Mean* (SD)**	*IQR*
Age (years)	42.00 (8.8)		17.88 (0.38)		17.88 (0.36)		17.88 (0.39)	
Body mass index ($kg.m^{-2}$)	*30.30*	*25.70-34.40*	*21.70*	*19.00-23.30*	*22.30*	*20.00-25.55*	*19.80*	*18.50-21.50*
Height (m)	1.58 (0.06)		1.65 (0.08)		1.59 (0.06)		1.71 (0.07)	
Weight (kg)	76.03 (17.29)		59.20 (11.53)		59.20 (13.27)		59.21 (9.72)	
Hip circumference (cm)	111.10 (18.562)		94.10 (14.01)		98.65 (15.64)		90.19 (10.08)	
Waist circumference (cm)	86.57 (16.608)		73.04 (11.94)		74.79 (14.41)		71.49 (8.91)	
Waist-to-hip ratio (WHR)	0.77 (0.12)		0.77 (0.11)		0.75 (0.12)		0.79 (0.08)	
Fat mass (grams)	28164.98 (12043.30)		10522.46 (9200.22)		15774.79 (10160.69)		5843.29 (4643.70)	
Lean mass (grams)	36565.36 (9978.57)		32379.29 (16379.77)		28029.07 (13421.17)		36284.72 (17742.41)	
Percentage fat mass (%)	39.34 (11.10)		18.86 (14.01)		28.12 (14.23)		10.62 (6.79)	

*All values are presented as means except BMI values, which are presented as medians together with interquartile range (IQR) because BMI was not normally distributed and was log transformed to normality

Medians +IQR are in italics

3.3 Summary of signals following association analysis

Both basic association (unadjusted) and linear regression (adjusted for covariates) analysis were carried out for all phenotypes in the various analysis strategies outlined earlier. MLMA analysis was used when datasets were combined (combined all and combined females) to account for relatedness. All of the results from association testing for the various analysis strategies can be found in Appendix E. In this section association results of the top ten signals were presented when:

- A previously associated GWAS- locus was replicated, or
- When the signal at a particular locus was supported by more than one SNP or,
- When a signal was observed across phenotypes, termed cross-phenotype associations according to (Lu et al., 2016) and
- When a signal was at Metabochip genome-wide significance ($P \leq 5 \times 10^{-7}$)

The signals presented in the tables do not always reflect the most significant *P-value* for a particular trait in that analysis but also satisfied one of the conditions set out above and they are adjusted for covariates (summarised beneath each table). As mentioned previously *P*-values that were trending towards significance as well as those that were genome-wide significant (in bold), were considered. All cross-phenotype associations are colour-coded in the tables by gene, whilst those loci only associated with one phenotype are not highlighted. The signals associated with one specific phenotype are reported in different sections (3.3.1-3.3.7) All unadjusted association (basic association tests) results can be found in Appendix E. The association results presented in Tables 3.3-3.10 were supported with Manhattan and LocusZoom plots where appropriate.

3.3.1 Body mass index

Variants in or near *APOH, CNTNAP5*, *ZFYVE9*, *TRPM7* correlated with BMI as reported in Table 3.3. These loci are discussed further in section 3.6 under cross-phenotype associations, as they were associated with other phenotypes.

Table 3.3 SNP associations with body mass index

	Gene symbol	Gene location	SNP ID	Chr	BP	A1*	A2	MAF	Beta	SE	*P*-value
											P_{adj}****
Combined all	*APOH | PRKCA*	intergenic	rs115012414	17	61724373	C	T	0,04	-0,03	0,01	7,53 x10^{-6}
	APOH | PRKCA	intergenic	rs77612309	17	61724990	C	T	0,04	-0,03	0,01	1,17 x10^{-5}
	SP110	intron	rs2114591	2	230758813	T	C	0,40	0,01	0,00	6,62 x10^{-5}
											P_{adj}***
Combined females	*APOH | PRKCA*	intergenic	rs115012414	17	61724373	C	T	0,03	-0,04	0,01	1,68 x10^{-5}
	APOH | PRKCA	intergenic	rs77612309	17	61724990	C	T	0,03	-0,04	0,01	2,85 x10^{-5}
	SP110	intron	rs2114591	2	230758813	T	C	0,4	0,01	0	9,06 x 10^{-5}
											P_{adj}*
Older females (caregivers)	*LOC728241 | CNTNAP5*	intergenic	rs6541885	2	123668890	A	G	0,04	-0,05	0,01	5,39 x10^{-5}
	LOC728241 | CNTNAP5	intergenic	rs4411698	2	123651577	C	T	0,04	-0,04	0,01	5,89 x10^{-5}
											P_{adj}**
Young adults	*ZFYVE9*	intron	rs2753399	1	52546945	A	G	0,07	0,07	0,01	1,66 x10^{-5}
											P_{adj}*
Young males	*TRPM7*	intron	rs17598819	15	48701919	T	C	0,01	0,12	0,03	5,17 x10^{-6}
	USP50 | TRPM7	intergenic	rs17598264	15	48632003	G	A	0,01	0,12	0,03	5,27 x10^{-6}
	TRPM7	utr	rs62021060	15	48640931	C	T	0,01	0,12	0,03	5,27 x10^{-6}
	TRPM7	intron	rs62017164	15	48655616	T	C	0,01	0,12	0,03	5,27 x10^{-6}
	TRPM7	intron	rs62017165	15	48655833	T	G	0,01	0,12	0,03	5,27 x10^{-6}
	TRPM7	intron	rs1060599	15	48661924	T	C	0,01	0,12	0,03	5,27 x10^{-6}
	TRPM7	intron	rs62017202	15	48736535	A	G	0,01	0,12	0,03	5,27 x10^{-6}

All genetic loci that are correlated with more than one phenotype are colour-coded in the tables, whilst those loci only associated with one phenotype are not highlighted. BMI values were logged transformed to normality for all the analysis. Chr- chromosome, BP[1]- Note that all BP positions are reported using NCBI Build 36 (hg18), utr-untranslated region, A1[2] (effect allele) is the minor allele in this study, A2-major allele, MAF-minor allele frequency, Beta- refers to per allele change in the phenotype in kg.m^{-2} where a positive beta value shows that the minor allele is associated with an increase in the phenotype and a negative value is associated with a decrease, SE- standard error, P_{adj-} *P*-value adjusted for various covariates. The combined all and combined females were further adjusted for 10 PC's upon merging to rule out any confounding from substructure amongst related individuals. Padj**** adjusted for relatedness, sex, age and 10 principal components, Padj***adjusted for relatedness, age, and 10 principal components, Padj**adjusted for sex and age, Padj*adjusted for age.

3.3.2 Waist circumference

SNPs in or near *SP110, NRXN3*, *LPAL2*, *ZFYVE9* and *TRPM7* were found to be associated with WC. Only a single variant rs10146149 near *NRXN3 (Neurexin 3)* was associated with WC in both the combined dataset and the combined females following adjustment for covariates outlined below Table 3.4. All of the variants associated with WC were suggestive of association. *SP110, LPAL2*, *ZFYVE9* and *TRPM7* were associated with other phenotypes besides WC (refer to section 3.4).

Table 3.4 SNP associations with waist circumference

Subject groups	Gene symbol	Gene location	SNP ID	Chr	BP	A1[1]	A2	MAF	Beta	SE	*P*-value
											P_{adj}****
Combined all	*SP110*	intron	rs2114591	2	230758813	T	C	0.40	1.62	0.37	1.10×10^{-5}
	NRXN3 | LOC100131580	intergenic	rs10146149	14	78578318	T	C	0.11	-2.40	0.59	4.49×10^{-5}
											P_{adj}***
Combined females	*SP110*	intron	rs2114591	2	230758813	T	C	0.40	2.13	0.47	6.40×10^{-6}
	NRXN3 | LOC100131580	intergenic	rs10146149	14	78578318	T	C	0.11	-3.08	0.75	4.13×10^{-5}
											P_{adj}*
Older female (caregivers)	*LPAL2*	intron	rs9364558	6	160849934	G	C	0.27	27.93	6.271	9.46×10^{-6}
	SLC22A3 | LPAL2	intergenic	rs115553347	6	160806782	G	C	0.04	56.29	13.89	5.48×10^{-5}
											P_{adj}**
Young adults	*ZFYVE9*	intron	rs2753399	1	52546945	A	G	0.01	8.93	2.05	1.41×10^{-5}
											P_{adj}*
Young males	*TRPM7*	utr	rs62021060	15	48640931	C	T	0.01	16.41	5.50	**6.19×10^{-8}**
	TRPM7	intron	rs62017164	15	48655616	T	C	0.01	16.41	5.50	**6.19×10^{-8}**
	TRPM7	intron	rs62017165	15	48655833	T	G	0.01	16.41	5.50	**6.19×10^{-8}**
	TRPM7	intron	rs1060599	15	48661924	T	C	0.01	16.41	5.50	**6.19×10^{-8}**
	TRPM7	intron	rs62017202	15	48736535	A	G	0.01	16.41	5.50	**6.19×10^{-8}**
	TRPM7	intron	rs62017207	15	48745151	C	T	0.01	16.41	5.50	**6.19×10^{-8}**
	TRPM7	intron	rs62017208	15	48746646	A	G	0.01	16.41	5.50	**6.19×10^{-8}**
	TRPM7	intron	rs62017209	15	48747259	A	G	0.01	16.41	5.50	**6.19×10^{-8}**
	TRPM7	intron	rs17520350	15	48684965	C	T	0.01	16.41	5.49	**6.19×10^{-8}**

In Plink the A1 (effect allele) is the minor allele[1]. Beta refers to per effect allele change in WC in mm

Padj****adjusted for relatedness, sex, age, height and 10 principal components
Padj***adjusted for relatedness, age, height and 10 principal components
Padj**adjusted for sex, age and height
Padj*adjusted for age and height

3.3.3 Hip circumference

Variants in or near *PPP1R3B | LOC100129150*, *APOH | PRKCA*, *WARS2, BDNFOS, LPAL2* and *CNTNAP5* were associated with HC. Only *BDNFOS (brain-derived neurotrophic factor opposite strand)* was associated with HC alone. When combining females only (young female adults and female caregivers) following adjustment for covariates, three SNPs were suggestively associated with HC (rs58174260, rs12574325 and rs16917135, shown in Table 3.5). SNPs rs12574325 and rs16917135 (P_{unadj} =6.42 x 10^{-5}) were associated with HC in the female caregivers (Appendix E). Following adjustment for age and height the two SNPs still remained suggestively associated with HC in the female caregivers, shown in Table 3.5.

Table 3.5 SNP associations with hip circumference

Subject groups	Gene symbol	Gene location	SNP ID	Chr	BP	A1[1]	A2	MAF	Beta	SE	*P*-value
											P_{adj}****
Combined all	*PPP1R3B | LOC100129150*	intergenic	rs11778774	8	9251517	G	A	0.01	6.64	1.60	3.25 x10^{-5}
	APOH | PRKCA	intergenic	rs115012414	17	61724373	C	T	0.04	-4.18	1.01	3.34 x10^{-5}
	APOH | PRKCA	intergenic	rs77612309	17	61724990	C	T	0.04	-4.11	1.01	4.34 x10^{-5}
	PPP1R3B | LOC100129150	intergenic	rs78933755	8	9245811	G	A	0.01	6.75	1.67	3.25 x10^{-5}
	PPP1R3B | LOC100129150	intergenic	rs73535332	8	9255047	C	G	0.01	6.97	1.77	3.25 x10^{-5}
	WARS2	intron	rs56750694	1	119382364	T	G	0.05	3.75	0.96	3.25 x10^{-5}
											P_{adj}***
Combined females	*PPP1R3B | LOC100129150*	intergenic	rs78933755	8	9245811	G	A	0.01	8.84	2.10	2.65 x10^{-5}
	BDNFOS	intron	rs58174260	11	27547828	T	G	0.04	4.92	1.22	5.61 x10^{-5}
	BDNFOS	intron	rs12574325	11	27569624	A	G	0.06	4.28	1.08	7.10 x10^{-5}
	BDNFOS	intron	rs16917135	11	27571281	T	C	0.06	4.28	1.08	7.10 x10^{-5}
											P_{adj}*
Older Female (caregivers)	*LPAL2*	intron	rs9364558	6	160849934	G	C	0.27	27.93	6.27	9.46 x10^{-6}
	SLC22A3 | LPAL2	intergenic	rs115553347	6	160806782	G	C	0.04	56.29	13.89	5.48 x10^{-5}
	LOC728241 | CNTNAP5	intergenic	rs4411698	2	123651577	C	T	0.04	-62.23	14.58	2.17 x10^{-5}
	LOC728241 | CNTNAP5	intergenic	rs6541885	2	123668890	A	G	0.04	-66.78	15.92	2.99 x10^{-5}

Older Females	*LOC400796 | SEC16B*	intergenic	rs6425446	1	176031200	A	G	0.41	25.07	5.99	3.13×10^{-5}
	PPP1R3B | LOC100129150	intergenic	rs78933755	8	9245811	G	A	0.01	106.40	25.51	3.29×10^{-5}
	PPP1R3B | LOC100129150	intergenic	rs35584813	8	9245040	C	T	0.01	102.40	25.07	4.79×10^{-5}
	PPP1R3B | LOC100129150	intergenic	rs73535332	8	9255047	C	G	0.01	109.30	27.03	5.69×10^{-5}
	BDNFOS	intron	rs12574325	11	27569624	A	G	0.06	51.01	12.69	6.25×10^{-5}
	BDNFOS	intron	rs16917135	11	27571281	T	C	0.06	51.01	12.69	6.25×10^{-5}
Young adults	*ZFYVE9*	intron	rs2753399	1	52546945	A	G	0.01	9.20	2.05	$\mathbf{P_{adj}}^{**}$ 8.29×10^{-6}
Young males	*USP50 | TRPM7*	intergenic	rs17598264	15	48632003	G	A	0.01	18.04	3.12	$\mathbf{P_{adj}}^{*}$ $\mathbf{1.28 \times 10^{-8}}$
	TRPM7	utr	rs62021060	15	48640931	C	T	0.01	18.04	3.12	$\mathbf{1.28 \times 10^{-8}}$
	TRPM7	intron	rs62017164	15	48655616	T	C	0.01	18.04	3.12	$\mathbf{1.28 \times 10^{-8}}$
	TRPM7	intron	rs62017165	15	48655833	T	G	0.01	18.04	3.12	$\mathbf{1.28 \times 10^{-8}}$
	TRPM7	intron	rs1060599	15	48661924	T	C	0.01	18.04	3.12	$\mathbf{1.28 \times 10^{-8}}$
	TRPM7	intron	rs62017202	15	48736535	A	G	0.01	18.04	3.12	$\mathbf{1.28 \times 10^{-8}}$
	TRPM7	intron	rs62017207	15	48745151	C	T	0.01	18.04	3.12	$\mathbf{1.28 \times 10^{-8}}$
	TRPM7	intron	rs62017208	15	48746646	A	G	0.01	18.04	3.12	$\mathbf{1.28 \times 10^{-8}}$
	TRPM7	intron	rs62017209	15	48747259	A	G	0.01	18.04	3.12	$\mathbf{1.28 \times 10^{-8}}$
	TRPM7	intron	rs17598819	15	48701919	T	C	0.01	18.05	3.12	$\mathbf{1.28 \times 10^{-8}}$

The effect allele is the minor allele ($A1^{1}$). Beta refers to the per effect allele change in HC in mm

Padj****adjusted for relatedness, sex, age, height and 10 principal components

Padj***adjusted for relatedness, age, height and 10 principal components

Padj**adjusted for sex, age and height

Padj*adjusted for age and height

Table 3.6 SNP associations with waist-to-hip ratio

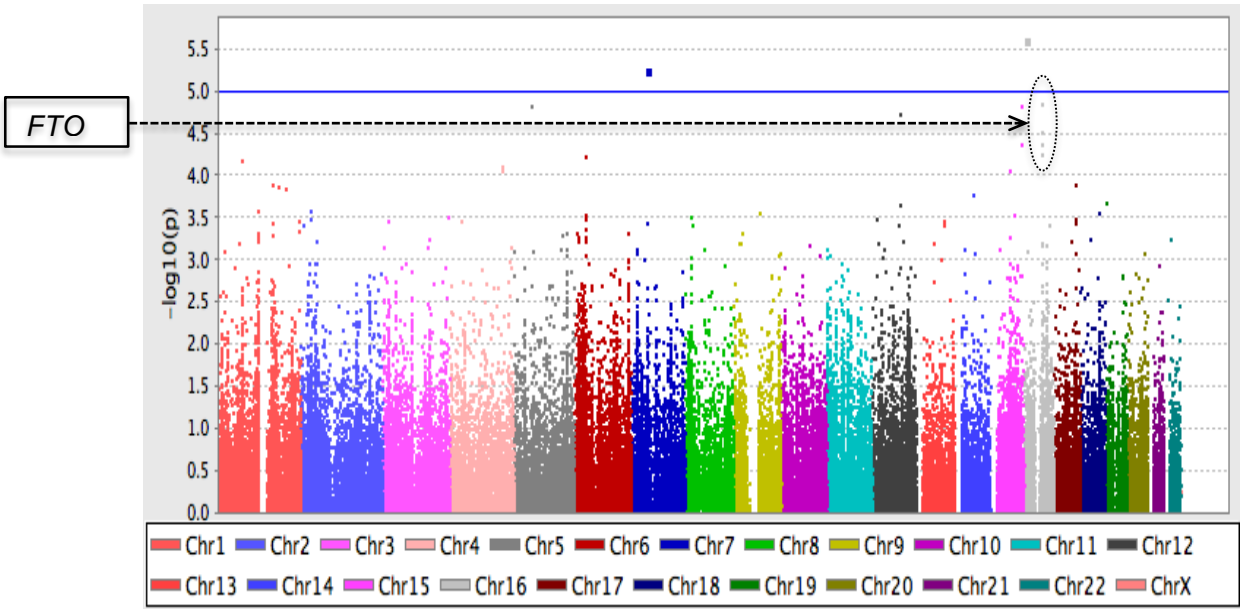
Subject groups	Gene symbol	Gene location	SNP ID	Chr	BP	A1*	A2	MAF	Beta	SE	*P*-value
											P_{adj}****
Combined all	*FTO*	intron	rs1861554	16	52607268	G	A	0.07	0.02	0.00	2.91×10^{-5}
	WARS2	utr	rs17023092	1	119375976	T	C	0.07	-0.02	0.00	9.11×10^{-5}
	WARS2	utr	rs74112264	1	119376488	A	G	0.07	-0.02	0.00	9.11×10^{-5}
	WARS2	utr	rs17023118	1	119381509	A	G	0.07	-0.02	0.00	9.11×10^{-5}
											P_{adj}***
Combined females	*FTO*	intron	rs1861554	16	52607268	G	A	0.07	0.02	0.01	2.75×10^{-5}
	FTO	intron	rs1861358	16	52602704	A	C	0.06	0.03	0.01	4.83×10^{-5}
	FTO	intron	rs2111116	16	52606753	A	G	0.06	0.03	0.01	4.83×10^{-5}
											P_{adj}**
Young adults	*WARS2*	intron	rs12095241	1	119431520	G	T	0.13	0.02	0.00	1.08×10^{-6}
	WARS2	intron	rs12088290	1	119385449	T	C	0.14	0.02	0.00	1.01×10^{-5}
											P_{adj}*
Young males	*LOC100129474 | SLC17A4*	intergenic	rs3923725	6	25842899	A	C	0.01	0.16	0.02	**6.31×10^{-13}**
	COBLL1	intron	rs115743734	2	165255601	A	G	0.01	0.09	0.01	**6.34×10^{-9}**

In Plink the A1 (effect allele) is the minor allele*. Beta refers to per effect allele change in WHR
Padj****adjusted for relatedness, sex, age and 10 principal components
Padj***adjusted for relatedness, age, and 10 principal components
Padj**adjusted for sex and age;
Padj*adjusted for age

3.3.4 Waist to hip ratio

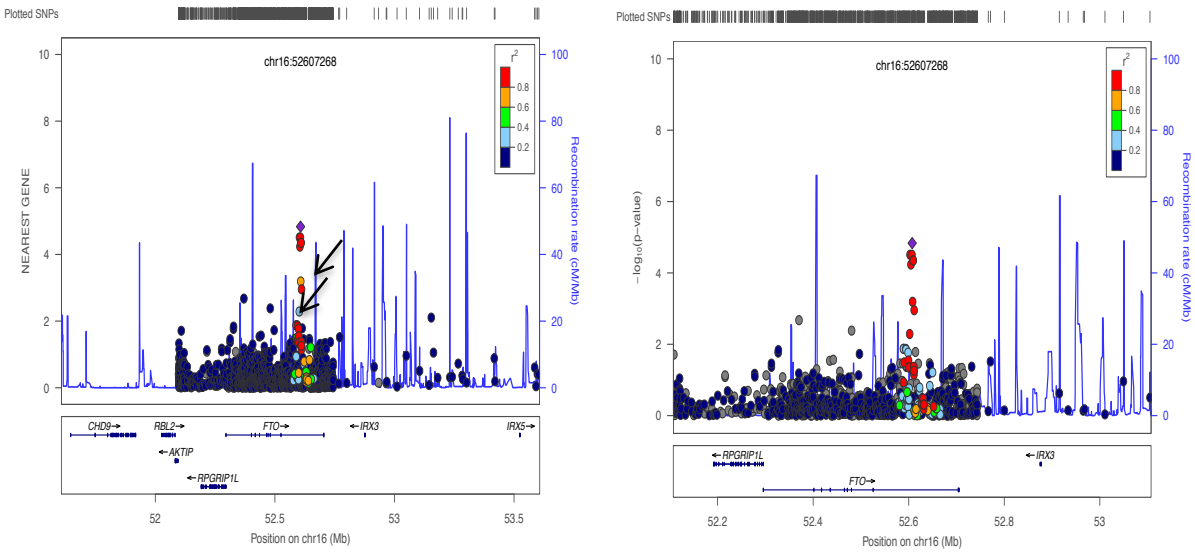
Variants in or near *FTO*, *WARS2, LOC100129474* | *SLC17A4* and *COBLL1* were associated with WHR. *WARS2* was also associated with HC, albeit with a different variant rs56750694. The suggestive association in *FTO (fat mass and obesity associated locus)* with rs1861554 was observed when combining the datasets, after adjustment for covariates. The association with *FTO* is strengthened by the observation that three intronic SNPs (reported in Table 3.6) in *FTO* are related to WHR when merging the 2 groups of females. This signal is driven by rs1861554 as illustrated in Fig. 3.6 (i). None of the signals observed at *FTO* were at GW significance, but rather trending toward significance following adjustment for covariates. Regional plots for the lead SNP are shown in Fig. 3.6 (ii-a) for the combined females against an African LD background and European (ii-b). There are highly correlated SNPs (three SNPs) with the index SNP against both the European and African LD structure, but there are some SNPs against the African LD background (indicated by black arrows) that have a lower correlation with the index SNP, but their *P*-values are not significant (10^{-3}).

Only variants in or near *LOC100129474* | *SLC17A4 (solute carrier family 17)* and *COBLL1* (*Cordon-Bleu WH2 repeat protein-like1*) were observed with WHR at genome-wide significance in the young males. The signal observed near *LOC100129474* | *SLC17A4* was the strongest, signified by the highest *P*-value observed in all the analyses, however together with *COBLL1* had a single SNP supporting the association.



(i)

a) YRI b) CEU



(ii)

Figure 3.6 Manhattan plot (i) and regional plot (ii) for variants near *FTO* and waist-to-hip ratio
(i) Showing the three variants suggestively associated with WHR in the combined females. The blue horizontal line indicates ($P< 5 \times 10^{-5}$) which is the cut-off used in this study indicating suggestive associations.
(ii) Regional plots for WHR and lead SNP rs1861554 in *FTO* in the combined female data set with LD structure shown against a YRI (a) and CEU (b) background. The recombination rates are indicated in blue. The three intronic SNPs are highly correlated as indicated by r^2 values in red. The black arrows represent variants that have lower *P*-values, and are not as highly correlated with the index SNP (purple) that might be important signals in narrowing down causal loci because they are indistinguishable against the European LD background. The closest genes are shown beneath the plots.

3.3.5 Fat mass

Several interesting variants in or near *SEC16B*, *CNTNAP5, SP110, ZFYVE9, TMEM18* and *NEGR1* were associated with fat mass. SNPs rs114285121, rs78501377, rs76275602, rs7840669 and rs78559588 in or near *TMEM18* (*transmembrane 18*) were only associated in the young female adults with fat mass following adjusted association analysis (Table 3.7). The variants were trending toward significance, shown in Fig. 3.7. All minor alleles effect positive changes to β values. Regional plots were not drawn for this locus, as LD information for European populations was not available. All of the other variants were associated with other phenotypes and were reported in section 3.4.

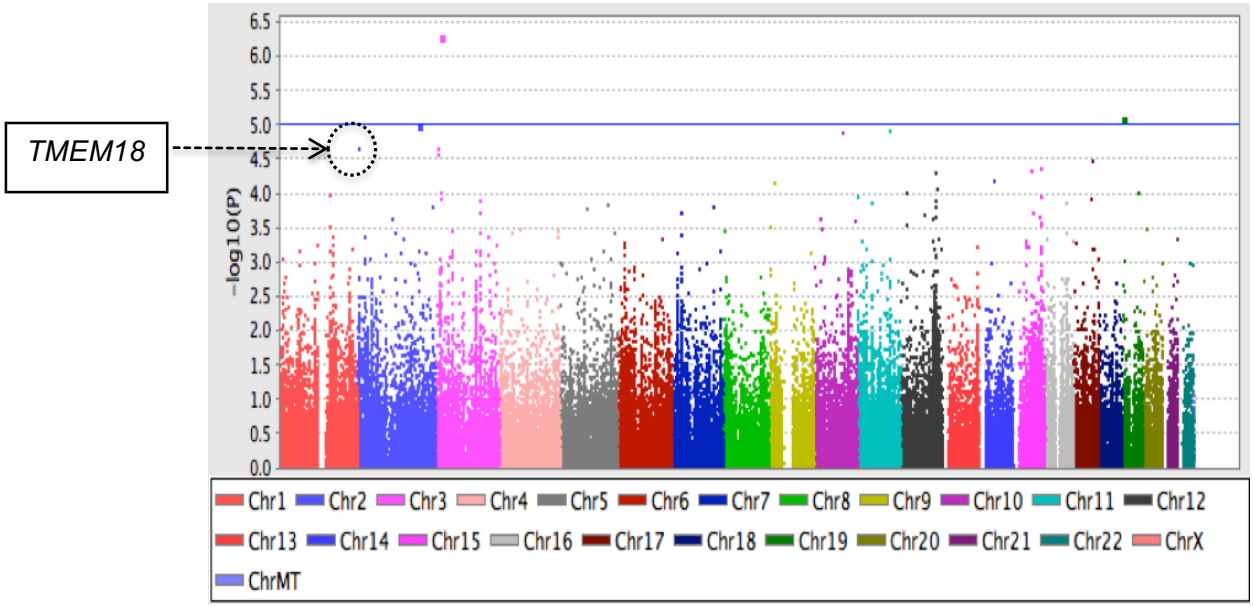


Figure 3.7 Manhattan plot for fat mass in the young adult females showing the association of SNPs in or near *TMEM18*.
The blue horizontal line is the cut-off for suggestive associations ($P \leq 10^{-5}$)

Table 3.7 SNP associations with fat mass

Subject groups	Gene symbol	Gene location	SNP ID	Chr	BP	A1[1]	A2	MAF	Beta	SE	*P*-value
											P_{adj}****
Combined all	*LOC400796 | SEC16B*	intergenic	rs6664268	1	176031329	C	T	0.22	-1.80	0.37	**9.48×10^{-7}**
	LOC400796 | SEC16B	intergenic	rs4075235	1	176034945	T	C	0.26	-1.65	0.35	1.84×10^{-6}
	LOC400796 | SEC16B	intergenic	rs4072161	1	176036310	G	T	0.48	1.45	0.31	3.18×10^{-6}
	LOC400796 | SEC16B	intergenic	rs6425446	1	176031200	A	G	0.42	1.36	0.31	1.01×10^{-5}
	LOC400796 | SEC16B	intergenic	rs11581129	1	176037754	A	G	0.17	-1.69	0.40	2.68×10^{-5}
	LOC400796 | SEC16B	intergenic	rs4311843	1	176035387	C	A	0.33	1.32	0.33	5.50×10^{-5}
											P_{adj}***
Combined females	*LOC400796 | SEC16B*	intergenic	rs6425446	1	176031200	A	G	0.41	1.81	0.40	5.55×10^{-6}
	LOC400796 | SEC16B	intergenic	rs4075235	1	176034945	T	C	0.26	-1.98	0.44	7.54×10^{-6}
	LOC400796 | SEC16B	intergenic	rs6664268	1	176031329	C	T	0.22	-2.06	0.46	9.63×10^{-6}
	LOC400796 | SEC16B	intergenic	rs4072161	1	176036310	G	T	0.47	1.71	0.40	2.27×10^{-5}
	LOC400796 | SEC16B	intergenic	rs4311843	1	176035387	C	A	0.32	1.73	0.42	4.47×10^{-5}
	LOC728241 | CNTNAP5	intergenic	rs6541885	2	123668890	A	G	0.03	-4.27	1.08	7.39×10^{-5}
	SP110	intron	rs2114591	2	230758813	T	C	0.40	1.53	0.39	8.27×10^{-5}
											P_{adj}*
Older Female (caregivers)	*LOC400796 | SEC16B*	intergenic	rs6425446	1	176031200	A	G	0.41	2.26	0.47	2.08×10^{-6}
	LOC728241 | CNTNAP5	intergenic	rs6541885	2	123668890	A	G	0.04	-5.94	1.29	4.48×10^{-6}
	LOC728241 | CNTNAP5	intergenic	rs4411698	2	123651577	C	T	0.04	-5.24	1.17	8.91×10^{-6}
	LOC400796 | SEC16B	intergenic	rs4072161	1	176036310	G	T	0.46	2.10	0.48	1.29×10^{-5}
	LOC400796 | SEC16B	intergenic	rs4348685	1	176034251	G	C	0.27	-2.26	0.53	2.43×10^{-5}
	LOC400796 | SEC16B	intergenic	rs4075235	1	176034945	T	C	0.27	-2.26	0.53	2.43×10^{-5}
	SP110	intron	rs2114591	2	230758813	T	C	0.39	1.99	0.47	2.83×10^{-5}
											P_{adj}**
Young adults	*ZFYVE9*	intron	rs2753399	1	52546945	A	G	0.01	7.33	1.56	3.26×10^{-6}
											P_{adj}*
Young females	*LOC727944 | TMEM18*	intergenic	rs114285212	2	630159	A	G	0.02	12.00	2.79	2.27×10^{-5}
	LOC727944 | TMEM18	intergenic	rs78501377	2	630682	C	T	0.02	12.00	2.79	2.27×10^{-5}

Young females	*LOC727944 | TMEM18*	intergenic	rs76275602	2	631877	A	G	0.02	12.00	2.79	2.27×10^{-5}
	LOC727944 | TMEM18	intergenic	rs78460669	2	633005	C	G	0.02	12.00	2.79	2.27×10^{-5}
	LOC727944 | TMEM18	intergenic	rs78559588	2	637656	C	T	0.02	12.00	2.79	2.27×10^{-5}
	LOC727944 | TMEM18	intergenic	rs78348389	2	638436	G	C	0.02	12.00	2.79	2.27×10^{-5}
	LOC727944 | TMEM18	intergenic	rs116093073	2	638520	A	G	0.02	12.00	2.79	2.27×10^{-5}
	LOC727944 | TMEM18	intergenic	rs114461922	2	642236	A	G	0.02	12.00	2.79	2.27×10^{-5}
											P_{adj}*
Young males	*LOC400796 | SEC16B*	intergenic	rs16852018	1	176098880	A	G	0.02	6.39	1.12	**2.41×10^{-8}**
	NEGR1 | LOC100132353	intergenic	rs72941254	1	72446505	A	C	0.05	3.28	0.67	1.64×10^{-6}
	NEGR1 | LOC100132353	intergenic	rs72941257	1	72451438	C	T	0.05	3.28	0.67	1.64×10^{-6}
	NEGR1 | LOC100132353	intergenic	rs72941270	1	72465674	A	T	0.05	3.28	0.67	1.64×10^{-6}
	NEGR1 | LOC100132353	intergenic	rs72941224	1	72411598	T	A	0.05	3.27	0.68	1.90×10^{-6}

In Plink the A1 (affect allele) is the minor allele[1]. Beta refers to per effect allele change in fat mass in kilograms (kgs)
Padj****adjusted for relatedness, sex, age, height and 10 principal components
Padj***adjusted for relatedness, age, height and 10 principal components
Padj**adjusted for sex, age and height
Padj*adjusted for age and height

3.3.6 Lean mass

Variants in or near *SP110* and *PPP1R3B* were associated with lean mass (Table 3.8). These variants are also associated with other phenotypes. The variants are all suggestive or trending towards suggestive significance.

Table 3.8 SNP associations with lean mass

Subject groups	Gene symbol	Gene location	SNP ID	Chr	BP	A1[1]	A2	MAF	Beta	SE	*P*-value
											P_{adj}****
Combined all	*SP110*	intron	rs2114591	2	230758813	T	C	0.40	0.77	0.18	9.89 x10^{-6}
											P_{adj}***
Combined females	*SP110*	intron	rs2114591	2	230758813	T	C	0.40	0.83	0.21	5.27 x10^{-5}
	PPP1R3B | LOC100129150	intergenic	rs73535324	8	9252643	A	C	0.01	3.79	0.92	9.89 x10^{-5}
	PPP1R3B | LOC100129150	intergenic	rs73535332	8	9255047	C	G	0.01	3.81	0.93	4.61 x10^{-5}
	PPP1R3B | LOC100129150	intergenic	rs78933755	8	9245811	G	A	0.01	3.64	0.90	5.29 x10^{-5}
	PPP1R3B | LOC100129150	intergenic	rs11778774	8	9251517	G	A	0.01	3.45	0.86	6.56 x10^{-5}
	PPP1R3B | LOC100129150	intergenic	rs35584813	8	9245040	C	T	0.01	3.46	0.88	8.06 x10^{-5}
	PPP1R3B | LOC100129150	intergenic	rs36090863	8	9246378	T	C	0.01	3.46	0.88	8.06 x10^{-5}
Older females (caregivers)											P_{adj}*
	PPP1R3B | LOC100129150	intergenic	rs73535324	8	9252643	A	C	0.01	4.70	1.10	2.15 x10^{-5}
	PPP1R3B | LOC100129150	intergenic	rs73535332	8	9255047	C	G	0.01	4.76	1.12	2.53 x10^{-5}
	PPP1R3B | LOC100129150	intergenic	rs11778774	8	9251517	G	A	0.01	4.31	1.04	3.77 x10^{-5}
	PPP1R3B | LOC100129150	intergenic	rs35584813	8	9245040	C	T	0.01	4.34	1.06	4.53 x10^{-5}
	PPP1R3B | LOC100129150	intergenic	rs36090863	8	9246378	T	C	0.01	4.34	1.06	4.53 x10^{-5}
	PPP1R3B | LOC100129150	intergenic	rs78933755	8	9245811	G	A	0.01	4.38	1.08	5.52 x10^{-5}

In Plink the A1 (affect allele) is the minor allele*. Beta refers to per effect allele change in lean mass in kg's.
Padj****adjusted for relatedness, sex, age, height and 10 principal components
Padj***adjusted for relatedness, age, height and 10 principal components
Padj*adjusted for age and height

3.3.7 Percentage fat mass

Variants in or near *SEC16B*, *APOH | PRKCA, CNTNAP5, NBEAL1 and NEGR1* were associated with PFM. Two SNPs rs7576822 and rs9678194 (Table 3.9 and Fig. 3.8) in and *NBEAL1 (neurobeachin like 1)* were seen to be associated with percentage fat mass in the young adults only, following the adjustment for covariates whilst variants in or near *SEC16B*, *APOH | PRKCA, CNTNAP5,* and *NEGR1* were associated with other

phenotypes as well. The association with *NBEAL1* was only observed in the young adults group.

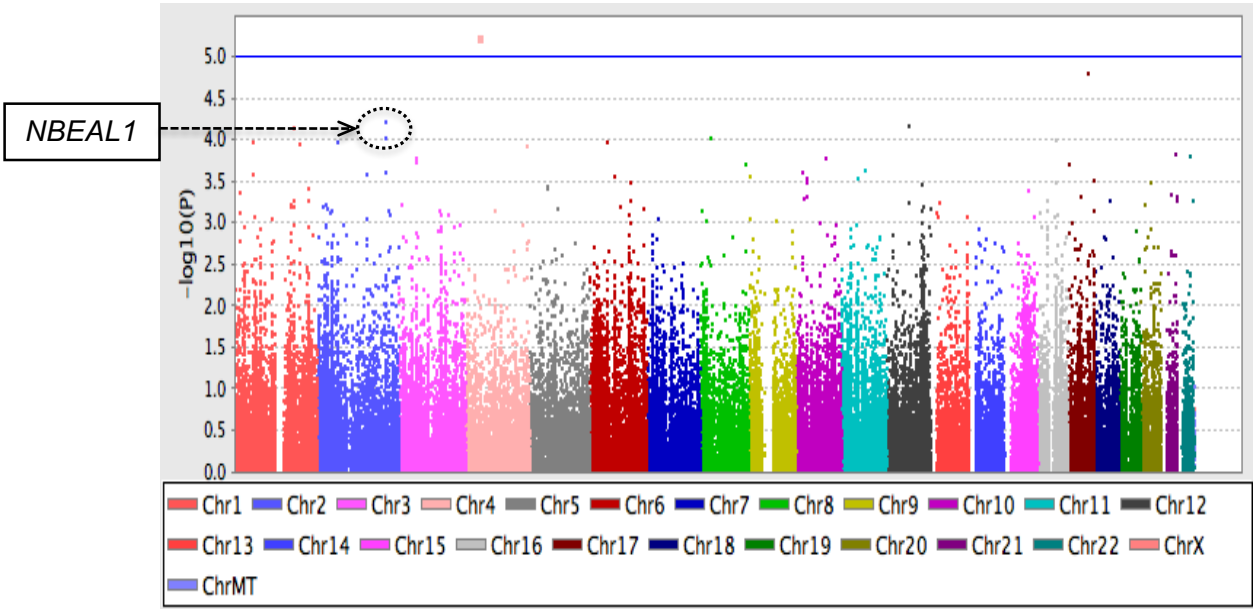


Figure 3.8 Manhattan plot for PFM in the young adult dataset where SNPs correlated with *NBEAL1* are trending towards significance.

Table 3.9 SNP associations with percentage fat mass

	Gene symbol	Gene location	SNP ID	Chr	BP	A1*	A2	MAF	Beta	SE	*P*-value
											P_{adj}****
Combined all	*LOC400796 | SEC16B*	intergenic	rs6664268	1	176031329	C	T	0,22	-1,35	0,27	**7,90 x10^{-7}**
	LOC400796 | SEC16B	intergenic	rs4075235	1	176034945	T	C	0,26	-1,27	0,26	**9,05 x10^{-7}**
	LOC400796 | SEC16B	intergenic	rs4072161	1	176036310	G	T	0,48	1,10	0,23	2,50 x10^{-6}
	LOC400796 | SEC16B	intergenic	rs2068973	1	176059338	A	G	0,43	1,01	0,23	1,35 x10^{-5}
	LOC400796 | SEC16B	intergenic	rs11581129	1	176037754	A	G	0,17	-1,28	0,30	2,00 x10^{-5}
	LOC400796 | SEC16B	intergenic	rs6425453	1	176085789	G	A	0,48	0,96	0,23	1,35 x10^{-5}
	LOC400796 | SEC16B	intergenic	rs6425446	1	176031200	A	G	0,42	0,95	0,23	1,35 x10^{-5}
	APOH | PRKCA	intergenic	rs115012414	17	61724373	C	T	0,04	-2,40	0,59	1,35 x10^{-5}
											P_{adj}***
Combined females	*LOC400796 | SEC16B*	intergenic	rs4075235	1	176034945	T	C	0,26	-1,38	0,31	8,89 x10^{-6}
	LOC400796 | SEC16B	intergenic	rs6425446	1	176031200	A	G	0,41	1,24	0,28	9,53 x10^{-6}
	LOC400796 | SEC16B	intergenic	rs2068973	1	176059338	A	G	0,42	1,21	0,28	1,49 x10^{-5}
	LOC400796 | SEC16B	intergenic	rs6664268	1	176031329	C	T	0,22	-1,38	0,33	2,31 x10^{-5}
	LOC400796 | SEC16B	intergenic	rs10913437	1	176062763	T	C	0,43	1,18	0,28	2,77 x10^{-5}
	LOC400796 | SEC16B	intergenic	rs4072161	1	176036310	G	T	0,47	1,18	0,28	3,23 x10^{-5}
	LOC400796 | SEC16B	intergenic	rs6425453	1	176085789	G	A	0,47	1,14	0,28	4,44 x10^{-5}
											P_{adj}*
Older females (caregivers)	*LOC728241 | CNTNAP5*	intergenic	rs4411698	2	123651577	C	T	0,04	-3,78	0,77	8,89 x10^{-6}
	LOC728241 | CNTNAP5	intergenic	rs6541885	2	123668890	A	G	0,04	-4,03	0,85	2,20 x10^{-6}
	LOC400796 | SEC16B	intergenic	rs2068973	1	176059338	A	G	0,43	1,44	0,31	4,17 x10^{-6}
	LOC400796 | SEC16B	intergenic	rs6425446	1	176031200	A	G	0,41	1,42	0,31	5,86 x10^{-6}
	LOC400796 | SEC16B	intergenic	rs4072161	1	176036310	G	T	0,46	1,43	0,31	6,32 x10^{-6}
	LOC400796 | SEC16B	intergenic	rs6425453	1	176085789	G	A	0,47	1,40	0,31	7,09 x10^{-6}
	LOC400796 | SEC16B	intergenic	rs10913437	1	176062763	T	C	0,43	1,38	0,31	1,07 x10^{-5}
	LOC400796 | SEC16B	intergenic	rs1854288	1	176069710	A	G	0,47	1,36	0,31	1,21 x10^{-5}
	LOC400796 | SEC16B	intergenic	rs12092449	1	176081551	C	T	0,47	1,35	0,31	1,48 x10^{-5}
											P_{adj}**

Young adults	*ZFYVE9*	intron	rs2753399	1	52546945	A	G	0,01	10,03	2,25	1,09 x10^{-5}
	NBEAL1	intron	rs7576822	2	203782462	T	C	0,02	4,52	1,12	6,13 x10^{-5}
	NBEAL1	intron	rs9678194	2	203760073	G	C	0,02	4,52	1,12	6,30 x10^{-5}
	LOC400796 | SEC16B	intergenic	rs16852018	1	176098880	A	G	0,01	5,45	1,37	7,40 x10^{-5}
											P_{adj}*
Young males	*LOC400796 | SEC16B*	intergenic	rs16852018	1	176098880	A	G	0,02	7,00	1,40	**8,51 x10^{-7}**
	NEGR1 | LOC100132353	intergenic	rs72941254	1	72446505	A	C	0,05	3,90	0,84	4,35 x10^{-6}
	NEGR1 | LOC100132353	intergenic	rs72941257	1	72451438	C	T	0,05	3,90	0,84	4,35 x10^{-6}
	NEGR1 | LOC100132353	intergenic	rs72941270	1	72465674	A	T	0,05	3,90	0,84	4,35 x10^{-6}
	NEGR1 | LOC100132353	intergenic	rs72941224	1	72411598	T	A	0,05	3,89	0,84	5,00 x10^{-6}

In Plink the A1 (affect allele) is the minor allele[1]. Beta refers to per effect allele change in PFM.

Padj****adjusted for relatedness, sex, age, height and 10 principal components

Padj***adjusted for relatedness, age, height and 10 principal components

Padj**adjusted for sex, age and height

Padj*adjusted for age and height

3.4 Sub-group and cross phenotype associations

Nine loci were found to be either suggestively or significantly associated with more than one trait. These loci are summarised in Table 3.10. The cross phenotype associations are discussed according to the genetic loci in common. The observed sub-group and cross phenotype associations are also presented in the Appendix E (Table E.11).

Table 3.10 Summary of cross-phenotype associations

Loci	Phenotypes						
	BMI	WC	HC	WHR	FM	LM	PFM
LOC400796 | SEC16B			x		x		x
*NEGR1 | LOC100132353**					x		x
TRPM7	x	x	x				
WARS2			x	x			
LOC728241 | CNTNAP5	x		x		x		x
SP110		x			x	x	
LPAL2		x	x				
PPP1R3B | LOC100129150			x			x	
ZFYVE9	x	x	x		x		x
APOH | PRKCA	x		x				x

3.4.1 *SEC16 homolog B (SEC16B)*

When both datasets were combined (after adjusting for covariates), three GW-significant signals were observed in or near *SEC16B* with lead SNP rs6664268 being the same for both fat mass and PFM. The variants rs6664268 and rs4075237 were GW significant for PFM in the combined dataset (Table 3.9), while only rs6664268 was GW significant for fat mass (Table 3.7). The association with fat mass and PFM was mirrored in the merged female group (female caregivers and young females) following adjustment for covariates. The association was driven by lead SNP rs6425446 (fat mass) and rs4075235 (PFM) in or near *SEC16B* as reported in Table 3.7 and illustrated in Fig. 3.9 and Table 3.9 and Fig. 3.10, respectively. Several SNPs were observed with *SEC16B* together with fat mass (lead SNP-rs6425446), and PFM (rs2068973), reported in Table 3.7 and 3.8, in the female caregivers after adjusting for age and height. Only a single SNP, rs16852018 was associated with *SEC16B* and PFM in the young adult group, after adjustment for sex, age and height (Table 3.9). Rs16852018 was also found at GW significance after adjustment for covariates with both fat mass and PFM in the young male adults with no signals for *SEC16B* observed with

any of the phenotypes in the young adult females. A single signal in or near *SEC16B* was observed for rs6425446 and HC in the female caregivers following adjustment for both age and height (Table 3.5).

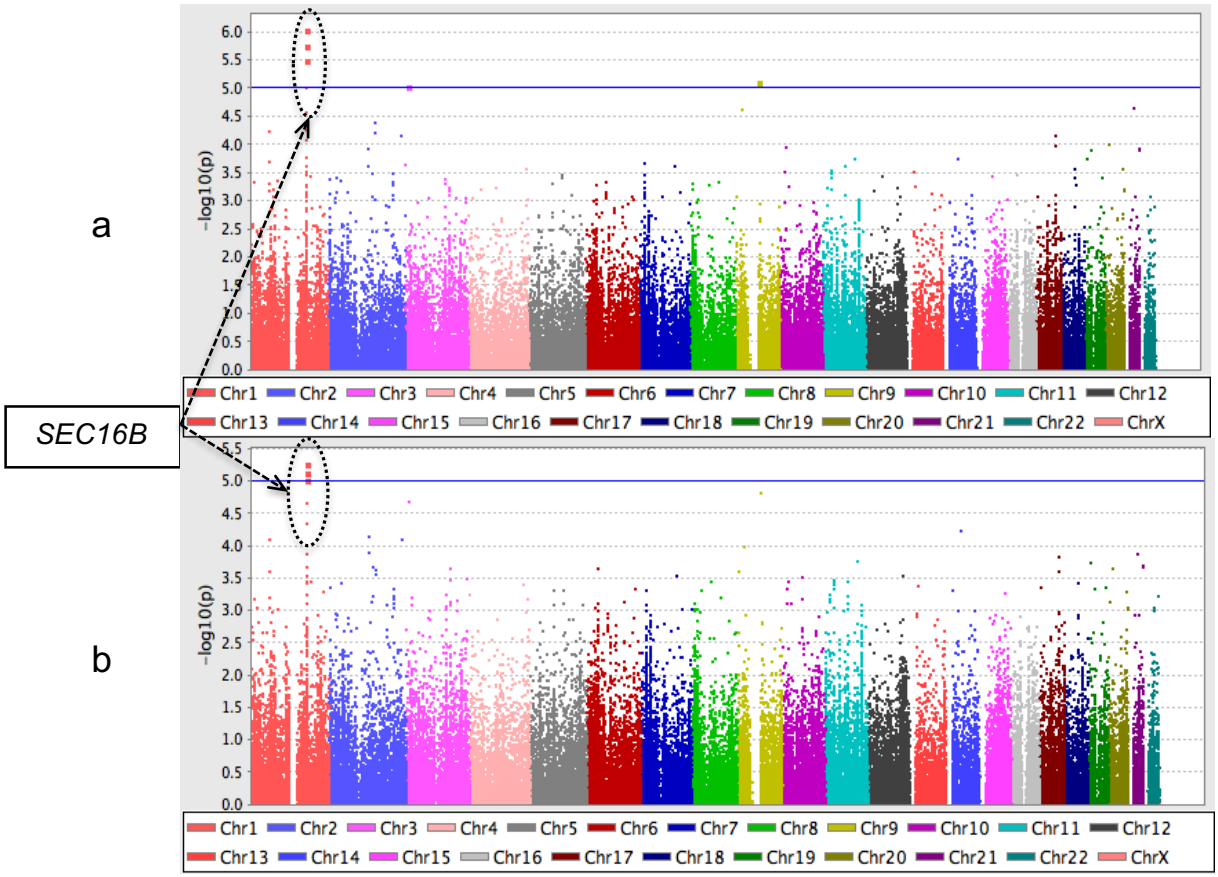


Figure 3.9 Manhattan plots for fat mass in the combined all (a) and combined female dataset (b), respectively

Illustrating the signals observed in or near *SEC16B* with the various sub-groups in the analysis, with a stronger signal for fat mass observed when all samples are combined (a). The horizontal lines in the Manhattan plots correspond to suggestive *P*-values ≤ 5.0 x 10^{-5} (blue)

β (beta) coefficients are reported for each SNP in the association results (Tables 3.4, 3.6 and 3.8) where the value reflects the per effect allele change in the phenotype. All the β (beta) values reported for *SEC16* associations effect a large change in the respective phenotypes. Some β values effect negative changes for the top signals, according to frequency of minor allele in African populations (by default PLINK uses the minor allele in association analysis), suggesting that the major allele is associated with increases in beta. These results showed that an increase in sample size (merging both datasets; or merging the females) increased the statistical power to detect the associations in or near *SEC16B*. Regional plots of lead

SNPs for fat mass and PFM in the combined datasets are illustrated and reported in Fig.s 3.11 - 3.13.

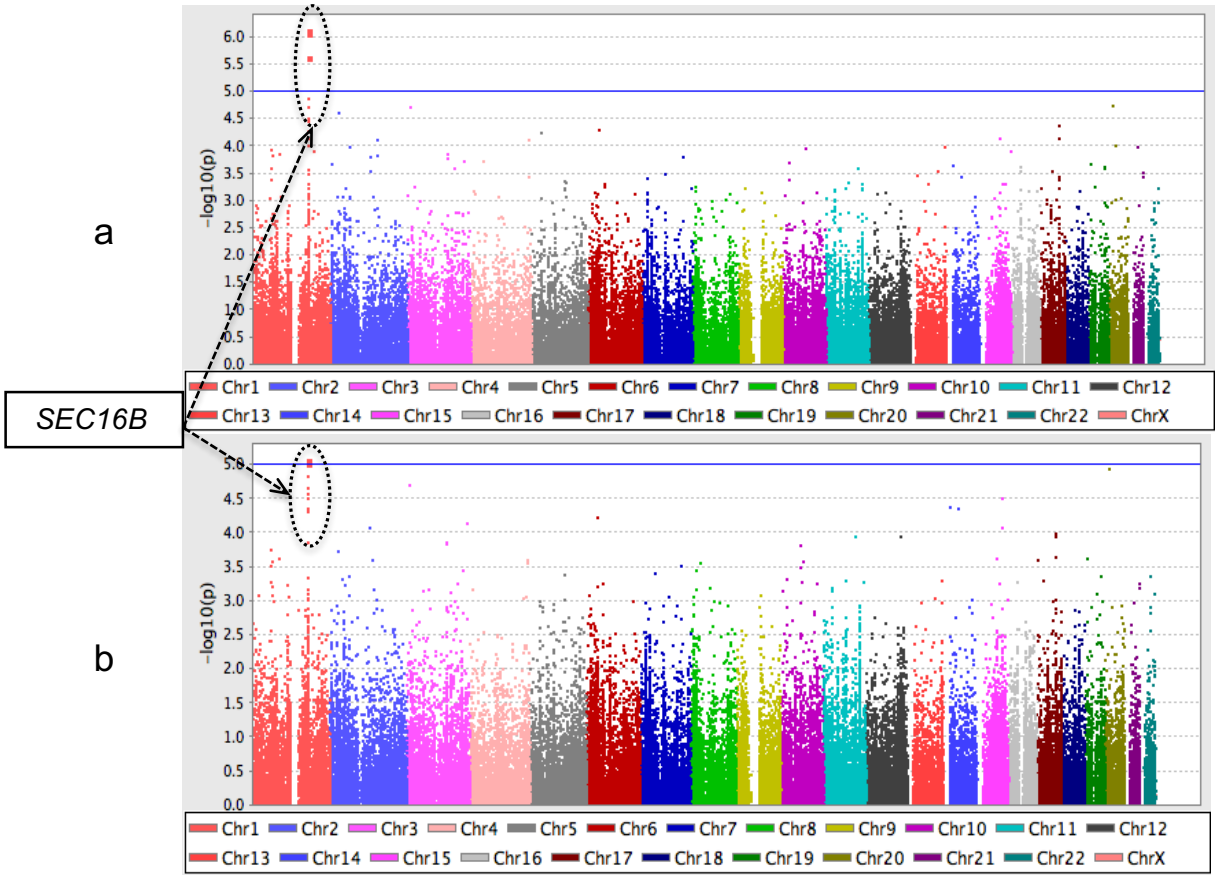


Figure 3.10 Manhattan plots for percentage fat mass in the combined all (a) and combined female dataset (b), respectively illustrating the strength of signals observed in the various sub-groups

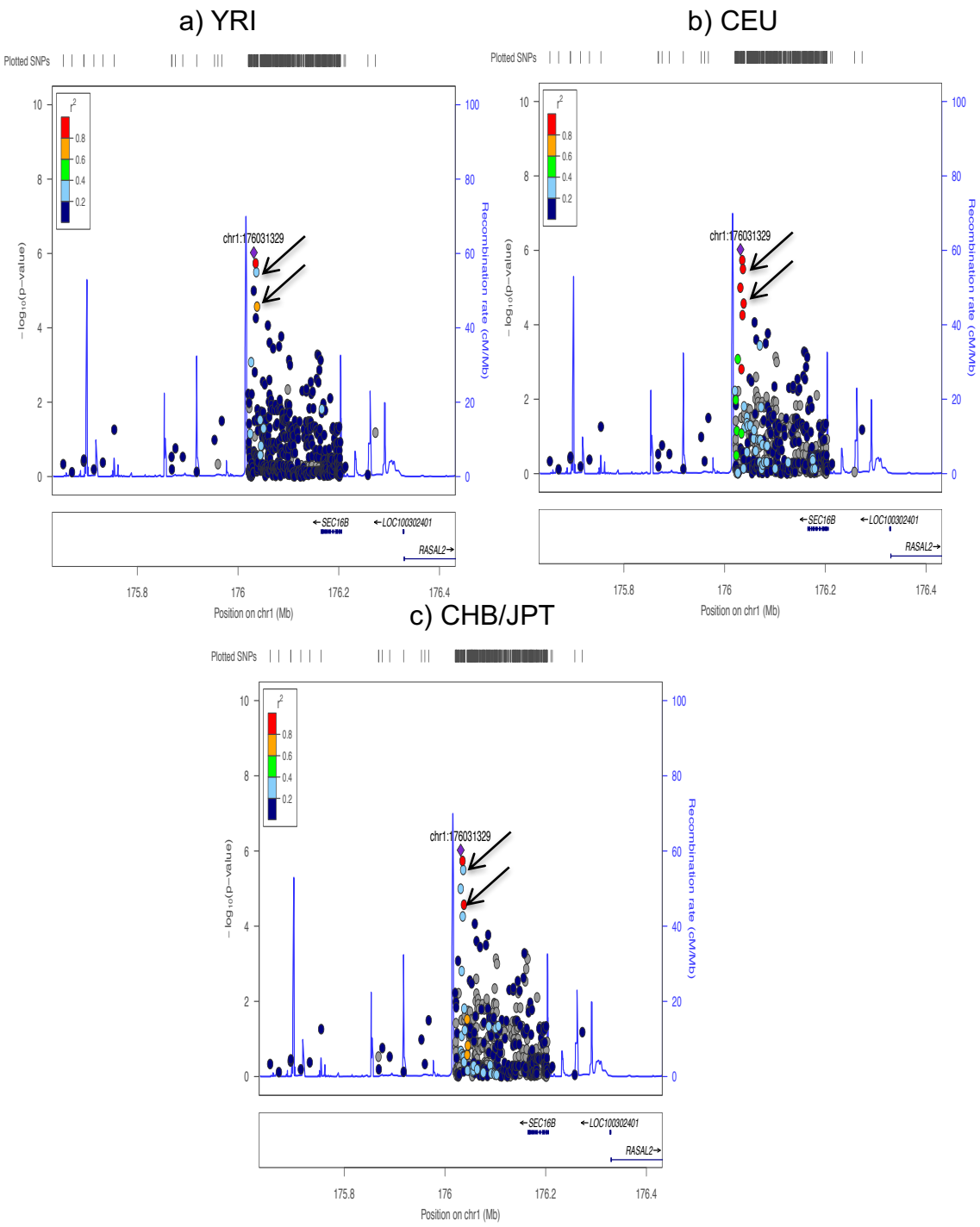


Figure 3.11 LocusZoom plots for fat mass and lead SNP rs6664268 in or near *SEC16B* in the combined dataset against various LD backgrounds

The presence of variants that have are trending towards significance that have a weaker correlation (as indicated by the arrow heads) with the index SNP (purple diamond), were observed against an African LD background. In Fig.s 3.11-3.13 the relationship of SNPs drawn against a European LD background, show that the cluster of associated SNPs are indistinguishable from one another, as reflected by their strong LD with the index SNP. Fig.s 3.11c-3.13c illustrates an intermediate effect of correlated SNPs against an Asian LD background. The recombination hotspots are illustrated in blue peaks, with neighbouring genes shown beneath the plot

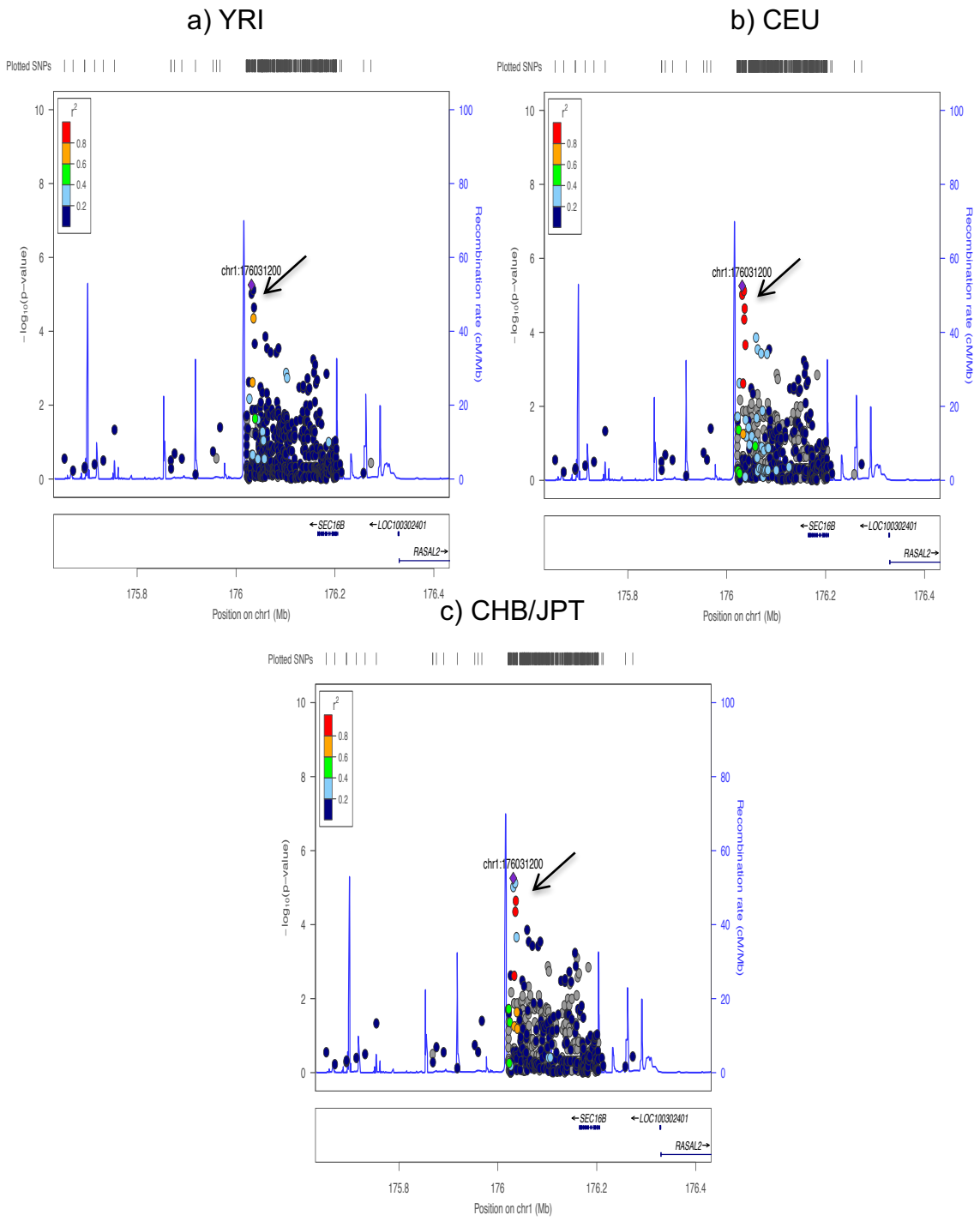


Figure 3.12 LocusZoom plots for fat mass and lead SNP rs6425446 in or near *SEC16B* in the combined females against various LD backgrounds

In this figure, the presence of two closely associated variants that have a weaker correlation (as indicated by the arrow) with the index SNP (purple diamond), were observed against an African LD background (a), this signal is indistinguishable against the European background (b), with an intermediate effect of correlated SNPs against an Asian LD background (c)

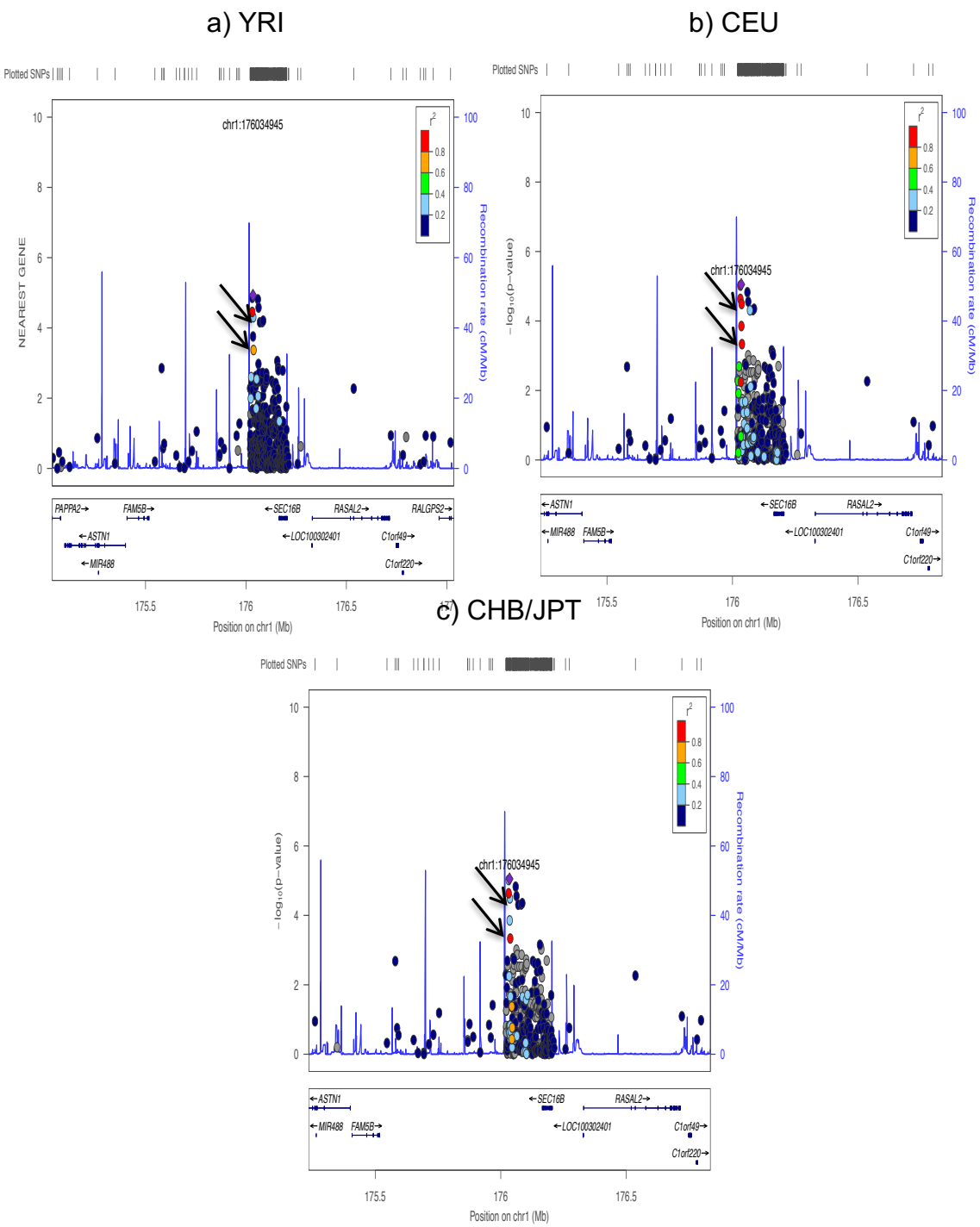


Figure 3.13 LocusZoom plots for PFM and lead SNP rs4075235 in or near *SEC16B* in the combined female dataset against various LD backgrounds

In this figure, the presence of two closely associated variants that have a weaker correlation (as indicated by the arrows) with the index SNP (purple diamond), were observed against an African LD background (a), this signal is indistinguishable against the European background (b), with an intermediate effect of correlated SNPs against an Asian LD background (c)

3.4.2 *Neuronal growth factor 1 (NEGR1)*

Variants in or near *NEGR1*, rs72941254, rs72941257, rs72941270 and rs72941224 (Table 3.7 and Table 3.9) were associated with fat mass and PFM in the young adult males only after adjusting for age and height (Fig. 3.14). The signal near *NEGR1* was only observed in the young male adults and variants were associated with positive effect allele changes as reflected by the β values. The association is stronger with fat mass (a) than with PFM (b) as illustrated by the *P*-values in Fig. 3.14. LocusZoom plots for the lead SNP are shown in Fig. 3.15 for the lead SNP in *NEGR1*.

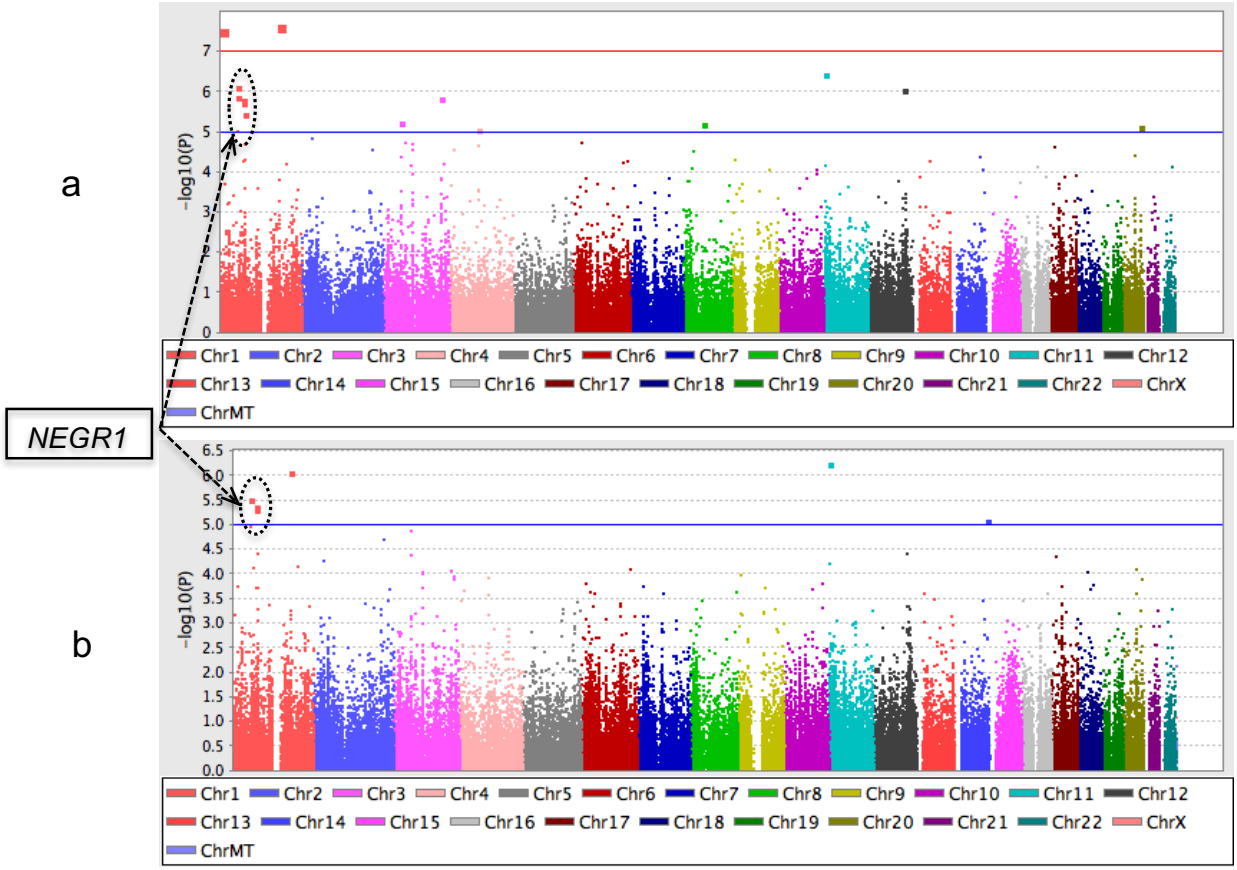


Figure 3.14 Manhattan plots for fat mass (a) and PFM (b) in the young male adults illustrating the cross-phenotype signals observed for *NEGR1*

The horizontal lines in the Manhattan plots correspond to suggestive *P*-values ≤ 5.0 x 10^{-5} (blue) and genome-wide *P*-values ≤ 5.0 x 10^{-7} (red)

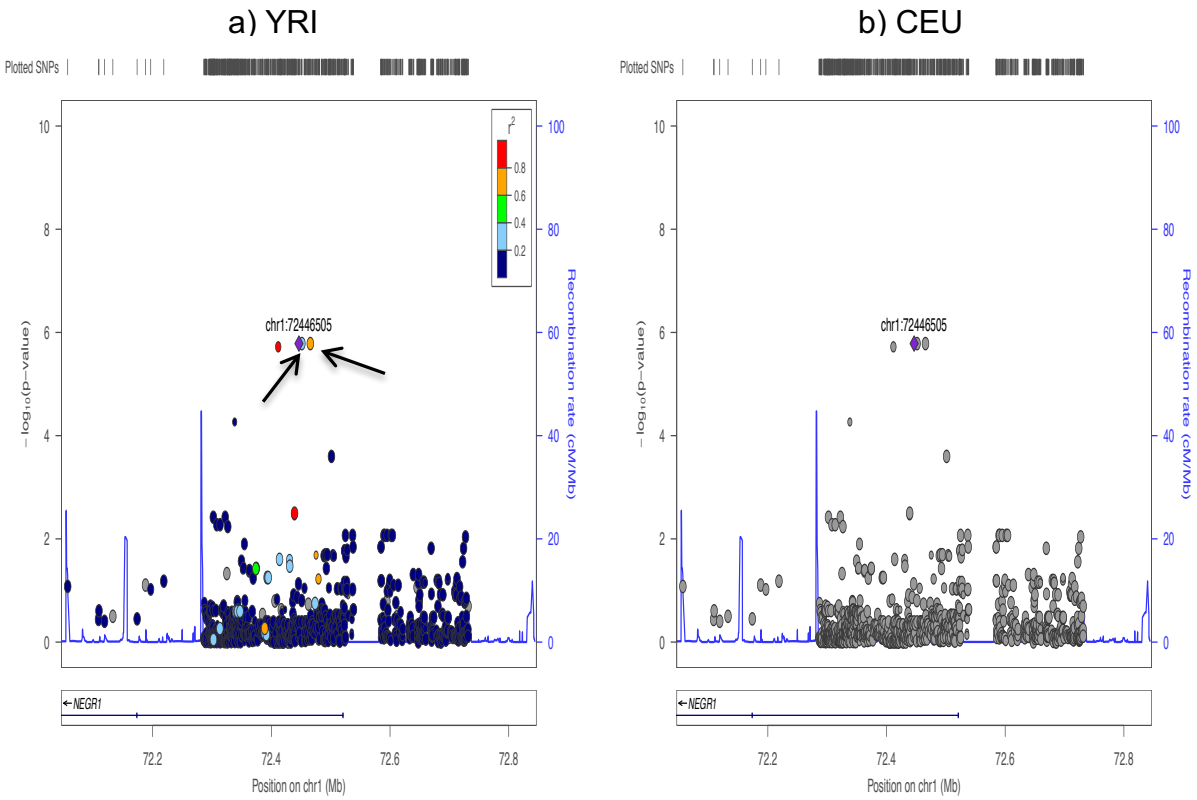


Figure 3.15 LocusZoom plots for fat mass and lead SNP rs72941254 near *NEGR1* in the young male adults with LD structure shown against a YRI (a) and CEU (b) background
In this figure, the presence of very closely associated variants that have a weaker correlation (as indicated by the arrows) with the index SNP (purple diamond), were observed against an African LD background (a), there was no available LD information to assess the correlation of SNPs against index SNP for a European background (b)

3.4.3 *Transient receptor potential melastatin 7 (TRPM7)*

SNPs in or near *TRPM7* were associated with WC (10 SNPS) and HC (10 SNPs) at GW significance in young males with rs62021060 as the lead SNP for WC and rs17598264 as the lead SNP for HC reported in Tables 3.4 and 3.5, respectively and illustrated in Fig. 3.16. Variants associated with BMI (7 SNPs, rs17598819, lead SNP) were trending toward GW significance, reported in Table 3.3. The effect alleles were associated with increases in the respective phenotypes as reflected by the positive β values. Regional plots are shown in Fig. 3.17, however there is no available African LD structure for the variants around *TRPM7*. The regional plots against a European LD background (Fig. 3.17b), shows that there is strong SNP coverage around the *TRPM7* locus with several SNPs in strong LD with the index SNP.

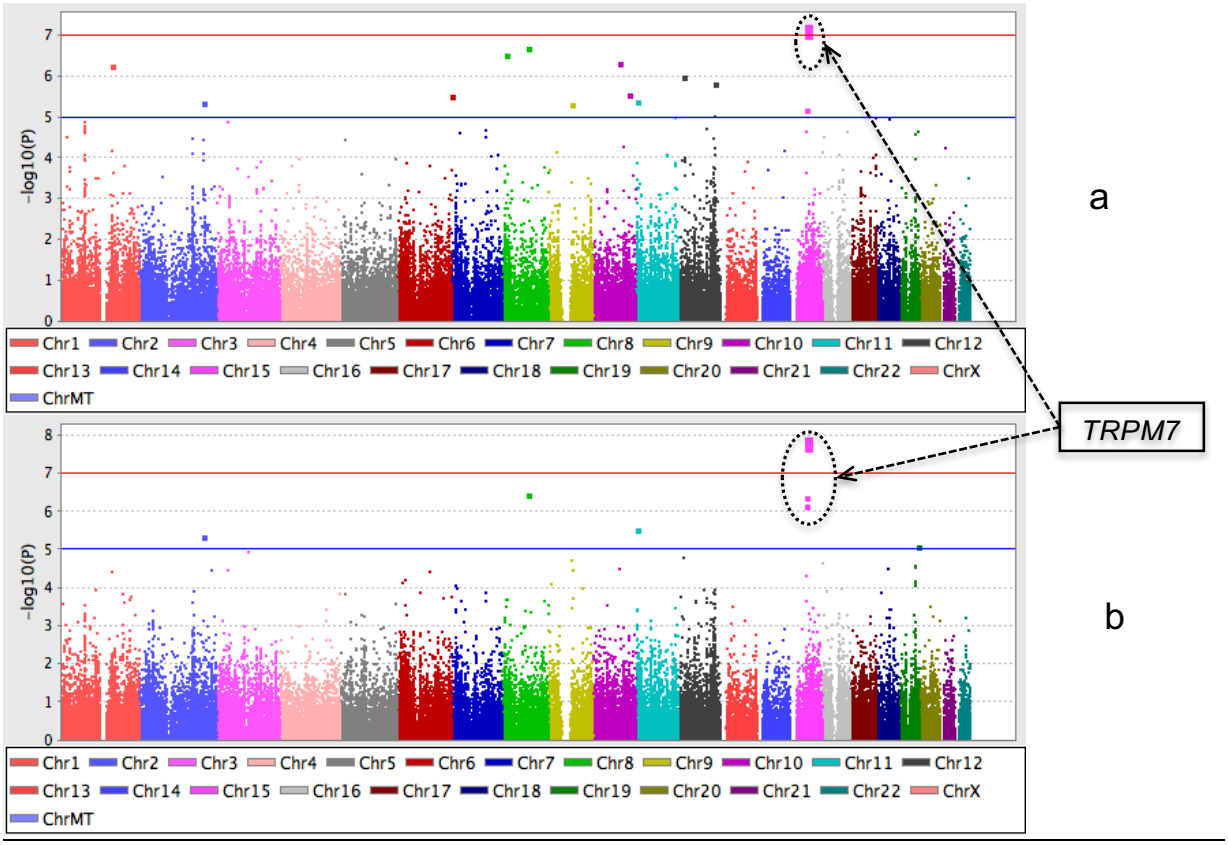


Figure 3.16 Manhattan plots for WC (a) and HC (b) in the young adult males where the cross-phenotype signals observed (at genome-wide significance) in or near *TRPM7* are shown

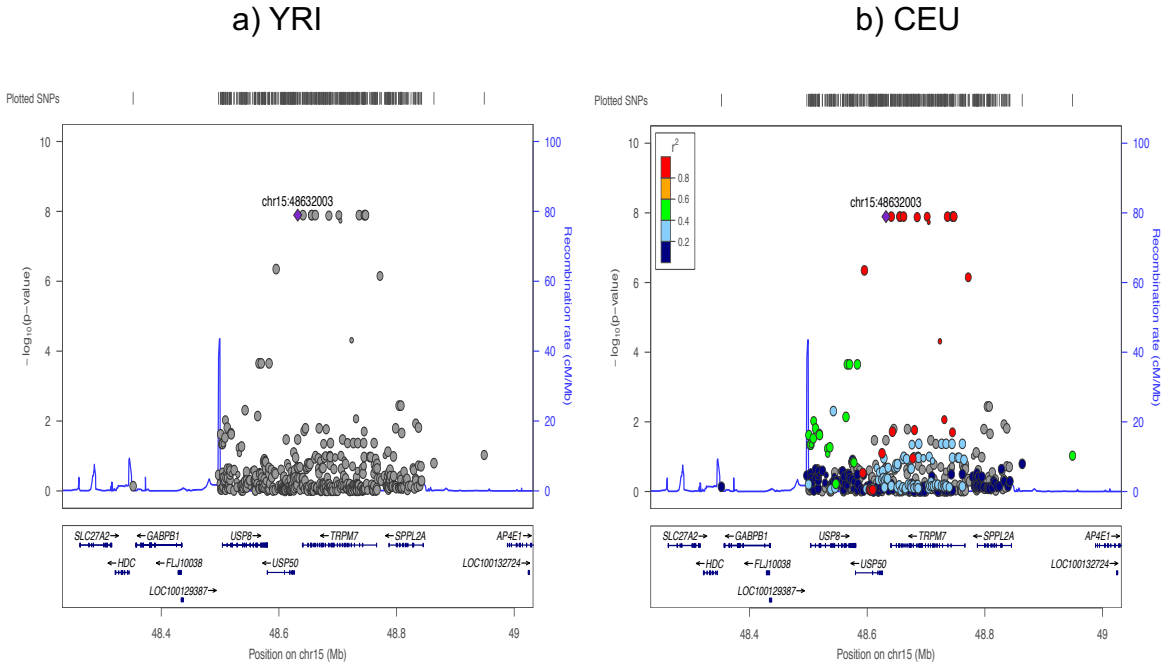


Figure 3.17 LocusZoom plots for HC and lead SNP rs17598264 near *TRPM7* in the young male adults against with LD structure shown against a YRI (a) and CEU (b) background

3.4.4 *Mitochondrial form of tryptophanyl-tRNA synthetase (WARS2)*

Three SNPs, rs17023092, rs74112264 and rs17023118 in the untranslated region (UTR) were associated with WHR (Fig. 3.18a), whilst a single intronic SNP rs56750694 was associated with HC following adjustment for covariates when the datasets are combined. Two intronic SNPs were associated with WHR and *WARS2*, rs12095241 and rs12088290 in the young adults shown Fig. 3.18 (b), where the signal observed is stronger than that observed when the datasets are combined.

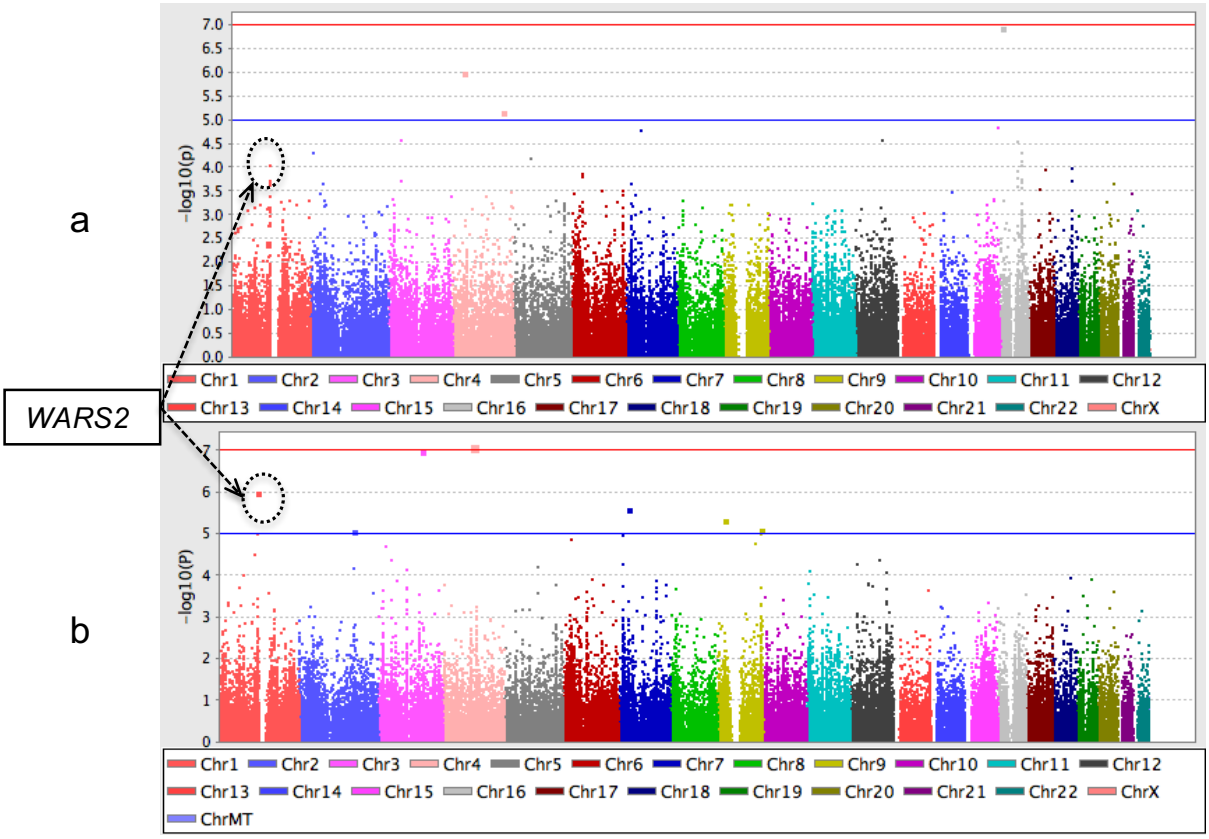


Figure 3.18 Manhattan plots for WHR in the young adults (a) and the combined all dataset (b) where the signals observed in the different sub-groups in or near *WARS2* are indicated
The signals in the young adults (b) are suggestive of significance, whilst the signal in the combined dataset (a) is attenuated compared to the young adults

3.4.5 *Contactin Associated Protein-Like 5 (CNTNAP5)*

Two SNPs (rs6541885 and rs4411698) were suggestively associated with BMI after adjusting for age, together with HC, fat mass and PFM after adjusting for both age and height in the female caregivers. The signals at *CNTNAP5* were mainly observed in the older female caregivers, with only a single SNP rs6541885 associated with fat mass in the combined females. The association signal near *CNTNAP5* is stronger in the caregivers with PFM (b) than with fat mass as illustrated in Fig. 3.19 (a). Negative β values were observed for all

minor allele variants associated with the intergenic region near *LOC72841* | *CNTNAP5* gene clusters. There was no available LD information around this region to draw regional plots.

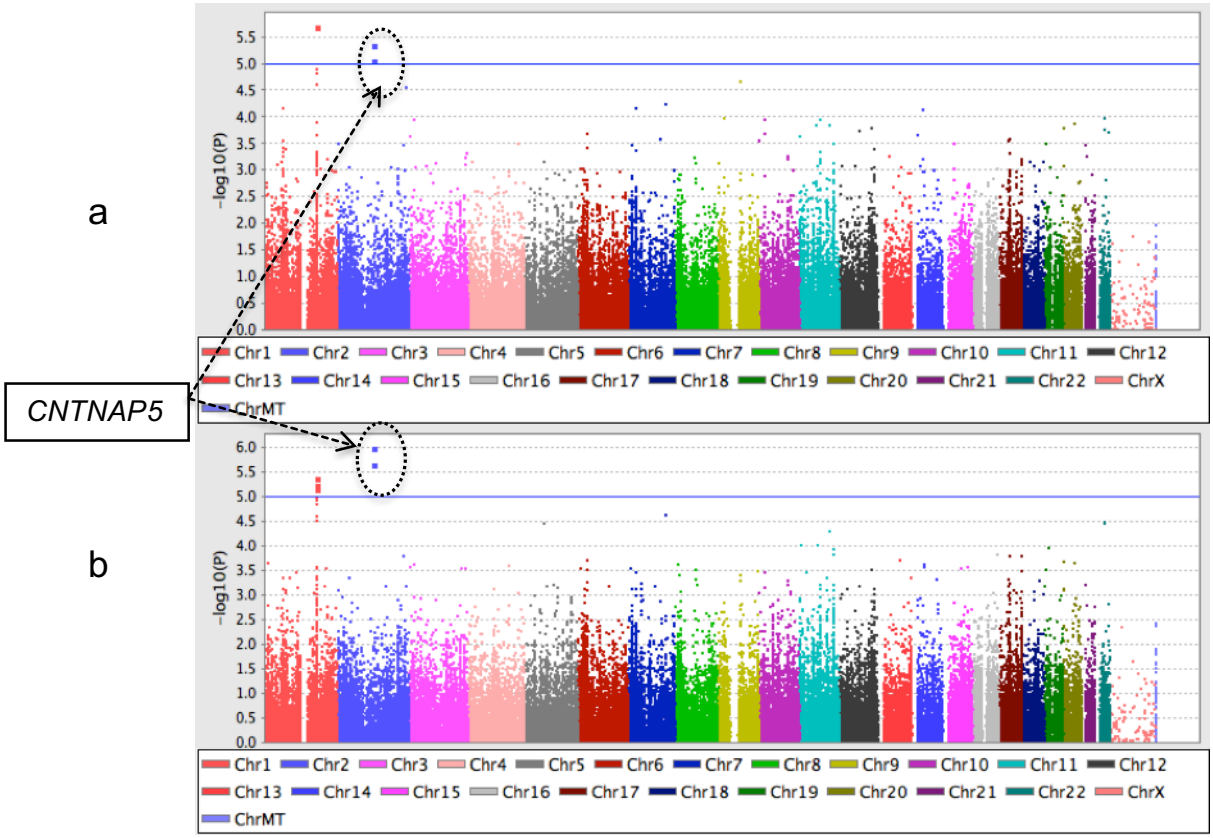


Figure 3.19 Manhattan plots showing the cross-phenotype association for *CNTNAP5* in the female caregivers The Manhattan plots refer to fat mass (a) and percentage body fat (b) where the suggestive association signals observed in or near *CNTNAP5* are more strongly associated with PFM than fat mass in the female caregivers

3.4.6 *SP110 nuclear body (SP110)*

A single intronic variant rs2114591 was associated with BMI, and was the strongest signal associated with WC and LM following adjustment for covariates, in the combined dataset. It was also the lead SNP correlated with WC and lean mass in the combined females. This SNP was associated with WC and fat mass in the female caregivers following adjusted linear regression, where the association with WC was stronger than with fat mass.

3.4.7 *Lipoprotein(a)-like 2) (LPAL2)*

After adjusting WC and HC for age and height, two SNPs, were trending towards significance in the female caregivers only, intronic SNP rs9364558 and rs115553347 an intergenic SNP in or near *SLC22A3* and *LPAL2*. This SNP was only observed in the female caregivers.

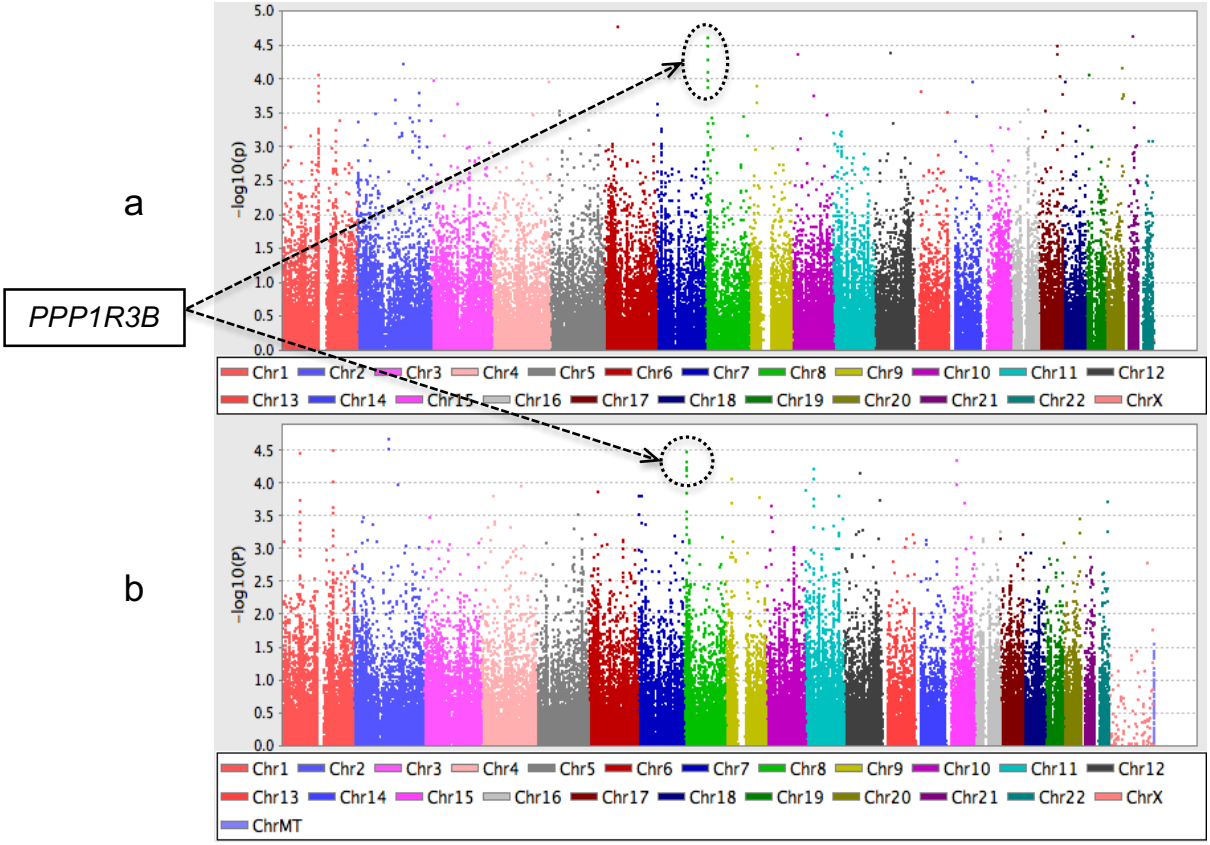


Figure 3.20 Manhattan plots showing the sub-group association with *PPP1R3B* and HC in combined dataset (a) and the female caregivers (b)

3.4.8 *Protein phosphatase 1 regulatory subunit 3B (PPP1R3B)*

Following adjustment for PC1 & 2, sex, age and height, three SNPs, rs11778774, rs78933755 and rs73535332, were trending towards significance with HC in the combined dataset illustrated in Fig. 3.20 (a). Two SNPs rs78933755 and rs35584813 were trending toward significance in the female caregivers and HC in Fig. 3.20 (b). A signal (5 SNPs) trending towards significance was correlated with lean mass in the combined females and female caregivers (Table 3.8) following adjustment for covariates but becomes attenuated when the datasets are combined.

3.4.9 *Brain-derived neurotrophic factor-opposite strand (BDNFOS)*

Following adjustment with age and height two SNPs were trending towards significance with HC in the female caregivers shown in Fig. 3.21 (a). When combining only the

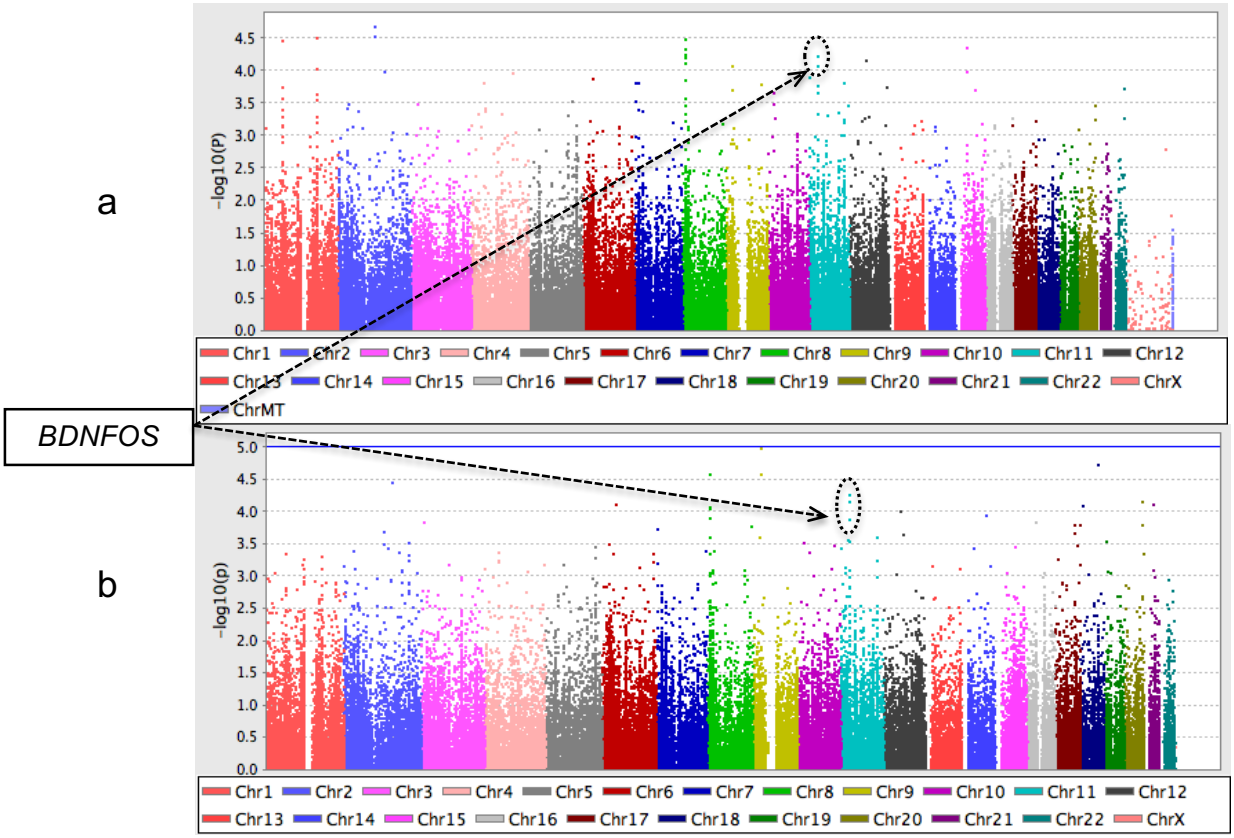


Figure 3.21 Manhattan plots showing the sub-group association with *BDNFOS* and HC in combined dataset (a) and the combined females

females (young female adults and female caregivers) following adjustment for first 10 PCS, age, and height, three SNPs were significant with HC, rs58174260, rs12574325 and rs16917135, shown in Fig. 3.21 (b).

3.4.10 *Zinc finger FYVE-type containing 9 (ZFYVE9)*

A single SNP, rs2753399 in *ZFYVE9* was associated with BMI, WC, HC and fat mass in the young adults following adjustment for covariates. Associations were then observed for this SNP rs2753399 with fat mass and PFM in the young male adults after adjusting for age and height.

3.4.11 *Apolipoprotein H (APOH)*

SNPs, rs115012414 and rs77612309 in the *APOH*| *PRKCA* gene cluster were suggestively associated with BMI, HC in the combined dataset. The association with BMI persisted in the combined female dataset only with both SNPs rs115012414 and rs77612309 trending

towards significance whilst the signal for HC was attenuated. The signal for *APOH* | *PRKCA* gene complex was not found in any of the other stratified datasets.

3.5 Age and sex-specific interactions

The dataset was stratified into various groups (summarised in Table 3.11) to establish if there were any age and sex-specific associations. Variants in or near *TRPM7, SLC17A7, COBLL1* and *NEGR1* were associated with different traits in the young males only whilst variants in or near *CNTNAP5* and *BDNFOS* were associated with females only (combined young and old).

Variants in or near *TMEM18* were associated with young females only whilst variants in or near *LPAL2* were associated with older female caregivers only. Also, when comparing older female caregivers to young adults (mixed sex), variants in or near *ZFYVE9*, *WARS2* and *NBEAL1* were only associated with young adults.

Table 3.11 Summary of age and/or sex-specific interactions

The main genetic loci (17) associated with body composition were compared to the group stratifications to assess age and sex interactions. The colour key is found below the table indicating the various stratifications in the dataset

Loci	Groups					
	Combined all	Combined females	Female caregivers	Young Adults	Young females	Young males
APOH | PRKCA	x	x				
LOC728241 | CNTNAP5		x	x			
ZFYVE9				x		
TRPM7						x
SP110	x	x	x			
NRXN3 | LOC100131580	x	x				
LPAL2			x			
PPP1R3B | LOC100129150	x	x	x			
WARS2				x		
BDNFOS		x	x			
LOC400796 | SEC16B	x	x	x	x		x
FTO	x	x				
LOC100129474 | SLC17A4						x
COBLL1						x
LOC727944 | TMEM18					x	
NEGR1 | LOC100132353						x
NBEAL1				x		

	Females only
	Young adults
	Young males
	Young females
	Older females

4. Discussion

This study used the Metabochip as a tool, the content of which includes the top-hits from GWAS of body composition traits including BMI (24 fine-mapping loci; 5065 replication SNPs), WHR (15 fine-mapping loci; 5056 replication SNPs), percentage body fat (1035 replication SNPs) and WC (2 fine-mapping loci; 1048 replication SNPs) (Voight et al., 2012). This tool has proved to be useful in many European and African American populations in the discovery and replication of variants contributing to obesity-related traits, but its use in sub-Saharan African (SSA) populations such as the Sowetan cohort has yet to be established, until now. A major aim of this study was to assess if association with obesity markers from GWAS in European populations were transferable to African populations. Genetic association analyses have replicated loci for increased obesity risk including *SEC16B, FTO, NEGR1* and *TMEM18* and led to the discovery of new signals that may correlate with an increased risk for obesity-related traits. Although the loci in which the associations occur have been replicated, different SNPs are implicated in Africans when compared with European, African-American and Asian populations. This could be attributable to ascertainment bias given that most of the SNPs on the Metabochip were informed from GWAS studies conducted in European populations. An important finding of this study was the observation of at least ten cross-phenotype associations (only the top 10 hits were reported). This is where a genetic locus previously associated with mainly variations in BMI have been shown to be associated with better surrogates of adiposity, DXA-derived body composition or traits associated with fat distribution.

4.1 Population Structure and PCA

PCA (PC1 and PC2) showed that the Bt20 dataset represents a homogenous group, as demonstrated by the clustering of individuals which confirms the demographics of urban Soweto, as observed by May et al. (2013). The Bt20 samples showed a close clustering, overlapping with other Bt20 participants (black Sowetan samples) and southeastern Bantu-speakers, thus implying a common genetic origin. This is expected for the black Sowetan samples as they were drawn from the larger Bt20 cohort. The Bt20 dataset together with the black Sowetan samples and southeastern Bantu samples, self-identify as being Sotho-speakers speaking southeastern Bantu languages. Southeastern Bantu languages belong to the Niger-Congo (Niger-Kordofanian; NK) ethno-linguistic group, being one of the four major language groups spoken in Africa (Li et al., 2014), along with Afroasiatic, Nilo-Saharan and Khoe (Wood et al., 2005). Both the Yoruban (west African) as well as the Luhya (east

African) from 1000G dataset speak a Niger-Congo language, which contribute to their closer clustering with the Bt20 dataset. The Masaai from Kenya (east African), who speak a Nilo-Saharan language, as expected, cluster further away from the Bt20 dataset. Further, the SWB, which clustered more closely to the Bt20 group, also speak a Niger-Congo language. Results from PCA show that language (based on self identification) has a high correlation with genetic variation. This correlation has been noted in other studies (Nettle and Harriss, 2003, Schlebusch et al., 2012, May et al., 2013, Li et al., 2014).

4.2 SNP and sample quality control

The results from quality control (illustrated in Fig. 3.4 and 3.5) showed that the greatest number of SNPs removed during QC were due to SNPs being rare or monomorphic in the Sowetan population. The increased number of monomorphic SNP observations reflects the limitations of using a tool designed for capturing common variation in European populations, in African populations. These results were consistent with the use of Metabochip in an African HapMap population (Yoruba), where ~24% of variants were found to be monomorphic (Crawford et al., 2013), also to a lesser extent in an African American dataset (Buyske et al., 2012). According to (Buyske et al., 2012) almost ~22% of polymorphic Metabochip SNPs have MAFs less than 0.025% in African Americans. Other differences such as genetic drift, environmental factors, recent mutations may also contribute to differences observed in the MAFs (Fu et al., 2011).

Some samples were removed due to poor DNA quality. Good quality DNA is imperative to successful genotyping, especially on large genotyping platforms (Adler et al., 2013). DNA quality and concentration is influenced to a large extent by handling, extraction method and possibly storage conditions (Tan and Yiap, 2009). The study utilised previously extracted DNA samples by salting-out, the quality of which may have been affected by user handling during the extraction process together with ageing reagents.

4.3 Adjustments for multiple testing

The failure to detect variants at loci that correlate with a particular trait can be due to false negatives associations. This arises as a result of effect sizes being too small causing associations not to reach GW significance. Therefore it was decided that the GWAS Bonferroni correction might be too strict and *P* values ≤ 5×10^{-5} were considered suggestive of association, as supported by other reported findings in the literature. The issue

surrounding the strictness of the Bonferroni correction has been debated over the years with some claiming that the correction is in fact too harsh (Perneger, 1998, Cordell and Clayton, 2005, Armstrong et al., 2009). Bonferroni correction is based on the assumption that each test is independent, but does not take into consideration the unknown correlation of SNPs that are in strong LD with each other (Armstrong et al., 2009). Other researchers have opted for the use of a false-discovery rate (FDR) correction, which measures the number of Type I errors (false positives) one would have to take into consideration for a result to be a real discovery and therefore reject the null hypothesis (Benjamini and Hochberg, 1995, Armstrong et al., 2009, Carlson et al., 2013).

4.4 Power to detect associations

The PowerGWAS/QT software, designed for use in GWAS of quantitative traits, was used to assess the power to detect genetic associations in this study. The power calculations were based on mean differences in BMI in this dataset, using effect sizes based on European populations as outlined by (Speliotes et al., 2010). These estimates were used, as the next-best proxy as no other data was available from African ancestry populations and the implications are further addressed in Chapter 5. Although the power analyses showed that the study was sufficiently powered to detect associations with moderate effect sizes, a limitation of the analysis was that effect size ranges were based on European estimates. The power to detect associations with other traits like WHR, WC, HC, fat mass, lean mass and PFM were not assessed in this study.

4.5 Association Analyses

4.5.1 *SEC16B*

In this study, variants in or near *SEC16B* were observed, with rs6664268 (fat mass- $P_{adj}=9.48 \times 10^{-7}$ and PFM- $P_{adj}=7.90 \times 10^{-7}$) the lead SNP in the merged dataset for both fat mass and PFM, while rs6425446 ($P_{adj}=5.55 \times 10^{-6}$) was the lead SNP for fat mass and rs4075235 ($P_{adj}=8.89 \times 10^{-6}$) with PFM when combining females only. Many of the variants observed during association analysis were suggestive of significance ($P \leq 10^{-5}$) with only variants in or near *SEC16B*, *TRPM7, COBLL1* and *SLC17A4* reaching Metabochip genome-wide significance. Evidence for strong signals in or near the *SEC16B* locus was illustrated in Manhattan plots for fat mass and PFM in the caregivers, merged dataset and combined females. There was a concomitant increase in the signal observed and number of SNPs associated in or near *SEC16B* with an increase in sample size (N=972-caregivers; N=1421-

combined females; N=1926-merged dataset). A single association signal in or near *SEC16B* was observed in the young males only. This may allude to the signal in or near *SEC16B* locus being associated with increased obesity risk with body composition measures (high levels of fat mass and PBF) in older individuals, however this remains to be established as older males were unavailable to assess the association.

The results from regional plots highlighted that the lead variants in or near *SEC16B* suggest the possibility of multiple independent association signals in or near this locus, with the presence of associated (albeit not at GW-significance) but not strongly correlated signals (Pruim et al., 2010). Also variants in or near *SEC16B* are indistinguishable from other SNPs that are highly correlated against a European LD background, with intermediate correlations observed in Asian populations. This illustrates the value of using diverse populations that display high levels of genetic diversity and low levels of LD in dissecting possible independent association signals and thus narrowing in on potential causal loci. Possible independent signals near this locus therefore warrant more studies with increased sample sizes in other diverse African populations.

SEC16B variants (index SNP rs543874 or rs10913469) in association with body composition have been observed in several continental populations including European (Thorleifsson et al., 2009, Speliotes et al., 2010, Bradfield et al., 2012), Asian (Hotta et al., 2009, Ng et al., 2010, Okada et al., 2012, Wen et al., 2012) and African Americans. In African Americans it was the most strongly associated signal observed in a fine-mapping study using the Metabochip (Gong et al., 2013). This locus is a widely replicated obesity risk locus associated with increased BMI in both adults and children (Zhao et al., 2009, Bradfield et al., 2012, Felix et al., 2015). Currently only one study, an Asian (Japanese) study (Hotta et al., 2011), has specifically assessed the association of *SEC16B* variants (rs10913469) with other body composition data. In this study they assessed CT-derived visceral fat area and subcutaneous fat area. This study did not find any associations with *SEC16B* or other obesity risk variants (12 variants at 10 obesity risk loci) examined, with the exception of a single association with *SH2B1* variants and visceral fat area. More recently a (mainly) European meta-analysis (Lu et al., 2016) on percentage body fat reported associations with *SEC16B* variant rs543874 and increases in percentage body fat. This study also highlighted and emphasised “cross-phenotype associations” in their study, where a trait previously associated with increased BMI, was also associated with another trait, in this case increases in percentage body fat. Besides this current study, *SEC16B* variants have not been examined/reported in a SSA population, nor has a genetic association study with body composition traits (DXA- derived fat mass and PFM) other than BMI, been carried out.

SEC16B encodes the long Sec16L and the short Sec16 proteins, which are required for transport of secretory molecules from the endoplasmic reticulum (ER) to the Golgi apparatus (Hotta et al., 2009, Budnik et al., 2011). Studies carried out in yeast models (*Saccharomyces cerevisiae*) have shown that mutations in *SEC16B* genes result in the absence of vesicles (Budnik et al., 2011). Since vesicles are important in the transport of molecules from the ER to the Golgi it implies an effect on the efficient transport of molecules. Hotta and collegues (2009) have postulated that the Sec16 protein plays a role in the transport of appetite-regulatory peptides such as neuropeptide Y and propiomelanocortin and may be involved in the regulation of these appetite-related peptides (Hotta et al., 2009). This protein is mainly expressed in adipose tissue and the gastrointestinal organs (Hotta et al., 2009). However, a study scomparing obesity related genes in Zucker diabetic fatty rats to that of its lean normoglycaemic counterpart showed that *SEC16B* was one of the only obesity-risk variants (of those tested) that was not expressed in the hypothalamus, but rather in subcutaneous adipose tissue, implying a more peripheral role in the regulation of obesity (Schmid et al., 2012).

4.5.2 *TRPM7*

This study reports the novel association of variants in the intron of *TRPM7* with measures of obesity. The signal for *TRPM7* was correlated with WC following linear regression analysis of WC in the young males. However when the dataset was stratified into males and females, the signal in *TRPM7* becomes stronger in the young males and no association signal was observed in the young females. This would imply that it is the young males driving the association in the young adults. Variants in *TRPM7* in the young males are associated with BMI (suggestive association) and with WC and HC at GW significance. The correlation of these variants with increases in WC and HC have yet to be replicated in a larger mixed sex cohort to assess if these variants are associated with males only.

TRPM7 is a member of the melastatin-related subfamily of proteins, also known to be associated with cell proliferation (Park et al., 2014). TRPM7 together with TRPM6 are ubiquitously expressed throughout the body and have been implicated in being involved with cellular magnesium homeostasis, which is critical for insulin and glucose metabolism (Song et al., 2009). A study examined whether there is an association between exonic variants in *TRPM6* and *TRPM7* and increased risk for type 2 diabetes but they did not find any significant associations (Song et al., 2009). A review (Liu et al., 2008a) highlighting the role of TRPM channel proteins in metabolic disease pointed to disruptions in the expression or

function of transient receptor potential (TRP) proteins that may result in increased cardiovascular risk in patients with metabolic syndrome. Some TRP molecules have been implicated in regulating adipogenesis, regulation of insulin secretion, lipid composition, and atherosclerosis with others including TRPM7 have been involved in hypertension (Liu et al., 2008a). Although these *TRPM7* variants have not been previously associated in GWAS with increases in WC and HC, they might play a role in affecting the expression of other TRP molecules that influence adipogenesis in young males.

4.5.3 *SLC17A7*

A variant intergenic to *LOC100129474* | *SLC17A7,* rs3923725 was associated with WHR only in the young males at genome-wide signficance. *SLC17A1* variants have been associated with cardiovascular disease risk factors where probable functions include sodium-dependent phosphate transmembrane transporter activity and symporter activity (White et al., 2016). *SLC17A7* is an important paralog (paralogs are genes related by duplication and evolve new functions which are related to the original gene) of *SLC17A1*.

4.5.4 *COBLL1*

A single variant in *COBLL1* was associated at GW-significance with WHR in young males. An intergenic region (*COBBL1-GRB14)* including this locus has recently been implicated in a GWAS as a novel locus that results in an increase in body fat percentage (Lu et al., 2016). This study showed that the body fat percentage increasing allele of *COBBL1-GRB14* was associated with reduced WHR_{adjBMI}, implying a gluteal rather than an abdominal fat storage mechanism of action. This was not observed in the current study with variant rs115743734, as it was only associated with increased unadjusted WHR in the males. It has also been noted that variants in or near *COBBL1-GRB14* have previously been associated with T2D risk, fasting insulin, triglycerides, HDL-cholesterol (Lu et al., 2016), indicating cross-phenotype associations.

4.5.5 *FTO*

SNPs in or near *FTO* are the most universally associated with obesity across all populations. Results from this study showed three intronic variants in *FTO* to be associated with WHR when merging the females, driven by lead SNP rs1861554. This may suggest that in this study variants in *FTO* might play a role in the distribution of body fat more so in females, however this finding needs to be replicated in a larger dataset. The distribution of body fat in females (more subcutaneous - specially deposited at the hips and thighs) differs to that of males who have more visceral fat, that is preferentially deposited around internal organs and

mostly accumulates around the waist (Randall et al., 2013). Varying hormone levels in males versus females are often associated with differences in body fat distribution which may also contribute to differences in body fat distribution (Randall et al., 2013). No significant associations were observed with *FTO* variants and BMI. *FTO* variants (mainly rs993906 and tagSNPs) have unequivocally been replicated against various ethnic backgrounds including a SSA population (Lombard et al., 2012) in both children and adults and with other body composition traits including WHR (Heid et al., 2010, Vasan et al., 2013).

Most functional studies (mouse models, *in vitro* and *in vivo*) related to *FTO*-GWAS identified variants have argued its link to mainly three candidate genes although there are other proposed mechanisms. These mechanisms aid in trying to unpack how *FTO* variants contribute to the obesity phenotype, which is described in a recent review article (Sandholt et al., 2015). The three candidates that have been identified include *RPGRIP1L (retinitis pigmentosa GTP-ase regulator- interacting protein-like 1)* (Peters et al., 2013), *IRX3* (*Iroquois homeobox protein 3*) and *IRX5* (*Iroquois homeobox protein 5*) (Smemo et al., 2014, Claussnitzer et al., 2015). The fine-mapping study on *FTO* variants in intron1 and *RPGRIP1L* in an African American population (N= 20 000) showed a reduction of the risk interval within the European LD block initially containing 104 SNPs down to 29 SNPs against an African LD background. Of the SNPs narrowed down, rs8050136 (in high LD with GIANT index SNP) affects a binding site, which resulted in reduced levels of mRNA expression for both *FTO* and *RPGRIP1L*. Since *RPGRIP1L* has been linked to the movement of Leptin receptors in the hypothalamus, it was postulated that the cellular response to Leptin could be influenced by *FTO* variants acting via *RPGRIP1L* (Peters et al., 2013, Sandholt et al., 2015). Using circular chromosome conformation capture (4C) experiments, (Smemo et al., 2014) have shown that intron 1 *FTO* variants interacts with the promoter region of homeobox gene *IRX3* and that obesity-risk SNPs in *FTO* are also associated with the expression of *IRX3* and not *FTO* in the brain. Further, when *Irx3* was knocked down in mice, the resulting phenotype in the mutant mice showed a reduction in body weight, and fat compared to its wildtype counterpart. The most recent study by (Claussnitzer et al., 2015) conferring a link to *IRX3/IRX5,* showed that a thymine (T) to cytosine (C) change in variant rs1421085, affects a conserved motiff for ARID5B which influences the expression of *IRX3/IRX5.* The subsequent increased expression of IRX3/IRX5 impacts the expression of other genes that affects the outcome of a preadipocyte, pushing its fate in favour of white adipocytes for storage over beige adipocytes associated with increased thermogenesis (generation of energy in the form of heat). When *IRX3* or *IRX5* was knocked down in white adipocytes from patients with rs1421085 C-allele, the effects were reversed, with an observed increase in thermogenesis (Claussnitzer et al.,

2015). It is enticing to pinpoint the causal interaction of *FTO* variants to anyone of these candidates, but the exact mechanism in contributing to obesity remains unknown.

4.5.6 *NEGR1*

Variants in or near *NEGR1* were associated with fat mass only in the young males. The regional plot with focus on the lead SNP rs72941254 near *NEGR1* best illustrates the possible multiple independent signals observed using African LD structure, unfortunately the LD information for European and Asian backgrounds were not available for comparison. The *NEGR1* variant rs2568958 was assessed in a mixed sex, adolescent subset of the Bt20 cohort (Chapter 2) and no significant associations with BMI were observed. Since this signal is only observed in young adult males (N~500), it is tempting to speculate that this signal near *NEGR1* plays more of an important role in contributing to adiposity in younger males, however this warrants replication in other African populations of increased sample size.

NEGR1 variants were amongst the first variants to be associated with increased BMI (Thorleifsson et al., 2009, Willer et al., 2009, Speliotes et al., 2010). *NEGR1* variants have been associated with two deletions (conserved non coding DNA) upstream of the gene, the smaller of which removes a binding site for NKX6.1, a strong transcriptional repressor (Willer et al., 2009, Wheeler et al., 2013). This has been implicated in contributing to the obesity phenotype. Neuronal growth regulator 1 (NEGR1*)* has been shown to be highly expressed in the hypothalamus (Willer et al., 2009) and is part of a family of cell adhesion molecules that have a putative role in regulating neuronal outgrowth (Schäfer et al., 2005). Animal model studies involving a loss of function of *Negr1* resulted in mutant mice with reduced weight and body size and energy output (Lee et al., 2012), suggesting a role in the regulation of energy balance as well as body weight. Another study focussing on the expression of *Negr1* in rats confirmed the strong expression of *Negr1* in the hypothalamus and demonstrated a role for *Negr1* in the regulation of energy expenditure by controlling the intake of specific nutrients (Boender et al., 2014). *NEGR1* has also been shown to be expressed in other tissues outside of the brain, including white adipose tissue and smooth muscle; also differential expression of *NEGR1* has been observed in subcutaneous adipose tissue (SAT) in lean versus obese siblings, suggesting that *NEGR1* might have a function in adipose tissue, possibly playing a role in regulating adipogenesis (Speakman, 2013).

A study investigating whether GWAS-identified loci for T2D and obesity have been subject to recent selective pressures showed that of all the obesity-related loci examined in the various continental populations, *NEGR1* showed the greatest differentiation among sub-Saharan

Africans (Klimentidis et al., 2011) together with evidence of selection from another African population (Ethiopian-Wolaita) (Tekola-Ayele et al., 2015). This can be seen as suggestive evidence of recent selection influencing genetic variants in or near *NEGR1* that could then potentially impact on metabolism and coping mechanisms for periods of famine and adapting to environments during human history (Tekola-Ayele et al., 2015).

4.5.7 *TMEM18*

Variants (lead SNP rs114285121) in or near *TMEM18* were only associated with young females and fat mass. The variant rs6548238 (index SNP in Europeans studies) in or near *TMEM18* was previously described in a candidate gene study of six obesity-associated variants in an adolescent mixed sex subset of Bt20 (outlined in Chapter 2) to be associated with increased BMI. Variants in or near *TMEM18* have been first described and replicated in adult as well as child cohorts (Thorleifsson et al., 2009, Willer et al., 2009, Zhao et al., 2009, Speliotes et al., 2010, Bradfield et al., 2012). No significant associations were observed with *TMEM18* variants and BMI in the adult Bt20 dataset, only the suggestive association of novel variants with fat mass in the young females. This possibly suggest that variants near *TMEM18* might play more of a significant role in contributing to overall adiposity earlier on in life, than in adulthood.

TMEM18 codes for a poorly described transmembrane protein, which has been suggested to be located in the nuclear membrane (Almén et al., 2010). It has been postulated that TMEM18 has a DNA-binding domain and binds transcription factors and thereby regulates transcription. However it remains unclear which genes are targeted (Speakman, 2013). TMEM18 seems to be a highly conserved protein and is ubiquitously expressed in the body (Almén et al., 2010, Speakman, 2013), but in humans it has been shown to be maximally expressed in adipose tissue (Bernhard et al., 2013). Some studies have shown increased expression in the hypothalamic appetite-regulating centre (Almén et al., 2010), but these results are localised to mouse models (Speakman, 2013). Its link to the obesity phenotype is not clear but it has been suggested that it might play a role in suppression of the genes implicated in energy expenditure or that its link to obesity is not via the central nervous system but might be related to peripheral functions in adipose tissue (Speakman, 2013).

4.5.8 *WARS2*

Three variants in the 3' untranslated region of *WARS2* (*tryptophanyl tRNA synthetase 2, mitochondrial)* were observed to be associated with WHR in the merged dataset. Variants in *TBIX15-WARS2* have previously been associated with increased WHR (Heid et al., 2010).

This study demonstrated that *TBIX15* showed definite depot-specific differences in adipose tissue expression in both humans and mice. Another study showed that although *TBIX15-WARS2* mRNA expression in SAT did not correlate with WHR, differential expression between VAT and SAT was observed for all the genes tested including *WARS2* (Schleinitz et al., 2014).

4.5.9 *CNTNAP5*

Two SNPs in the *contactin associated protein–like 5* (*CNTNAP5)* gene had a negative effect on HC, fat mass and PFM in the female caregivers. Only a single SNP, rs6541885 showed an association with fat mass in the merged females. No associations were observed in the combined dataset. These variants have not been described previously to be associated with obesity. The latest GIANT study, assessed the functional relevance of the 97 variants associated with increased BMI, by manually reviewing the literature around the 457 genes within 500 kb and r^2> 0.2 of these variants. They characterised the resulting variants according to function. The *CNTNAP5* gene was one of those surrounding genes and has been implicated in lipid biosynthesis and metabolism (Locke et al., 2015).

Variants in *CNTNAP2* have been implicated in a child with syndromic obesity (Vuillaume et al., 2014, Yazdi et al., 2015), while a recent GWAS on obesity and T2D risk in an Australian Aboriginal population has noted suggestive associations with BMI and variants in *CNTNAP2* (Anderson et al., 2015). CNTNAP2 is found in myelinated axons where it is required for localisation of the potassium-voltage gated channel protein KCN1 (Anderson et al., 2015). Mutations in mouse *Cntnap2* have resulted in obesity by disrupting the localisation of the mouse equivalent KCN1 (Kv1.1) (Yazdi et al., 2015). Perhaps variants in or near *CNTNAP5* contribute a similar effect towards body composition through a possible involvement with protein KCN1.

4.5.10 *SP110, LPAL2, PPP1R3B, BDNFOS, NBEAL1* and *NRXN3*

A single variant, rs2114591 in or near *SP110 (SP110 nuclear body protein) was* associated with WC (lead SNP), BMI and lean mass in the combined dataset and WC, BMI, fat mass and lean mass in the combined females. Variants in *SP110* have been reported in a study on childhood obesity in a Hispanic population that was suggestive of association with increased BMI (Comuzzie et al., 2012). A variant rs6727879 near *SP110* has also recently been shown to be nominally associated ($P<10^{-7}$) with SAT in a GWAS of sex-specific loci linked with abdominal and visceral fat distribution (Sung et al., 2015).

Two variants in or near *LPAL2* (*lipoprotein, Lp(A)-Like 2)* were associated with WC in the female caregivers. This locus together with *SLC22A3* and *LPA* form a gene complex, variants of which have previously been associated with increased risk for CHD (Trégouët et al., 2009).
Variants in or near *PPP1R3B (protein phosphatase 1 regulatory subunit 3B)* were trending towards significance in the combined dataset with HC. This locus has previously been identified as a susceptibility locus in a GWAS of obesity-related fatty liver disease conducted in a European population (Speliotes et al., 2011).

When combining females (young and caregivers) variants in or near *brain-derived neurotrophic factor opposite stand* (*BDNFOS)* were associated with HC. Several variants in or near BDNF have been associated with BMI across populations (Thorleifsson et al., 2009, Speliotes et al., 2010, Okada et al., 2012, Wen et al., 2012). A recent study has shown that the *BDNF* variant rs12291063 (C-allele) disrupts binding and transactivation via a transcriptional regulator and is associated with lower ventromedial hypothalamic *BDNF* expression and with obesity (Mou et al., 2015). BDNF is a pivotal player in the energy homeostasis pathway by being a downstream regulator of the propiomelanocortin pathway and could contribute to the obesity phenotype via this mechanism of action.

Variants in or near *NBEAL1 (neurobeachin like 1)* were tentatively associated with PFM in the young adults. Variants near this locus have been reported in a study on childhood obesity in a Hispanic population and were associated with increased fat mass change and deposition and energy storage (Comuzzie et al., 2012).

A single variant near *NRXN3* (*neurexin 3)* was associated with WC in the combined females and the merged dataset. The variant, rs10146997 in or near *NRXN3* has previously been described and replicated as a novel locus associated with increased WC in a European population (Heard-Costa et al., 2009, Bille et al., 2011). *NRXN3* encodes a protein that plays a role in cell adhesion in the vertebrate nervous system (Bille et al., 2011), but the mechanism by which it modifies abdominal fat deposition is unknown. Even though the association with WC near *NRXN3* was replicated in this study, there is a need for this signal to be replicated in a larger cohort.

4.6 Replication of GWAS associations in African populations

The majority of association signals that have been reported in this study show a replication of obesity risk loci identified in previous studies and has addressed the transferability of GWAS identified signals in a locus-wide manner. This suggests that GWAS-identified variants of

body composition are tagged by different SNPs in an African cohort. According to (Lu and Loos, 2013) this type of comparison is more detailed and takes into account differences in genetic LD backgrounds between populations of varying ancestries, however they also note that these studies are not always possible due to lack of data. Most of the regional plots constructed in this study drawn using r^2 (proxy for LD) and recombination rates from African populations illustrated the presence of areas of increased recombination activity near the top association signals observed. It is tempting to postulate that the same SNPs that are present in European populations are not being replicated because the LD blocks may have been broken up into smaller haplotype blocks in African populations due to recombination (decay in LD). Due to the smaller African LD blocks, the tagging efficiency of the Metabochip array is decreased. However, it must be noted that the allele frequencies of these variants may be different between European and African populations and our study had a limited sample size. Other studies have observed the "dilution of effects" (Carlson et al., 2013) when the transferability of signals from European to African American populations or African American/European to indigenous African populations have been examined (Ng et al., 2012, Gong et al., 2013, Adeyemo et al., 2015). It has been postulated that besides the difference in sample sizes influencing the ability to detect smaller effects, differences in environment, especially diet and physical activity, may attenuate the relationship between causative SNPs and obesity (Adeyemo et al., 2015). This may be true for differences in signals observed in African American vs. African populations, where although they share related ancestries, the environment (diet, physical activity, etc.) is very different (Adeyemo et al., 2015). Also it is possible that variants that have been replicated in European populations have failed quality control criteria and have been discarded due to low MAFs in indigenous African populations.

Many previously reported associations with BMI were not replicated in this SSA cohort. This may partly be due to the low tagging efficiency of the Metabochip array in an African population given that the assay design is based on European LD structure. Varying allele frequencies of variants between European and African populations may have contributed to lower statistical power to detect associations in this study. The authors of the MalariaGEN consortium stated that the lack of replication of GWAS signals in African populations questions the validity of previously reported associations, stating that real associations may fail to replicate due to overestimation of effect size ('winner's curse'), variation in frequency of effect allele between populations, variations in LD between index SNP and causative SNP and the overall complexity of the disease (allelic heterogeneity or epistasis) (Jallow et al., 2009). This is particularly relevant for this study where only locus-wide, rather than SNP-SNP interactions, were replicated. Other studies have also noted the attenuation of signals due to low LD when replicating European signals in African and other continental populations

(Jallow et al., 2009, Buyske et al., 2012, Carlson et al., 2013, Lu and Loos, 2013, Adeyemo et al., 2015).

The lack of statistical power to detect genome-wide associations due to small effects and small sample size is acknowledged in this study. This combined dataset (N=1926) might not have been sufficient to detect small effects for BMI as have been previously reported (Speliotes et al., 2010). Large, meta-analyses are needed to detect the small effect sizes from surrogate measures of adiposity such as BMI. The latest GIANT consortium meta-analyses have combined sample sizes of 339 224 individuals in which 97 BMI-associated loci (GW significance) were discovered and replicated, 56 of which, were novel. Only five of the 97 loci demonstrated strong evidence of several independent association signals, and also highlighted that these BMI-associated loci had significant effects on other metabolic phenotypes (Locke et al., 2015). This idea of cross-phenotype associations was recently echoed in a GWAS meta-analysis of percentage body fat (Lu et al., 2016).

4.7 BMI is not a good indicator of obesity

No significant association signals for BMI were observed in the larger adult-based Metabochip study. However three BMI-associated variants in the smaller adolescent subset of the Bt20 study (see Chapter 2) were replicated in a SNP-SNP candidate gene approach. This is possibly due to these genetic variants having a greater influence on BMI earlier in life, than later where environmental influences have more of an effect. The associations may also not have been replicated in the adult study because of the difference in approach (candidate gene versus hypothesis free approach) together with the difference in the number of markers, with six variants tested in the adolescent study as opposed to the ~200 000 SNP markers on the Metabochip. Therefore the conditions to satisfy statistical significance is much lower in a candidate gene approach rather than a hypothesis free approach. Also, most of the SNPs tested in the candidate gene study, failed QC in the Metabochip study.

Most GWAS for obesity-risk variants have focussed on BMI as a measure of obesity, however BMI is not the best measure of adiposity as it does not discriminate between fat mass and fat free mass (Tan et al., 2014). More accurate measures of body adiposity have been suggested including percentage fat mass and total fat mass (Comuzzie and Allison, 1998, Tan et al., 2014). This study focussed on the genetic association between SNPs captured on the Metabochip and a number of measures of adiposity including BMI, WC, HC, WHR, fat mass, fat free mass (excluding BMC) and PFM.

Other studies have shown that estimates of body composition by DXA may be better indicators of obesity (Comuzzie et al., 2012, Tan et al., 2014). It is well known that BMI (determined from twin and family studies) exhibits high heritability estimates ranging from 40-70% in European populations as mentioned in the literature review, but the heritability of other body composition phenotypes are less well described. A study of 554 participants (492 European and 48 African-American sib pairs) that assessed the heritability of body composition measures using DXA showed that fat mass, fat-free mass and PFM (0.71, 0,60 and 0.64, respectively) are highly heritable traits (Hsu et al., 2005). This suggests that these phenotypes have a strong genetic component and are well suited for genetic association studies.

4.8 Age- and sex-specific signals

There have been some suggestive sex- and age-specific associations in this dataset. Variants in or near *TRPM7, SLC17A7, COBLL1* and *NEGR1* were associated with anthropometry in males only, while variants in or near *CNTNAP5* and *BDNFOS* were associated with females only (combined young and old). Variants in or near *TMEM18* were associated with young females only whilst variants in or near *LPAL2* were associated with older female caregivers only. Also when comparing older female caregivers to young adults (mixed sex), variants in or near *ZFYVE9*, *WARS2* and *NBEAL1* were only associated with young adults of mixed sex.

An important consideration is that certain phenotypes may be modulated by variations in age and sex more than others. The GIANT meta-analyses have shown that there were no significant sex-specific and SNP-BMI associations (Winkler et al., 2015) while phenotypes associated with the distribution of fat like waist-to-hip ratio, waist circumferences and hip circumferences, displayed sex- and age-specific SNP associations in European populations (Heid et al., 2010, Randall et al., 2013, Shungin et al., 2015, Winkler et al., 2015). In this study age- and sex-specific SNP effects were observed for fat mass, PFM and WHR. Very few genetic associations were observed for BMI, so the existence of age- and sex-specific SNP associations with BMI could not be properly assessed.

5. Conclusion

In summary, 17 loci associated with either increased anthropometric measures (BMI, WC, HC, WHR) or increased DXA-derived body composition measures (fat mass, PFM), were observed in this study. Some of the signals that are reported in this study, have not previously been associated with obesity elsewhere. In addition, several cross-phenotype associations were observed in this study, where variants previously associated with BMI were also associated with other body fat-associated phenotypes. This study also demonstrated that more specific measures of body fat, such as those generated by DXA produced stronger SNP associations than did composite variables like BMI.

Age- and sex-specific effects of genetic variants on obesity were highlighted in the recent GIANT meta-analyses of BMI, body fat distribution and percentage body fat. It was also noted in Chapter 2 and in the Metabochip study that some of the genetic variants associated with obesity measures have stronger associations earlier on in life or in a particular sex. However, these correlations remain to be confirmed. Larger, longitudinal studies of body composition traits and genetic variants, and taking into account sex and age are necessary to better assess these sex- and age-specific gene effects.

GWAS have led to the discovery of potentially functional intronic non-coding variants. The question of how these intronic/intergenic variants contribute to biological mechanisms resulting in diseases is a major challenge. Recent endeavours such as the Encyclopedia of DNA elements (ENCODE) project (Consortium, 2012) aim to elucidate the biological relevance of these variants together with others. As illustrated by the example of *FTO,* polymorphisms in intronic regions can have “long reaching” effects on genes, and play a role in regulating gene expression. Intronic variants have also been shown to contain regulators of alternative splicing and other regulatory elements entrenched within these regions (Cooper, 2010). Caution has to be exercised with regards to the direct functional relevance of the SNP being captured by GWAS as opposed to being in LD with a ‘yet to be discovered’ causal SNP (Cooper, 2010).

This study has demonstrated the value of:

- Using more accurate measures of adiposity, such as DXA-derived fat mass in eliciting suggestive association signals even when optimum sample sizes are not achieved.
- Using an African population and capitalising on LD decay to narrow-down signals within loci to possibly hone in on causal variants or African-specific variants.

This study contributes important information to what we currently understand about the genetic contributions to body composition in an African population, and suggests that a GWAS using a larger cohort and with an African-specific SNP array may uncover further novel, causal gene variants.

Chapter 4

Heritability estimates (h^2) derived from Metabochip data for body composition phenotypes

1. Introduction

The study described in the previous chapter involved the genome-wide analysis of SNPs from a chip array and their association with anthropometric variables within our study population. Recent advances in the downstream use of such array data has led to the development of mathematical models that allow chip arrays to be used for measuring the level of heritability of various phenotypes (Yang et al., 2011a). The level of heritability of anthropometric variables, and particularly fat mass, has not been studied in African populations. Therefore the current chapter will describe the use of the Metabochip array for assessing the heritability of various anthropometric measurements in our study population.

The familial heritability of complex diseases such as obesity demonstrates that genetic factors do play a role in their development. An estimation of heritability tells us about the proportion of variation of a phenotype that is due to genetic factors. According to Visscher (2008) heritability is more formally defined as the proportion of total variance in a population (V_P) for a given trait that is due to genetic variance. There are two types of heritability, narrow-sense (h^2) and broad-sense (H^2). Narrow-sense heritability is phenotypic variation resulting from additive genetic effects, whilst broad-sense heritability is phenotypic variation that is due to a combination of additive, dominant and epistatic genetic effects. These three genetic effects are defined thus:

- *Additive* - sum of the contribution of allelic variation over many SNPs
- *Dominance* - is non-additive and is characterised by the heterozygotic phenotype not being the midpoint between the 2 different homozygotic phenotypes (Van Asselt et al., 2006).
- *Epistatic*- arises due to non-additive effects and is the result of genetic interactions between alleles at different loci (E-Resource2).

Calculations of heritability include only additive effects i.e. h^2 because the non-additive effects described above do not normally explain phenotypic similarities between family members.

A recent meta-analysis reporting on variance components for almost 18 000 traits and including almost ~14,5 million partly-dependent twin pairs, indicated that the heritability for most traits (69%) showed mostly additive genetic effects and that almost half of all the traits cluster around functional domains (Polderman et al., 2015), implying that the heritability of most traits are not random and can mostly be explained by narrow-sense heritability. This study included global populations but mentioned that African studies were clearly underrepresented. Recent GWAS have shown that several genetic variants play a role in complex traits, each with small effects, but that the additive effect of those variants is much smaller than the reported heritability for that trait (Visscher, 2008, Manolio et al., 2009, Yang et al., 2010, Hemani et al., 2013, Polderman et al., 2015). This issue of the missing heritability (discussed in the literature review) is a contentious one with some believing that missing heritability is simply not being captured with current tools while others claim that it lies in non-additive variation, inflated heritability estimates or parent-of-origin effects (Zaitlen et al., 2013).

Genetic markers captured on GWAS arrays are not only valuable in finding links between SNPs and biological mechanisms underlying diseases, but also offer the opportunity to estimate heritability for traits of interest (Visscher, 2008). Identity-by descent (IBD) describes the phenomenon whereby related individuals have a higher probability of sharing identical genotypes at polymorphic loci when compared to unrelated individuals and due to this show more likeness for particular phenotypes (Thompson, 2013). Therefore relatedness can be inferred through IBD by assessing pair-wise comparisons of genetic markers, although many markers are required. The inferred relatedness is then correlated with similarities in phenotypes and estimates of heritability can then be determined (Visscher et al., 2010). An important study by Yang et al. (2010) proposed the use of the total amount of phenotypic variance captured by *all SNPs* on commercial SNP arrays and using this to estimate heritability. Yang and colleagues (2011a) later developed a method to assess heritability from SNP arrays termed Genome-Wide Complex Trait Analysis (GCTA).

Missing heritability can be described as the difference between h^2 and the portion of phenotypic variance accounted for by *genome-wide significant* GWAS SNPs (h^2_{gwas}) (Zaitlen et al., 2013). The method described by Yang et al. (2011a) aims to explain the contribution of *all SNPs* (genotyped SNPs and SNPs in LD with genotyped SNPs) on the array to the variance (termed h^2_g), including even those SNPs that are not significantly associated with the trait, and in this way account for more of the missing heritability. If related individuals are included in the analyses the estimate of the genetic variance could be biased by shared environmental effects and thus be inflated (Yang et al., 2011a). However removal of related

individuals also results in a loss of sample size, which results in large standard errors (Zaitlen et al., 2013). An approach to overcome this issue has been proposed Zaitlen et al. (2013) where the method allowed h^2 to be estimated with the inclusion of both closely and distantly related individuals. Further, the approach using a new variance component method produces both estimates of h^2 and h^2_g, so they are directly comparable within the same population (Zaitlen et al., 2013).

Heritability estimates have been reported for body composition traits (summarised in Table 4.1), including height (0.40-0.80), weight (0.52-0.59), BMI (twin studies 0.40-0.90; 0.24-0.81 for family studies), WC (0.31-0.76), HC (0.41-0.45) and WHR (0.31-0.76), showing high levels of heritability (Rose et al., 1998, Elks et al., 2012, Murrin et al., 2012, Nan et al., 2012, Liu et al., 2013, Randall et al., 2013, Locke et al., 2015, Shungin et al., 2015, Sung et al., 2015, Winkler et al., 2015). Heritability estimates for body composition traits measured by DXA, CT or bioelectrical impedance, are less common with estimates ranging from 0.47-0.64 for PFM, 0.48-0.71 for fat mass and 0.49-0.72 for lean mass (Luke et al., 2001, Hsu et al., 2005), and again show a high genetic contribution to the variance of these traits. These estimates include reports across ancestries with several studies reporting on sex differences in estimates of heritability for WC and WHR (mostly adjusted for BMI), with larger measures of heritability reported in females over males (Heid et al., 2010, Liu et al., 2013, Randall et al., 2013, Shungin et al., 2015, Sung et al., 2015, Winkler et al., 2015).

Since heritability estimates for traits relating to body composition are unknown in SSA populations, and cannot be extrapolated from other populations we decided to use the Metabochip genotype data in an attempt to capture narrow-sense heritability or h^2. Metabochip data can be used to specifically ask, how much of the phenotypic variance for a single trait or traits can be explained by all the SNPs captured on the Metabochip (h^2). We used GCTA as adapted by Zaitlen et al. (2013) for use in our dataset.

2. Materials and Methods

GCTA[3] vs.1.24 together with R vs.3.2.2 (Development Core Team, 2008), were used for the analysis. GCTA software estimates a relatedness matrix based on pairwise genetic

[3] Any further mention of GCTA will refer to GCTA vs.1.24 Yang, J., Lee, S. H., Goddard, M. E.&Visscher, P. M. (2011a). GCTA: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics,* 88**,** 76-82.

covariance. In GCTA, the estimation of the proportion of variance is captured by the SNPs on the Metabochip. This is achieved by fitting the various body composition traits with a linear mixed model using the relatedness matrix as a variance term. All of the analysis was conducted on the Wits cluster (www.cream-ce.core.wits.ac.za), as the analyses were computationally intensive.

Narrow sense-heritability estimates were generated for weight, height, BMI, WC, HC, WHR, fat mass, fat-free mass and PFM using Metabochip data for different scenarios. We calculated heritability estimates where 'all individuals' refers to the relationships from all pairs of individuals. In this scenario these estimates are neither an indication of h^2 or h^2_g. With regards to 'related individuals', in this scenario only those individuals with kinship coefficients >0.05 are used and is an estimation of h^2. Lastly, for 'unrelated individuals', the dataset was pruned to only include individuals that had kinship values of <0.05 and is an estimate of h^2_g. With the Zaitlen method the full relatedness matrix using all individuals was fitted with two variance components, one that resulted in an estimate of h^2 and one that resulted in an estimate of h^2_g, with both these components being estimated simultaneously (Zaitlen et al., 2013).

The scripts required for the analysis were modified and adapted from Zaitlen et al. (2013) with assistance from G. Hemani (Bristol University, UK). Fig. 4.1 was drawn using R vs.3.2.2 with assistance from G. Hemani. The final scripts required to generate estimates of h^2 were optimised and automated for all the phenotypes in the combined dataset, to generate a single output file. Bearing in mind that GCTA requires a large dataset for the analysis to reduce the standard error, we used the combined dataset containing both related and unrelated individuals. All scripts used in the analyses are shown in Appendix F.

2.1 Merging, basic QC, and generation of GRMs

1. The merged dataset together with the phenotype file were used as input files to conduct the SNP heritability analysis. The merging of the datasets (caregivers and young adults) were described in detail in Chapter 3 and Appendix C.
2. The merged dataset (prunedbmerge2.*) was further subjected to QC (by removing markers with MAF <0.01, markers showing significant deviations from HWE $P < 1$ x

10^{-6}, as well as markers missing more than 1% of their genotypes) to ensure a clean dataset, bearing in mind that QC prior to this was conducted on individual batches and not the merged dataset. Relatedness based on IBS was assessed within the merged dataset.

3. Following the removal of SNPs that did not meet QC requirements - a cleaned merged file was generated (clean.*).
4. The next step involved the generation of a GRM in GCTA that consisted of all related and unrelated individuals ('all individuals' using clean.*) and the following commands:

```
gcta64 \
--bfile ${plinkfile} \
--make-grm \
--maf 0.01 \
--out ${allfile}
```

where `plinkfile = clean.*; allfile` = the output file of GRM1.

The possibility of any duplicates in the dataset upon merging were assessed using the script *`find_duplicates.R`* where IBS cut-offs > 0.80 denoted duplicate individuals.

5. Possible duplicates were removed in PLINK with the `--remove` command together with the `duplicate_geno.txt` file, which generated a new cleaned file without duplicates `(clean2.*).`
6. A new GRM was generated without any duplicates.
7. The GRM generated was adjusted for 10 principal components (PCs), to account for any possible substructure in the dataset.
8. The R-script *`remove_unrelateds.R`* was run to create GRMs for 'related individuals', 'unrelated individuals'.

Steps 1-8 were automated using the script *`geno_data.sh`*.

2.2 Generation of phenotype files

A phenotype file generation script was modified to include all phenotypes including height, weight, BMI, WC, HC, WHR, fat mass, lean mass and PFM, together with the covariates PCs, age and sex, in the R-script, *`pheno_data_all.R`*.

2.3 Running and automating the analysis in GCTA

Once the appropriate data files were generated, the analyses were run in GCTA using the modified script run_allpheno_analysis.sh, which uses a restricted maximum likelihood (REML) algorithm to perform regression, assuming additive inheritance. The analyses were adjusted for covariates, PCs, age and sex (generated in the previous step). Each phenotype was analysed where, 1) all individuals were considered, 2) only related individuals were considered, 3) only unrelated individuals were considered, 4) combining GRMs from all individuals and related individuals (Zaitlen method).

The comparative h^2 results from Ngcungcu (2013) were calculated using the ASSOC option within Statistical Analysis for Genetic Epidemiology (SAGE) vs.6.01 assuming an additive model of inheritance and were adjusted for the necessary confounders described in Table 4.1. The African Programme on Genes in Hypertension cohort (APOGH) cohort is made up of random black South African nuclear families that self identify as being Nguni or Sotho speakers (Ngcungcu, 2013). The family data includes one or more children (>16 years) together with one or both parents if available.

3. Results

Further QC was performed when the datasets were merged where duplicate individuals (N=4), together with additional SNPs (N=390), were removed, resulting in N=125 388 SNPs and N=1923 individuals available for analysis in the combined dataset. Fig. 4.1 illustrates the distribution of relationships between unrelated (Fig. 4.1a) and related (Fig. 4.1b) individuals, where 'Unrelateds' refers to all pairs of individuals with kinships ≤ 0.05 and 'Relateds' refers to all pairs of individuals with kinships ≥ 0.05. The figure shows a clear clustering of individuals around 0.5, indicating parent-child or full-sibling relationships, together with other relationships distributed from 0.05-0.3 in the related individuals (Fig. 4.1b). As previously described in Chapter 3, kinship scores (derived from IBS) of ≥ 0.80 refer to duplicates or monozygotic twins, ≥ 0.5 implies first-degree relatives, ≥0.25 implies second-degree relatives and ≥ 0.125 third degree relatives. This clustering observed in Fig. 4.1b is suggestive of the study design that includes mother-child pairs, together with other cryptic relationships that exist within the dataset.

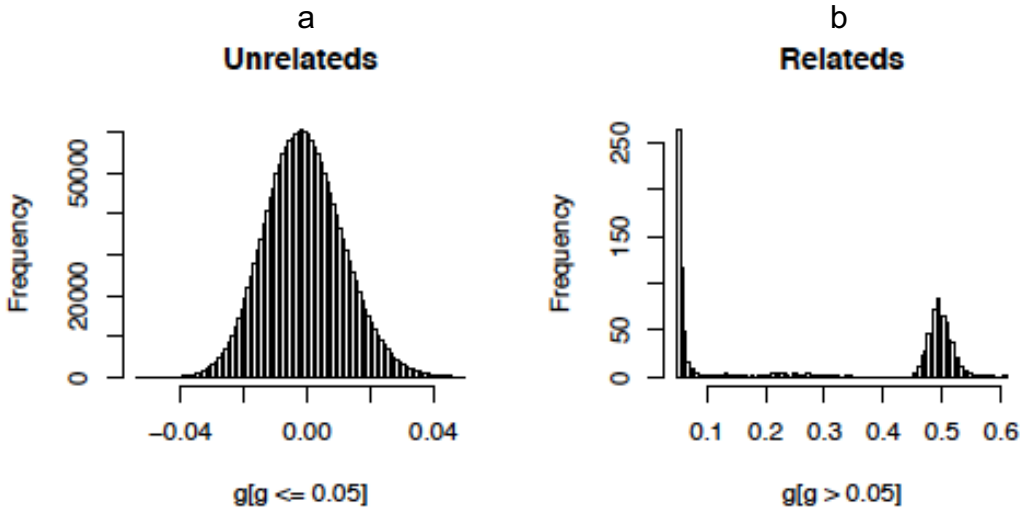


Figure 4.1 Graph showing the frequency (y-axis) versus kinship values (x-axis) illustrating the distribution of kinship scores in unrelated (a) and related individuals (b).

Results of the narrow sense heritability estimates (h^2) are presented in Table 4.1. These estimates and the GRMs were calculated for four different scenarios to better understand the relatedness in the dataset. In Table 4.1 ALL samples, refers to the relationships from all individuals, thus reflecting kinships containing a heterogeneous mix of related and unrelated individuals. RELATED, uses all individuals, but sets everyone with a kinship value of <0.05 to that of 0. This ensures that the REML algorithm then only seeks to maximise the variance between different relatedness groups, similar to that of traditional pedigree analysis. UNRELATED, uses a dataset that was pruned to only include individuals that have kinship values of <0.05. This analysis reduces the dataset by almost half. The ZAITLEN method fits two GRM's simultaneously: the GRM generated from *all samples* and the GRM generated for *related samples only*. This algorithm attempts to partition the genetic variance into SNP heritability in one component (h^2_g) and the remaining variance that is due to genetic effects not captured by SNPs together with common environment (h^2).

Table 4.1 Narrow sense heritability estimates (h^2) for different analysis strategies together with published data

	All[#]	N	Relateds[#]	N	Unrelateds[#]	N	Zaitlen Method[#]	N	SAGE Estimates	Published estimates
	$h^2+h^2_g$ (SE)		h^2 (SE)		h^2_g (SE)		h^2_g / h^2 (SE)		[§]h^2 (SE)	[$]$h^2_{published}$
Height (mm)	0.50 (0.05)	1921	0.84 (0.06)	1921	1.30×10^{-1} (0.1)	1109	0.83 (0.06)	1921	0.76 (0.07)*	0.40-0.80
Weight (kg)	0.27 (0.05)	1919	0.56 (0.06)	1919	1.00×10^{-5} (0.1)	1107	0.57 (0.06)	1919	0.38 (0.09)**	0.52-0.59
BMI ($kg.m^{-2}$)	0.23 (0.05)	1919	0.48 (0.07)	1919	2.00×10^{-6} (0.1)	1107	0.40 (0.07)	1919	0.26 (0.01)**	0.24-0.90
Waist circumference (cm)	0.20 (0.05)	1898	0.40 (0.07)	1898	1.00×10^{-5} (0.1)	1096	0.40 (0.07)	1897	0.35 (0.10)**	0.31-0.76
Hip circumference (cm)	0.26 (0.05)	1899	0.52 (0.07)	1899	2.00×10^{-6} (0.1)	1097	0.52 (0.07)	1899	0.42 (0.09)**	0.41-0.45
WHR	0.10 (0.05)	1897	0.17 (0.07)	1897	1.00×10^{-5} (0.1)	1096	0.18 (0.07)	1897	0.46 (0.09)**	0.31-0.76
Fat mass (kg)	0.15 (0.06)	1711	0.52 (0.08)	1711	2.00×10^{-6} (0.1)	993	0.58 (0.07)	1711	did not assess	0.48-0.71
Lean mass (kg)	0.47 (0.06)	1711	0.79 (0.07)	1711	1.00×10^{-2} (0.1)	993	0.79 (0.07)	1711	did not assess	0.49-0.72
PFM	0.18 (0.06)	1711	0.49 (0.07)	1711	4.00×10^{-6} (0.1)	993	0.54 (0.07)	1711	did not assess	0.47-0.64

h^2 – is equivalent to the heritability estimated from IBS according to a certain threshold, h^2_g- heritability explained by genotyping SNPs and h^2_g / h^2 is the heritability explained by the Zaitlen method, where the method allows for both variance components to be fitted simultaneously; [#]Adjusted for first 10 principal components, age and sex.

[§]h^{2-} Heritability data from MSc. Thesis (Ngcungcu, 2013) *adjusted for age and sex; **adjusted for age, sex, smoking and drinking; calculated using SAGE (Statistical Analysis for Genetic Epidemiology) software vs.6.01 in a black Sowetan population.

[$]$h^2$ estimates reported from published pedigree, family-based and population studies where pedigree studies = twin design, full-siblings and siblings and family-based and population studies = where SNPs are used to construct GRMs and h^2 is based on the additive model (Rose et al., 1998, Luke et al., 2001, Hsu et al., 2005, Elks et al., 2012, Murrin et al., 2012, Nan et al., 2012, Liu et al., 2013, Randall et al., 2013, Locke et al., 2015, Shungin et al., 2015).

With the exception of (Luke et al., 2001) all of the other estimates have been derived from populations of European descent

The results in Table 4.1 show that the Zaitlen method (h^2_g/h^2) reflects the upper bound of the heritability estimates with h^2_g/h^2 ranging from 0.18 for WHR - 0.83 for height. These values are very similar to the estimates when only considering related individuals (h^2 = 0.17-0.84). The highest heritability estimates from related individuals were observed for height and lean mass, (h^2 = 0.84 and 0.79, respectively). The heritability estimates from related individuals are most similar to traditional pedigree analysis estimates, which are comparable with estimates observed in European populations for most traits, with the exception of WHR. Very low estimates of heritability were estimated for h^2_g, using the unrelated individuals for all traits. When all samples were considered, irrespective of the relatedness, the heritability estimates reflects an average between heritability estimates generated separately for related and unrelated individuals.

The comparative h^2 results from Ngcungcu (2013) showed that the heritability estimates ranged from 0.26 for BMI to 0.76 for height. These heritability estimates were based on the variances in phenotype using the APOGH, living in metropolitan Johannesburg (Ngcungcu, 2013). The heritability estimates shown for this study are lower than the heritability estimates from our dataset using the Zaitlen method with the exception of WHR.

4. Discussion

In this part of the study GCTA was used in conjunction with SNP data from Metabochip to measure the narrow-sense heritability (h^2) of body composition traits including BMI, weight, height, WC and HC. We also focussed on DXA-derived fat mass, lean mass and PFM. GCTA was used to estimate additive genetic effect of fine-mapping and replication SNPs captured on this array. Most GWAS association studies have been carried out on measures of body size (BMI) and body fat distribution (WC and WHR) with the latest GWAS meta-analyses showing that only 2,7% and 1,36% of the variation in BMI and body fat distribution is explained by the identified SNPs (Sandholt et al., 2015).

The method proposed by Yang et al. (2010) for assessing heritability using GWAS arrays requires data from unrelated individuals. Our dataset comprised a large number of mother-child pairs together with some unrelated individuals. The method proposed by Zaitlen et al. (2013) which fits two variance components at the same time, allowed for the use of both closely and distantly related individuals and was therefore used to accommodate our experimental design.

Heritability was estimated using different scenarios as described earlier, with the analysis of unrelated individuals (h^2_g) most closely resembling the approach used by Yang et al. (2010) The heritability estimates calculated in this scenario was very low for all traits when compared to estimates of heritability captured by published data. Published h^2_g estimates from GWAS arrays for BMI, height and WHR were reported as 0.16, 0.42 and 0.13, respectively (Yang et al., 2011b, Vattikuti et al., 2012) whereas Zaitlen and colleagues reported equivalent h^2_g of 0.23, 0.40 and 0.14 for BMI, height and WHR, respectively (Zaitlen et al., 2013). There was no available published h^2_g for the other traits. A possible reason for this low heritability is that the current study was underpowered in terms of the number of SNPs (~125 000 markers on a fine-mapping and replication array) contributing to the variances captured on the array. Yang and co-workers (2010) used ~295 000 GWAS SNPs while Zailten and colleagues (2013) used about ~ 300 000 SNPs, each contributing small effects to variance of phenotypes tested. It is also possible that SNPs captured by Metabochip only contribute a small fraction to the phenotypic variance of body composition traits in African populations. In other words, the Metabochip array is designed to reflect variation based on LD patterns in European populations and might not accurately tag variants in African populations because of the lower LD. Yang and colleagues have stated that that the missing heritability not captured by GWAS SNPs regarding height can be explained by incomplete linkage between SNPs and the causal variants (Yang et al., 2010).

Although Zaitlen et al. (2013) reported high heritabilities for the 23 traits examined in their study, their estimates were much lower than most estimates captured from traditional twin or family studies. The estimates reported were 0.54, 0.58 and 0.74 for BMI, height and WHR, respectively. The estimates for our study using the Zaitlen approach ranged from 0.18 (WHR) -0.83 (height) and were mostly within the range given in the literature from twin or family studies, with the exception of WHR. The estimates obtained using the Zailten approach were very similar to the estimates achieved when only related individuals were used in the analysis. Other studies cited in (Zaitlen et al., 2013) have shown that using the approach described by Yang et al. (2011a) results in overestimates of h^2 when using related individuals, which is reflected in our estimates when only related individuals are considered. This is why we employed the Zaitlen approach, but still achieved the upper bound of heritability estimates. Possible reasons include the unaccounted for epistatic interactions in related individuals leading to overinflated heritability estimates which has been described previously (Falconer and Mackay, 1995), or due to ethnic differences in body fat distirbution and limitations of current dataset.

Using the Zaitlen approach, estimates of heritability (h^2_g/h^2) reported in the current study for BMI (0.40) and WHR (0.18) were lower than the estimates reported by Zaitlen et al. (2013) for BMI (0.54) and WHR (0.75). This could be due to differences within and between populations contributing to the variances in phenotypes as well as difference in tools (GWAS vs. Metabochip) used to estimate the heritability. Similar, although higher estimates of heritability were reported in our study as compared to h^2 estimates from family-studies reported in Ngcungcu (2013) with exception of WHR. Despite the inconsistencies, the APOGH study together with the current study remain the only estimates of heritability for body composition available for SSA populations.

Very low estimates of heritability for WHR were reported in this study for (h^2_g/h^2) 0.18, (h^2) 0.17, (h^2_g) $1x10^{-5}$. Zaitlen et al. (2013) reported estimates of (h^2_g/h^2) 0.75, (h^2) 0.19 and (h^2_g) 0.14 for WHR. The estimates reported by Zaitlen et al. (2013) for WHR using related individuals were similar to the estimates reported in this study. Vatikutti and colleagues (2012) also reported low estimates (h^2_g) of WHR (0.13). These low and varying estimates of WHR are most likely due to WHR not being an accurate measure of body fat distribution and that each component, WC or HC independently, have separate factors including genetic and environmental contributing to them.

The accuracy of heritability estimates is reflected by the standard error, with a low standard error being optimal. The standard error is influenced mainly by sample size and confounding variables. However Visscher and Goddard (2015) have stated that sometimes even with very large sample sizes, large standard errors are observed, which may be a reflection of the experimental design. The estimates reported in our study have large standard errors (especially in the unrelated individuals) because of the small sample size, as GCTA requires sample sizes of 3000 or more (Zaitlen et al., 2013).

A recent publication by Kumar et al. (2016) has postulated that GCTA is flawed in its ability to assess heritability because of two main reasons:

- Heritability estimates produced by GCTA are unreliable because of the inaccurate estimation of genetic relationship matrices due to the fact that the number of SNPs used exceeds the sample size for the phenotype data.

- That GCTA doesnot properly control for population stratification, which results in unreliable estimates of heritability. They postulate that GCTA 'over fits' normal GWAS and results in higher heritability estimates.

However, despite these criticisms the methods used to determine heritability in the current study give values that are comparable to the literature. Furthermore, population structure was taken into account in our study by including 10 PCs as a fixed effect in the linear analysis as well as removing outliers during the individual batch QC process. It should be noted that the sample size used in this study was small and this may affect the accuracy of our estimates of heritability. It is recommended that N>3000 should be used when using the GCTA methodology (Yang et al., 2011a, Zaitlen et al., 2013). Further investigations should be performed using a larger sample size to confirm these findings.

5. Conclusion

As with most studies surrounding heritability estimates, one has to be cautious in its interpretation. This study shows that within the Bt20 cohort, there is a substantial genetic contribution to the variation in traits related to body composition as captured by the Metabochip. The captured heritability is similar to those reported for European populations and warrants further investigations using larger cohorts of unrelated (so contamination from shared common environment can be eliminated) individuals and using a SNP array that captures common variation and the appropriate LD structure for African populations. The main strength of this study is that it provided the opportunity to estimate narrow-sense heritability for body composition traits in a population for which heritability estimates are unavailable.

Chapter 5

Concluding remarks

5.1 Rationale of the thesis

This study aimed to expand our current knowledge regarding the genetic contribution to obesity risk in a sub-Saharan African population by:

1. Using phenotypes describing various aspects of body composition, in particular DXA-derived measures and their association with genetic markers.
2. Estimating the narrow-sense heritability of phenotypes contributing to body composition using Metabochip data, and to assess if these estimates were similar to those observed in other populations.
3. Determining whether variants associated with increased obesity risk in European populations, played a similar role in a South African black population.

Obesity rates are increasing not only globally but also within the African continent, with obesity being more common in females than males (Dalal et al., 2011, Ng et al., 2014). This was observed within this study with a median BMIs of ~30kg/m^2 in older females, and a higher BMI measure observed in young females compared to their young male counterparts.

Heritability estimates of traits related to obesity (BMI, WC, WHR, fat mass, PFM) have shown a strong genetic contribution across European populations. Estimates of heritability allow a comparison of the importance of genetic and non-genetic factors contributing to a trait but because it is a ratio it cannot inform the size of the component (Visscher, 2008). Genetic components imply susceptibility of an individual to developing obesity (in this case) in a given environment, and identifying these genetic factors has been the focus of many studies (Anderson et al., 2015).

However a majority of what we currently know about genetic factors linked to obesity is focussed on European and some Asian studies, with African populations remaining understudied. There are advantages in studying African populations, in that the high genetic diversity and low linkage disequilibrium allows for narrowing down of loci where signals have been identified (Teo et al., 2010). This could potentially unveil causal variants. The main motivation for this study is the lack of genetic studies, either association or heritability estimates, of traits contributing to obesity in diverse sub-Saharan African populations where obesity is a considerable problem.

5.2 Summary of findings

The overall aim of the study was to assess whether genetic variants contributing to obesity risk, captured in European populations were the same in African populations. Both SNP-SNP as well as obesity-susceptibility loci replications were reported within the Bt20 cohort. Three out of six obesity risk variants in or near, *TMEM18* (rs6548238), *SH2B1* (rs7498665) and *GNPDA2* (rs10938397), were associated with increased BMI using a candidate gene approach, in an adolescent subset of the Bt20 cohort. It was suggested that these variants might be correlated with the regulation of body size in African adolescents (Pillay et al., 2015).

Most studies have focussed on using BMI as an indicator of obesity risk together with surrogate measures of fat distribution, including WC, HC, WHR, and more recently WHR_{adjBMI}. This measure has been used to assess the effects of fat distribution that is separate from overall adiposity (Andersen and Sandholt, 2015). In this study more precise measures of body composition were used to conduct fine-mapping and replication using the Metabochip. These measures included DXA-derived total fat mass, total lean mass and PFM together with the conventional surrogates for obesity risk (BMI, WC, HC and WHR).

Obesity-susceptibility loci (with novel variants rather than previously associated SNPs) were replicated in or near *SEC16B*, *NEGR1*, *FTO*, *TMEM18*, *WARS2, NRXN3, and SP110* together with new signals. These associations were reported with fat mass, PFM and WHR. Cross-phenotype associations for anthropometric and body composition traits were observed for four of the five of the replicated loci perhaps indicating a common biological pathway. These loci have previously been associated with increases in BMI or measures of fat distribution. Associations were reported between *SEC16B, NEGR1* and *TMEM18* with increased fat mass and PFM while *FTO* and *WARS2* were associated with increased WHR. The strongest signals (supported by more than one variant) observed in this study were with variants in or near *SEC16B* with fat mass and PFM.

The narrow-sense heritability estimates calculated in this study demonstrated that within the Bt20 cohort, there was a substantial genetic contribution (ranging from 0,10-0,50) to the variation in body composition traits, captured by the Metabochip, but that the effects of shared common environment amongst individuals cannot be excluded.

5.3 Implications of findings

In this study some obesity-risk loci that are common to European populations have been replicated, together with the same SNP (as with the candidate gene study in Chapter 2), however some of the variants within other replicated loci are novel to this study (described in Chapter 3). This is an example of the genetic heterogeneity that characterises complex disease phenotypes (Rosenberg et al., 2010) that may only be identified when conducting fine-mapping in a population other than the discovery population. Obesity risk loci that are common to populations suggest common disease aetiologies between populations, however the presence of novel associations within populations indicates that population-specific variants contribute to the disease process (Fu et al., 2011). Such findings have been reported in populations of East Asian ancestry with traits such as T2D, systemic lupus erythematosus, ulcerative colitis and height (Fu et al., 2011). Such studies are important for genetic risk prediction, and can be advantageous in cross-ethnicity fine-mapping studies.

Conducting fine-mapping studies in populations of diverse ancestries such as African populations are beneficial due to the high genetic diversity and the low linkage disequilibrium. High genetic diversity and low LD can be exploited by fine-mapping to refine disease/trait - associated loci, and thus hone in on actual causal variants as well as identifying population-specific variants. Fine-mapping studies involve the genotyping of extra variants at identified GWAS loci that may result in the unveiling of SNPs that are more strongly associated with the trait than the index SNPs reported from GWAS (Wu et al., 2013). This can be used to further localise causal variants. The associations reported in this study have not been previously linked to obesity and are likely to be African-specific variants and might include causal variants. Replication in an independent SSA population using an independent genotyping platform would be necessary to confirm this. Other studies (Buyske et al., 2012, Musunuru et al., 2012, Wu et al., 2013) have demonstrated that for lipid-associated traits, fine-mapping in non-European populations resulted in the discovery of additional associations that may be rare or absent in discovery populations, usually European. A previous study identified more signals per locus in African American populations and more signals overall than in European populations, due to the greater diversity in non-European populations (Wu et al., 2013). Most trans-ethnic fine-mapping studies (Buyske et al., 2012, Musunuru et al., 2012, Wu et al., 2013) together with the current study have demonstrated the benefit of fine-mapping populations of varying ethnic backgrounds to identify population-specific variants as well as refining association signals.

"Allelic heterogeneity" is characterised by different variants at the same gene/locus having an effect on the same phenotypes (Wu et al., 2013) and applies to both monogenic and complex disorders. Studies have demonstrated the occurrence of allelic heterogeneity with complex traits like T2D (Voight et al., 2010, Sim et al., 2011), CAD (Peden and Farrall, 2011) and lipid-associated traits (Wu et al., 2013) in different populations. In this study the presence of possible independent signals at *SEC16B, NEGR1* and *FTO* were suggested (illustrated in regional plots) and further studies including deep-sequencing of these loci, are necessary to confirm these findings.

To date most obesity risk loci have been associated with variations in BMI, as previously mentioned. In this study we have shown possibly African-specific variants associated with more than one phenotype – cross-phenotype associations. Several overlaps were highlighted in this study with the same genetic variant(s) being associated with more than one phenotype, which may be suggestive of sharing common underlying pathways with highly related traits. Pleiotropy refers to the contribution of a gene (broad-sense) or an allele (narrow-sense) to more than one phenotype (Mackay et al., 2009), which is common to complex genetic disorders, but is generally defined in terms of "seemingly" unrelated clinical outcomes. A recent study by Lu et al. (2016) has demonstrated cross-phenotype associations with BMI and percentage body fat in eight loci including *SEC16B*. They also showed that genetic variants that have a larger effect on percentage body fat than BMI, are also associated with other cardiometabolic traits including fasting glucose, lipid levels, T2D and CAD. In this study, cross-phenotype associations with fat mass and PFM were observed more so with loci previously associated with BMI (an indicator of overall body size) rather than loci involved in fat distribution (WC, HC, WHR), with the exception of *FTO*. Variants in *FTO* have previously been shown to be associated with phenotypes other than BMI, in particular WHR (Heid et al., 2010, Vasan et al., 2013). This suggests that genetic variants contributing to overall adiposity are connected to a common biological pathway and are distinct to those variants contributing to fat distribution.

Variants in or near *SEC16B*, *NEGR1*, *TMEM18* and *FTO* previously associated with BMI were shown to be associated with fat mass, PFM and WHR in this current study. The use of more precise measures of adiposity, particularly fat mass and PFM, resulted in an increase in power (also reported in Lu, et al., 2016) to detect these correlations. It has been reported that neither BMI nor other anthropometric measurements such as WC or HC, are accurate predictors of body fatness in most cross-sectional studies (Baumgartner et al., 1995, Müller et al., 2010, Okorodudu et al., 2010, Jenkins and Campbell, 2014). Jenkins and Campbell (2014) have reported that large-scale genomic studies have focussed on measures that do

not adequately represent the degree of body fatness and that their on-going use adds to the current uncertainties about genetic contribution to obesity. Also, even though studies utilising direct measures of body fatness such as DXA have shown little benefit, this may only be due to the use of small sample sizes because of the inherent costs involved. The authors go on to mention that the gain in statistical power and interpretation possibilities with the use of more direct phenotypes to describe adiposity potentially outweigh the additional costs of collecting these phenotypes.

The narrow-sense heritability estimates for body composition show that these traits have high heritability estimates. This means that genetic factors have a large part to play in the variance of these phenotypes, however the possible contribution of shared environment cannot be excluded. Estimates of heritability allow us to ask new questions about the biology of traits and allow a comparison of the importance of genetic versus non-genetic factors in a particular surrounding (Visscher et al., 2008). A lot of heterogeneity is observed in studies estimating heritability. Heritability estimates for BMI range from 0.47-0.90 and 0.21-0.81 in twin and family studies, respectively. According to Elks et al. (2012) study design factors contribute almost 47% to the heterogeneity observed. The authors also postulate that variations in BMI are dependent on age of the cohort, with a greater influence in the heritability of BMI in childhood than later on in life.

5.4 Overall limitations of the study

There were some challenges in conducting a fine-mapping and replication study in an SSA setting some of which were addressed earlier, but include:

- **Statistical power and sample size calculations**

  Normally power calculations for general quantitative traits require summary statistics (mean and variance/standard deviation) of the trait/s. Due to the fact that in genetic studies effect sizes of quantifiable traits are traditionally assessed from heritability estimates, normal requirements for power calculations are not sufficient (Feng et al., 2011). This is because heritability is difficult to translate into summary statistics. The power calculations conducted in this study may not have been suitable for an African population due to the fact that heritability estimates and effect sizes were based on European estimates (Speliotes et al., 2011), however they were the only estimates available for use.

Sample size is an important factor in being able to determine statistically significant associations (Visscher et al., 2012). The lack of ability to detect most of the previously replicated obesity loci captured by Metabochip is mostly due to false negatives or the lack of statistical power that is due to sample size. Also, due to the fact that BMI is not a precise indicator of obesity, the effect sizes of SNPs that contribute to this heterogeneous trait are small, and therefore large sample sizes are required for these small-effect size variants to be statistically significant. The use of more precisely measured phenotypes (discussed earlier) associated with variants with larger effect sizes may negate the need for extremely large sample sizes (Lu et al., 2016), and may be the reason why associations with DXA-measured estimates of adiposity, were mainly replicated in this study.

- **Experimental design**

One of the challenges of this study was the use of related individuals. Although relatedness was accounted for by including a kinship matrix during both the association and heritability analysis, there may still have been a reduction in the power to detect associations or an effect of shared common environment, respectively.

Another challenge in this study was unequal proportion of males and females in the study, with males being underrepresented. We also had an unequal proportion of young and older female adults upon merging females only. Including older males in the study to match that of the female caregivers would have been advantageous to assess the effects of age and sex in the older adults more accurately, as these factors are important modifiers of body composition.

- **Metabochip as a tool to capture variation in African populations**

As previously highlighted, a large proportion of SNP data could not be used, as these variants are monomorphic, and therefore uninformative in this population. This is a major limitation of using arrays capturing common variation in European populations, in populations of diverse genetic ancestries. Despite this limitation, genetic associations were still reported, aided by the use of more precisely defined phenotypes.

The fact that the Metabochip is not a GWAS chip but rather a chip with genome-wide representation of closely linked markers, is a limitation for PCA. The use of unlinked markers that are independent of each other is a prerequisite for PCA. PCA may have been limited by the varying sample sizes of the population groups (HapMap and 1000 Genomes, BSO, SEB and SWB) used to construct the plots. Since the Bt20 dataset was

much larger than the other populations, it would have been ideal to draw a subset of the Bt20 to match that of the other 1000G datasets. Also the number of SNPs available after merging and pruning that were common to all datasets were based on 13500 and 12100 markers for the female caregivers and young adults, respectively and may not have been ideal to tease out genetic relations amongst individuals. Having noted the limitations of PCA, it is also important to report that PCA was not performed to elucidate the genetic variation among African populations, but rather to establish if any substructure existed in the dataset (based on PC1 and 2) and for those individuals to be removed, prior to performing the association analysis.

- **Use of Genomic Control (GC)**

During association analysis the Genomic Control (GC) was assessed in PLINK as a way to address substructure following QC. GC controls for confounding that may arise as a result of substructure by increasing the threshold required for statistical significance, thereby controlling the number of false positives (Cordell and Clayton, 2005). However drawbacks of this method include the fact that there is no control in place for false negatives, which leads to a loss in power to detect real associations. Results from the GC (as indication by the GI test statistic; lambda) showed no evidence of inflation of lambda that could be attributed to population stratification. Scores further away from one suggests that there is an over-inflation of the test statistic that can be attributed to false differences in allele frequencies, which may arise due to cryptic relatedness, population structure or errors during genotyping (Yang et al., 2011c). GI scores of close to one in this dataset for all phenotypes shows that the PCA carried out was effective in the identification of outliers, proving its usefulness even though limitations were present.

- **Lack of information for African populations**

Another challenge and limitation of this study and one that is due to the paucity of information on obesity risk variants on the African continent, is the need for replication of the association signals observed. In order for these correlations to be validated, they should be replicated in other African populations using independent technologies. Even though African-American or West African cohorts may be available for replication, there are differences in genetic substructure of SSA populations. These differences were illustrated in principle component analysis plots shown earlier in the results sections and reported in a project describing the variation of African populations on the subcontinent – African Genome Variation Project (AGVP), which must be taken into consideration (Gurdasani et al., 2015). Nevertheless, these populations remain the only options for

replication purposes however very few, if any studies have the matching detailed phenotypes (e.g. DXA-derived body composition) that were used in this current study.

The challenges that still remain that are relevant to GWAS in general, is the issue of missing heritability. With regard to BMI, the current 97 loci that have been associated with this trait only accounts for 2.7% of the variation in BMI (Locke et al., 2015). GWAS are not sufficient or efficient enough by virtue of their design in capturing rare variants that might confer larger effects, and whole genome sequencing is a better tool to capture these rare variants. Perhaps investigating the impact of dominance and epistasis on narrow-sense heritability estimates in closely related individuals may also contribute to this missing heritability (Zaitlen et al., 2013). Zaitlen et al. (2013) have reported inflated narrow-sense heritability estimates for T2D and CAD, when they included the effects of dominance and epistasis in regression models. Smaller effects could be captured by expanding sample sizes (hence a drive towards more consortia efforts), or implementing better study designs. There is logic in all these arguments, and future studies should draw on all these suggestions and include studies with phenotypes that have more biological relevance. Using precisely measured phenotypes may increase the ability to detect genetic variants conferring risk for obesity and capturing more of the heritability contributing to the phenotypic variance.

5.5 Future Work

The next step would be to replicate and perform a meta-analysis of the results from this study with other Metabochip and GWAS data in African populations. A causal variant is described by Visscher (Visscher et al.) as "……..an unknown variant that has a direct or indirect functional effect on disease risk, rather than a variant that is associated with disease risk through LD - it is the variant that causes the observed association signal". The strategies for possible ways of identifying causal genetic variants in general are challenging and have been summarised in Fig. 5.1 (adapted from Kumar et al., 2012). These strategies make use of powerful bioinformatics tools and resources, most of which are easily available web-based applications, together with pathway analysis tools and expression quantitative loci (e-QTL) to rank the importance of gene loci that have been discovered by GWAS (Kumar et al., 2012). Such software are useful in assigning the functionality of genetic variants using a number of approaches (Chami and Lettre, 2014). Some software rank genetic variants according to whether they are evolutionarily conserved, with conserved variants more likely to be important functionally (Lindblad-Toh et al., 2011). Others, use prediction methods based on conservation, physical and chemical changes that DNA variants may have on amino acids

(Chami and Lettre, 2014) including 'Sorting tolerant from intolerant (SIFT) algorithms' (Kumar et al., 2009) and Polyphen-2 (Adzhubei et al., 2010). It is also possible to test associated genotypes *in silico* to see if they have an effect on regulatory mechanisms as with e-QTL (Chami and Lettre, 2014). Initiatives such as ENCODE have been used to define regulatory sequences in human cell lines, to establish if non-coding SNPs fall within such regions (Consortium, 2012).

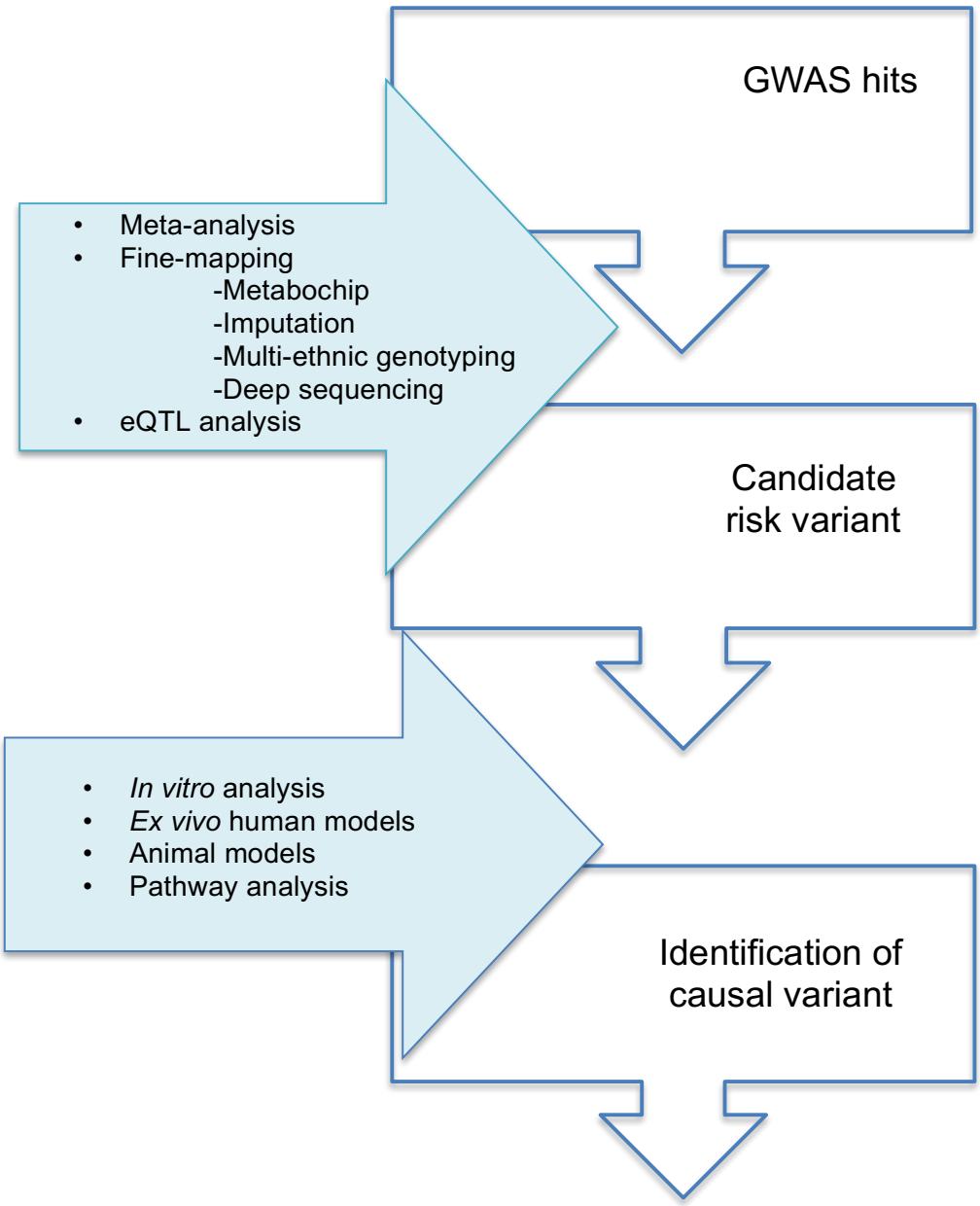


Figure 5.1 Summary of strategies to identify causal variants.
Adapted from (Kumar et al., 2012). GWAS signals can be followed up by using a meta-analysis or fine-mapping strategy. Then e-QTL or pathway analysis is helpful in ranking causative genes and these can be used to ask questions about how they relate to disease and identifying causal variants. These potential causal variants then have to be tested experimentally.

Also outlined in Figure 5.1, is the use of diverse ethnic populations for conducting GWAS or fine-mapping or imputation studies. Given, the current challenges of conducting research on the African continent, the use of the Metabochip to establish if European gene associations with obesity could be applied to African populations, was the most suitable choice at the time. This is mainly due to the fact that genetic analyses of African populations are particularly suited to fine-mapping disease loci in Caucasians, where this is difficult to achieve due to higher LD. However, an African-specific GWAS chip reflecting African LD structure would be the most appropriate tool to discover African-specific variants as well as adequately test replication of European GWAS signals.

One approach to advance our understanding of the biological mechanisms underlying a phenotype under investigation is to better refine a trait. These “sub-phenotypes" could be defined by severity of disease, age of onset, or the site and/or pattern of symptoms, such as that observed with inflammatory bowel disease, for example. In this way, we may detect associations with variants contributing different effects to sub-phenotypes that would otherwise be overlooked by considering all cases,simultaneously, as the same phenotype. However, by focusing on specific sub-phenotypes we reduce sample size, and thus will lose power to map loci contributing homogeneous effects to the unified phenotype (Morris et al., 2010). We are aware of the benefits of sub-phenotyping, however with restrictions in our current sample size we may loose power to detect associations.

In summary, this study using more precisely measured obesity-related phenotypes, has contributed new findings on obesity-associated gene variants in an African population. With new research initiatives like H3Africa (www.h3africa.org), whose mandate is to support genomics research related to human diversity and disease biology in Africa, it is possible that more African-based genomic research will be undertaken and will lead to the discovery of new and specific disease causing variants in African populations.

References

Reference List

1000 Genomes Project, C. (2010). A map of human genome variation from population-scale sequencing. *Nature,* 467**,** 1061-1073.

1000 Genomes Project, C. (2015). A global reference for human genetic variation. *Nature,* 526**,** 68-74.

Adeyemo, A., Chen, G., Zhou, J., Shriner, D., Doumatey, A., Huang, H.&Rotimi, C. (2010). FTO genetic variation and association with obesity in West Africans and African Americans. *Diabetes,* 59**,** 1549-1554.

Adeyemo, A.&Rotimi, C. (2014). What does genomic medicine mean for diverse populations? *Mol Genet Genomic Med,* 2**,** 3-6.

Adeyemo, A. A., Tekola-Ayele, F., Doumatey, A. P., Bentley, A. R., Chen, G., Huang, H., Zhou, J., Shriner, D., Fasanmade, O.&Okafor, G. (2015). Evaluation of Genome Wide Association Study Associated Type 2 Diabetes Susceptibility Loci in Sub Saharan Africans. *Frontiers in genetics,* 6.

Adler, A. J., Wiley, G. B.&Gaffney, P. M. (2013). Infinium Assay for Large-scale SNP Genotyping Applications. *Journal of visualized experiments: JoVE*.

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S.&Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nature methods,* 7**,** 248-249.

Almén, M. S., Jacobsson, J. A., Shaik, J. H. A., Olszewski, P. K., Cedernaes, J., Alsiö, J., Sreedharan, S., Levine, A. S., Fredriksson, R.&Marcus, C. (2010). The obesity gene, TMEM18, is of ancient origin, found in majority of neuronal cells in all major brain regions and associated with obesity in severely obese children. *BMC medical genetics,* 11**,** 1.

Andersen, M. K.&Sandholt, C. H. (2015). Recent Progress in the Understanding of Obesity: Contributions of Genome-Wide Association Studies. *Current obesity reports***,** 1-10.

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. *Nat Protoc,* 5**,** 1564-73.

Anderson, D., Cordell, H. J., Fakiola, M., Francis, R. W., Syn, G., Scaman, E. S. H., Davis, E., Miles, S. J., Mcleay, T.&Jamieson, S. E. (2015). First Genome-Wide Association Study in an Australian Aboriginal Population Provides Insights into Genetic Risk Factors for Body Mass Index and Type 2 Diabetes. *PLoS One,* 10**,** e0119333.

Andreasen, C. H., Mogensen, M. S., Borch-Johnsen, K., Sandbæk, A., Lauritzen, T., Almind, K., Hansen, L., Jørgensen, T., Pedersen, O.&Hansen, T. (2009). Studies of CTNNBL1 and FDFT1 variants and measures of obesity: analyses of quantitative traits and case-control studies in 18,014 Danes. *BMC medical genetics,* 10**,** 17.

Ardlie, K. G., Kruglyak, L.&Seielstad, M. (2002). Patterns of linkage disequilibrium in the human genome. *Nature Reviews Genetics,* 3**,** 299-309.

Armstrong, D. L., Reiff, A., Myones, B. L., Quismorio, F. P., Klein-Gitelman, M., Mccurdy, D., Wagner-Weiner, L., Silverman, E., Ojwang, J. O.&Kaufman, K. M. (2009). Identification of new SLE-associated genes with a two-step Bayesian study design. *Genes and immunity,* 10**,** 446-456.

Armstrong, M. E. G., Lambert, M. I., Sharwood, K. A.&Lambert, E. V. (2006). Obesity and overweight in South African primary school children-the Health of the Nation Study. *Journal of Endocrinology, Metabolism and Diabetes in South Africa,* 11**,** p. 52, 54, 56-57, 62-63.

Barker, D. J. P., Eriksson, J. G., Forsen, T.&Osmond, C. (2002). Fetal origins of adult disease: strength of effects and biological basis. *International journal of epidemiology,* 31**,** 1235-1239.

Barrett, J. C., Fry, B., Maller, J.&Daly, M. J. (2005). Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics,* 21**,** 263-265.

Baumgartner, R. N., Heymsfield, S. B.&Roche, A. F. (1995). Human body composition and the epidemiology of chronic disease. *Obesity Research,* 3**,** 73-95.

Benjamini, Y.&Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 289-300.

Berndt, S. I., Gustafsson, S., Magi, R., Ganna, A., Wheeler, E., Feitosa, M. F., Justice, A. E., Monda, K. L., Croteau-Chonka, D. C., Day, F. R., et al. (2013). Genome-wide meta-analysis identifies 11 new loci for anthropometric traits and provides insights into genetic architecture. *Nat Genet,* 45, 501-12.

Bernhard, F., Landgraf, K., Klöting, N., Berthold, A., Büttner, P., Friebe, D., Kiess, W., Kovacs, P., Blüher, M.&Körner, A. (2013). Functional relevance of genes implicated by obesity genome-wide association study signals for human adipocyte biology. *Diabetologia,* 56, 311-322.

Bille, D. S., Banasik, K., Justesen, J. M., Sandholt, C. H., Sandbæk, A., Lauritzen, T., Jørgensen, T., Witte, D. R., Holm, J.-C.&Hansen, T. (2011). Implications of central obesity-related variants in LYPLAL1, NRXN3, MSRA, and TFAP2B on quantitative metabolic traits in adult Danes. *PLoS One,* 6, e20640.

Bochukova, E. G., Huang, N., Keogh, J., Henning, E., Purmann, C., Blaszczyk, K., Saeed, S., Hamilton-Shield, J., Clayton-Smith, J.&O'rahilly, S. (2010). Large, rare chromosomal deletions associated with severe early-onset obesity. *Nature,* 463, 666-670.

Bodurtha, J. N., Mosteller, M., Hewitt, J. K., Nance, W. E., Eaves, L. J., Moskowitz, W. B.&Al.., E. (1990). Genetic analysis of anthropometric measures in 11-year-old twins: the Medical College of Virginia Twin Study. . *Pediatr Res,,* 28, 1-4.

Boender, A. J., Van Gestel, M. A., Garner, K. M., Luijendijk, M. C. M.&Adan, R. a. H. (2014). The obesity‐associated gene Negr1 regulates aspects of energy balance in rat hypothalamic areas. *Physiological reports,* 2, e12083.

Bogardus, C. (2009). Missing heritability and GWAS utility. *Obesity (Silver Spring),* 17, 209-210.

Boyle, A. P., Hong, E. L., Hariharan, M., Cheng, Y., Schaub, M. A., Kasowski, M., Karczewski, K. J., Park, J., Hitz, B. C.&Weng, S. (2012). Annotation of functional variation in personal genomes using RegulomeDB. *Genome research,* 22, 1790-1797.

Bradfield, J. P., Taal, H. R., Timpson, N. J., Scherag, A., Lecoeur, C., Warrington, N. M., Hypponen, E., Holst, C., Valcarcel, B., Thiering, E., et al. (2012). A genome-wide association meta-analysis identifies new childhood obesity loci. *Nat Genet,* 44, 526-31.

Budnik, A., Heesom, K. J.&Stephens, D. J. (2011). Characterization of human Sec16B: indications of specialized, non-redundant functions. *Scientific reports,* 1.

Burgio, E., Lopomo, A.&Migliore, L. (2015). Obesity and diabetes: from genetics to epigenetics. *Molecular biology reports,* 42, 799-818.

Buyske, S., Wu, Y., Carty, C. L., Cheng, I., Assimes, T. L., Dumitrescu, L., Hindorff, L. A., Mitchell, S., Ambite, J. L., Boerwinkle, E., et al. (2012). Evaluation of the metabochip genotyping array in African Americans and implications for fine mapping of GWAS-identified loci: the PAGE study. *PLoS One,* 7, e35651.

Cameron, N., De Wet, T., Ellison, G. T.&Bogin, B. (1998). Growth in height and weight of South African urban infants from birth to five years: The Birth to Ten Study. *American journal of human biology,* 10, 495-504.

Cargill, M., Altshuler, D., Ireland, J., Sklar, P., Ardlie, K., Patil, N., Lane, C. R., Lim, E. P., Kalyanaraman, N.&Nemesh, J. (1999). Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat Genet,* 22, 231-238.

Carlson, C. S., Matise, T. C., North, K. E., Haiman, C. A., Fesinmeyer, M. D., Buyske, S., Schumacher, F. R., Peters, U., Franceschini, N., Ritchie, M. D., et al. (2013). Generalization and dilution of association results from European GWAS in populations of non-European ancestry: the PAGE study. *PLoS Biol,* 11, e1001661.

Chacon-Cortes, D.& Griffiths, L. R. (2014). Methods for extracting genomic DNA from whole blood samples: current perspectives. *Journal of Biorepository Science for Applied Medicine,* 2014, 1-9.

Chakravarti, A. (1999). Population genetics—making sense out of sequence. *Nat Genet,* 21**,** 56-60.

Chambers, J. C., Elliott, P., Zabaneh, D., Zhang, W., Li, Y., Froguel, P., Balding, D., Scott, J.& Kooner, J. S. (2008). Common genetic variation near MC4R is associated with waist circumference and insulin resistance. *Nat Genet,* 40**,** 716-718.

Chami, N.& Lettre, G. (2014). Lessons and implications from genome-wide association studies (GWAS) findings of blood cell phenotypes. *Genes,* 5**,** 51-64.

Chimusa, E. R., Daya, M., Möller, M., Ramesar, R., Henn, B. M., Van Helden, P. D., Mulder, N. J.&Hoal, E. G. (2013). Determining ancestry proportions in complex admixture scenarios in South Africa using a novel proxy ancestry selection method. *PLoS One,* 8**,** e73971.

Cho, Y. S., Go, M. J., Kim, Y. J., Heo, J. Y., Oh, J. H., Ban, H.-J., Yoon, D., Lee, M. H., Kim, D.-J.&Park, M. (2009). A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits. *Nat Genet,* 41**,** 527-534.

Clarke, G. M., Anderson, C. A., Pettersson, F. H., Cardon, L. R., Morris, A. P.&Zondervan, K. T. (2011). Basic statistical analysis in genetic case-control studies. *Nat Protoc,* 6**,** 121-133.

Claussnitzer, M., Dankel, S. N., Kim, K.-H., Quon, G., Meuleman, W., Haugen, C., Glunk, V., Sousa, I. S., Beaudry, J. L.&Puviindran, V. (2015). FTO obesity variant circuitry and adipocyte browning in humans. *New England Journal of Medicine,* 373**,** 895-907.

Cole, T. J., Flegal, K. M., Nicholls, D.&Jackson, A. A. (2007). Body mass index cut offs to define thinness in children and adolescents: international survey. *BMJ,* 335**,** 194

Collins, F. S., Green, E. D., Guttmacher, A. E.&Guyer, M. S. (2003). A vision for the future of genomics research. *Nature,* 422**,** 835-847.

Collins, F. S., Guyer, M. S.&Chakravarti, A. (1997). Variations on a theme: cataloging human DNA sequence variation. *Science,* 278**,** 1580.

Comuzzie, A. G.&Allison, D. B. (1998). The search for human obesity genes. *Science,* 280**,** 1374-1377.

Comuzzie, A. G., Cole, S. A., Laston, S. L., Voruganti, V. S., Haack, K., Gibbs, R. A.&Butte, N. F. (2012). Novel genetic loci identified for the pathophysiology of childhood obesity in the Hispanic population.

Consortium, E. P. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature,* 489**,** 57-74.

Consortium, H. A., Rotimi, C., Abayomi, A., Abimiku, A., Adabayeri, V. M., Adebamowo, C., Adebiyi, E., Ademola, A. D., Adeyemo, A., Adu, D., et al. (2014). Research capacity. Enabling the genomic revolution in Africa. *Science,* 344**,** 1346-8.

Cooper, D. N. (2010). Functional intronic polymorphisms: buried treasure awaiting discovery within our genes. *Human genomics,* 4**,** 1.

Cordell, H. J.&Clayton, D. G. (2005). Genetic association studies. *The Lancet,* 366**,** 1121-1131.

Crawford, D. C., Goodloe, R., Brown-Gentry, K., Wilson, S., Roberson, J., Gillani, N. B., Ritchie, M. D., Dilks, H. H.&Bush, W. S. Characterization of the Metabochip in diverse populations from the International HapMap Project in the Epidemiologic Architecture for Genes Linked to Environment (EAGLE) project. Pac Symp Biocomput, 2013. World Scientific, 188-199.

Croteau‐Chonka, D. C., Marvelle, A. F., Lange, E. M., Lee, N. R., Adair, L. S., Lange, L. A.&Mohlke, K. L. (2011). Genome‐Wide Association Study of Anthropometric Traits and Evidence of Interactions With Age and Study Year in Filipino Women. *Obesity (Silver Spring),* 19**,** 1019-1027.

Crowther, N. J.&Ferris, W. F. (2010). The impact of insulin resistance, gender, genes, glucocorticoids and ethnicity on body fat distribution. *Journal of Endocrinology, Metabolism and Diabetes of South Africa,* 15**,** 115-120.

Crowther, N. J.&Norris, S. A. (2012). The current waist circumference cut point used for the diagnosis of metabolic syndrome in sub-Saharan African women is not appropriate. *PLoS One,* 7**,** e48883.

Dalal, S., Beunza, J. J., Volmink, J., Adebamowo, C., Bajunirwe, F., Njelekela, M., Mozaffarian, D., Fawzi, W., Willett, W.&Adami, H.-O. (2011). Non-communicable diseases in sub-Saharan Africa: what we know now. *International journal of epidemiology,* 40**,** 885-901.

De Bakker, P. I. W., Yelensky, R., Pe'er, I., Gabriel, S. B., Daly, M. J.&Altshuler, D. (2005). Efficiency and power in genetic association studies. *Nat Genet,* 37**,** 1217-1223.

Den Hoed, M., Ekelund, U., Brage, S., Grontved, A., Zhao, J. H., Sharp, S. J., Ong, K. K., Wareham, N. J.&Loos, R. J. F. (2010). Genetic Susceptibility to Obesity and Related Traits in Childhood and Adolescence Influence of Loci Identified by Genome-Wide Association Studies. *Diabetes,* 59**,** 2980-2988.

Development Core Team, R. (2008). R: A language and environment for statistical computing. *Foundation for Statistical Computing*.

Devlin, B., Roeder, K.&Wasserman, L. (2001). Genomic control, a new approach to genetic-based association studies. *Theoretical population biology,* 60**,** 155-166.

Dina, C., Meyre, D., Gallina, S., Durand, E., Körner, A., Jacobson, P., Carlsson, L. M. S., Kiess, W., Vatin, V.&Lecoeur, C. (2007). Variation in FTO contributes to childhood obesity and severe adult obesity. *Nat Genet,* 39**,** 724-726.

Ding, T.&Schloss, P. D. (2014). Dynamics and associations of microbial community types across the human body. *Nature,* 509**,** 357-360.

Dorajoo, R., Blakemore, A. I. F., Sim, X., Ong, R. T., Ng, D. P. K., Seielstad, M., Wong, T. Y., Saw, S. M., Froguel, P.&Liu, J. (2012). Replication of 13 obesity loci among Singaporean Chinese, Malay and Asian-Indian populations. *Int J Obes (Lond),* 36**,** 159-163.

Dudbridge, F.&Gusnanto, A. (2008). Estimation of significance thresholds for genomewide association scans. *Genet Epidemiol,* 32**,** 227.

Duren, D. L., Sherwood, R. J., Czerwinski, S. A., Lee, M., Choh, A. C., Siervogel, R. M.&Chumlea, W. C. (2008). Body composition methods: comparisons and interpretation. *Journal of diabetes science and technology,* 2**,** 1139-1146.

Ehret, G. B. (2010). Genome-wide association studies: contribution of genomics to understanding blood pressure and essential hypertension. *Current hypertension reports,* 12**,** 17-25.

El-Sayed Moustafa, J. S.&Froguel, P. (2013). From obesity genetics to the future of personalized obesity therapy. *Nat Rev Endocrinol,* 9**,** 402-13.

Elks, C. E., Den Hoed, M., Zhao, J. H., Sharp, S. J., Wareham, N. J., Loos, R. J. F.&Ong, K. K. (2012). Variability in the heritability of body mass index: a systematic review and meta-regression. *Frontiers in endocrinology,* 3.

Ellis, K. J. (2005). Whole-body counting and neutron activation analysis. *Human Body Composition***,** 51-62.

Eu-Ahsunthornwattana, J., Miller, E. N., Fakiola, M., Jeronimo, S. M. B., Blackwell, J. M., Cordell, H. J.&Wellcome Trust Case Control, C. (2014). Comparison of methods to account for relatedness in genome-wide association studies with family-based data.

Falconer, D.&Mackay, T. (1995). I ntroduction to Q uantitative Genetics. *Longman,* 19**,** 1.

Fall, T.&Ingelsson, E. (2014). Genome-wide association studies of obesity and metabolic syndrome. *Molecular and cellular endocrinology,* 382**,** 740-757.

Farooqi, S. (2012). Genetic, molecular and physiological mechanisms involved in human obesity. *Society for Endocrinology Medal Lecture*.

Feeley, A., Musenge, E., Pettifor, J. M.&Norris, S. A. (2012). Changes in dietary habits and eating practices in adolescents living in urban South Africa: The birth to twenty cohort. *Nutrition,* 28**,** e1-e6.

Feeley, A., Pettifor, J. M.&Norris, S. A. (2009). Fast-food consumption among 17-year-olds in the Birth to Twenty cohort. *South African Journal of Clinical Nutrition,* 22.

Feeley, A. B. B. (2012). *THE IMPACT OF DIETARY HABITS AND PRACTICES DURING ADOLESCENCE ON THE RISK OF OBESITY: THE BIRTH TO TWENTY COHORT.* Faculty of Health Sciences, University of the Witwatersrand.

Felix, J. F., Bradfield, J. P., Monnereau, C., Van Der Valk, R. J. P., Stergiakouli, E., Chesi, A., Gaillard, R., Feenstra, B., Thiering, E.&Kreiner-Møller, E. (2015). Genome-wide

association analysis identifies three new susceptibility loci for childhood body mass index. *Hum Mol Genet*, ddv472.
Feng, S., Wang, S., Chen, C.-C.&Lan, L. (2011). GWAPower: a statistical power calculation software for genome-wide association studies with quantitative traits. *BMC Genet,* 12, 12.
Frayling, T. M., Timpson, N. J., Weedon, M. N., Zeggini, E., Freathy, R. M., Lindgren, C. M., Perry, J. R. B., Elliott, K. S., Lango, H.&Rayner, N. W. (2007). A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science,* 316, 889-894.
Fu, J., Festen, E. A.&Wijmenga, C. (2011). Multi-ethnic studies in complex traits. *Hum Mol Genet*, ddr386.
Gibbs, R. A., Belmont, J. W., Hardenbol, P., Willis, T. D., Yu, F., Yang, H., Ch'ang, L.-Y., Huang, W., Liu, B.&Shen, Y. (2003). The international HapMap project. *Nature,* 426, 789-796.
Gillman, M. W. (2005). Developmental origins of health and disease. *The New England journal of medicine,* 353, 1848.
Goedecke, J. H., Jennings, C. L.&Lambert, E. V. (2005). Obesity in South Africa. *Chronic diseases of lifestyle in South Africa.*
Goldberg, A. D., Allis, C. D.&Bernstein, E. (2007). Epigenetics: a landscape takes shape. *Cell,* 128, 635-638.
Gong, J., Schumacher, F., Lim, U., Hindorff, L. A., Haessler, J., Buyske, S., Carlson, C. S., Rosse, S., Buzkova, P., Fornage, M., et al. (2013). Fine Mapping and Identification of BMI Loci in African Americans. *Am J Hum Genet,* 93, 661-71.
Graham, C., Mullen, A.&Whelan, K. (2015). Obesity and the gastrointestinal microbiota: a review of associations and mechanisms. *Nutr Rev*, nuv004.
Guh, D. P., Zhang, W., Bansback, N., Amarsi, Z., Birmingham, C. L.&Anis, A. H. (2009). The incidence of co-morbidities related to obesity and overweight: a systematic review and meta-analysis. *BMC Public Health,* 9, 1.
Gurdasani, D., Carstensen, T., Tekola-Ayele, F., Pagani, L., Tachmazidou, I., Hatzikotoulas, K., Karthikeyan, S., Iles, L., Pollard, M. O.&Choudhury, A. (2015). The African genome variation project shapes medical genetics in Africa. *Nature,* 517, 327-332.
Han, B., Kang, H. M.&Eskin, E. (2009). Rapid and accurate multiple testing correction and power estimation for millions of correlated markers. *PLoS Genet,* 5, e1000456.
Hartl, D. L.&Clark, A. G. (1997). *Principles of population genetics*, Sinauer associates Sunderland.
Hassanein, M. T., Lyon, H. N., Nguyen, T. T., Akylbekova, E. L., Waters, K., Lettre, G., Tayo, B., Forrester, T., Sarpong, D. F., Stram, D. O., et al. (2010). Fine mapping of the association with obesity at the FTO locus in African-derived populations. *Hum Mol Genet,* 19, 2907-16.
Heard-Costa, N. L., Zillikens, M. C., Monda, K. L., Johansson, A., Harris, T. B., Fu, M., Haritunians, T., Feitosa, M. F., Aspelund, T.&Eiriksdottir, G. (2009). NRXN3 is a novel locus for waist circumference: a genome-wide association study from the CHARGE Consortium. *PLoS Genet,* 5, e1000539.
Hebebrand, J., Hinney, A., Knoll, N., Volckmar, A. L.&Scherag, A. (2013). Molecular genetic aspects of weight regulation. *Dtsch Arztebl Int,* 110, 338-44.
Heid, I. M., Jackson, A. U., Randall, J. C., Winkler, T. W., Qi, L., Steinthorsdottir, V., Thorleifsson, G., Zillikens, M. C., Speliotes, E. K., Magi, R., et al. (2010). Meta-analysis identifies 13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution. *Nat Genet,* 42, 949-60.
Hemani, G., Knott, S.&Haley, C. (2013). An evolutionary perspective on epistasis and the missing heritability. *PLoS Genet,* 9, e1003295.
Hennig, B. J., Fulford, A. J., Sirugo, G., Rayco-Solon, P., Hattersley, A. T., Frayling, T. M.&Prentice, A. M. (2009). FTO gene variation and measures of body mass in an African population. *BMC medical genetics,* 10, 21.
Heymsfield, S. (2005). *Human body composition*, Human kinetics.

Hiura, Y., Tabara, Y., Kokubo, Y., Okamura, T., Miki, T., Tomoike, H.&Iwai, N. (2010). A genome-wide association study of hypertension-related phenotypes in a Japanese population. *Circulation Journal,* 74**,** 2353-2359.

Horikoshi, M., Mägi, R., Van De Bunt, M., Surakka, I., Sarin, A.-P., Mahajan, A., Marullo, L., Thorleifsson, G., Hägg, S.&Hottenga, J.-J. (2015). Discovery and fine-mapping of glycaemic and obesity-related trait loci using high-density imputation. *PLoS Genet,* 11**,** e1005230.

Hotta, K., Kitamoto, T., Kitamoto, A., Mizusawa, S., Matsuo, T., Nakata, Y., Hyogo, H., Ochi, H., Kamohara, S.&Miyatake, N. (2011). Computed tomography analysis of the association between the SH2B1 rs7498665 single-nucleotide polymorphism and visceral fat area. *Journal of human genetics,* 56**,** 716-719.

Hotta, K., Nakamura, M., Nakamura, T., Matsuo, T., Nakata, Y., Kamohara, S., Miyatake, N., Kotani, K., Komatsu, R.&Itoh, N. (2009). Association between obesity and polymorphisms in SEC16B, TMEM18, GNPDA2, BDNF, FAIM2 and MC4R in a Japanese population. *Journal of human genetics,* 54**,** 727-731.

Hsu, F. C., Lenchik, L., Nicklas, B. J., Lohman, K., Register, T. C., Mychaleckyj, J., Langefeld, C. D., Freedman, B. I., Bowden, D. W.&Carr, J. J. (2005). Heritability of body composition measured by DXA in the diabetes heart study. *Obesity Research,* 13**,** 312-319.

Jallow, M., Teo, Y. Y., Small, K. S., Rockett, K. A., Deloukas, P., Clark, T. G., Kivinen, K., Bojang, K. A., Conway, D. J.&Pinder, M. (2009). Genome-wide and fine-resolution association analysis of malaria in West Africa. *Nat Genet,* 41**,** 657-665.

Janssen, I., Katzmarzyk, P. T.&Ross, R. (2004). <Waist circumference and not BMI explains obesity related health risk.pdf>. *Am J Clin Nutr,* 79**,** 379–84.

Jenkins, A.&Campbell, L. V. (2014). Future management of human obesity: understanding the meaning of genetic susceptibility.

Jiao, H., Arner, P., Hoffstedt, J., Brodin, D., Dubern, B., Czernichow, S., Van't Hooft, F., Axelsson, T., Pedersen, O.&Hansen, T. (2011). Genome wide association study identifies KCNMA1 contributing to human obesity. *BMC medical genomics,* 4**,** 51.

Joubert, J., Norman, R., Bradshaw, D., Goedecke, J. H., Steyn, N. P., Puoane, T.&Collaboration, S. a. C. R. A. (2007). Estimating the burden of disease attributable to excess body weight in South Africa in 2000. *South African Medical Journal,* 97**,** 683-690.

Kaar, J. L., Crume, T., Brinton, J. T., Bischoff, K. J., Mcduffie, R.&Dabelea, D. (2014). Maternal obesity, gestational weight gain, and offspring adiposity: the exploring perinatal outcomes among children study. *The Journal of pediatrics,* 165**,** 509-515.

Kang, S. J., Chiang, C. W., Palmer, C. D., Tayo, B. O., Lettre, G., Butler, J. L., Hackett, R., Adeyemo, A. A., Guiducci, C., Berzins, I., et al. (2010). Genome-wide association of anthropometric traits in African- and African-derived populations. *Hum Mol Genet,* 19**,** 2725-38.

Kengne, A. P., Echouffo-Tcheugui, J.-B., Sobngwi, E.&Mbanya, J.-C. (2013). New insights on diabetes mellitus and obesity in Africa–Part 1: prevalence, pathogenesis and comorbidities. *Heart***,** heartjnl-2012-303316.

Kilpeläinen, T. O., Zillikens, M. C., Stančáková, A., Finucane, F. M., Ried, J. S., Langenberg, C., Zhang, W., Beckmann, J. S., Luan, J. A.&Vandenput, L. (2011). Genetic variation near IRS1 associates with reduced adiposity and an impaired metabolic profile. *Nat Genet,* 43**,** 753-760.

Klimentidis, Y. C., Abrams, M., Wang, J., Fernandez, J. R.&Allison, D. B. (2011). Natural selection at genomic regions associated with obesity and type-2 diabetes: East Asians and sub-Saharan Africans exhibit high levels of differentiation at type-2 diabetes regions. *Human genetics,* 129**,** 407-418.

Kosoy, R., Nassir, R., Tian, C., White, P. A., Butler, L. M., Silva, G., Kittles, R., Alarcon-Riquelme, M. E., Gregersen, P. K.&Belmont, J. W. (2009). Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America. *Human mutation,* 30**,** 69-78.

Kruger, H. S., Micklesfield, L. K., Wright, H. H., Havemann-Nel, L.&Goedecke, J. H. (2015). Ethnic-specific cut-points for sarcopenia: evidence from black South African women. *Eur J Clin Nutr*.

Kumar, P., Henikoff, S.&Ng, P. C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc,* 4**,** 1073-1081.

Kumar, V., Wijmenga, C.&Withoff, S. From genome-wide association studies to disease mechanisms: celiac disease as a model for autoimmune diseases. Seminars in immunopathology, 2012. Springer, 567-580.

Lander, E. S. (1996). The new genomics: global views of biology. *Science,* 274**,** 536-539.

Lau, E., Carvalho, D., Pina-Vaz, C., Barbosa, J.-A.&Freitas, P. (2015). Beyond gut microbiota: understanding obesity and type 2 diabetes. *Hormones (Athens, Greece),* 14**,** 358-369.

Laurie, C. C., Doheny, K. F., Mirel, D. B., Pugh, E. W., Bierut, L. J., Bhangale, T., Boehm, F., Caporaso, N. E., Cornelis, M. C.&Edenberg, H. J. (2010). Quality control and quality assurance in genotypic data for genome‐wide association studies. *Genet Epidemiol,* 34**,** 591-602.

Lear, S. A., James, P. T., Ko, G. T.&Kumanyika, S. (2010). Appropriateness of waist circumference and waist-to-hip ratio cutoffs for different ethnic groups. *Eur J Clin Nutr,* 64**,** 42-61.

Lee, A. W. S., Hengstler, H., Schwald, K., Berriel-Diaz, M., Loreth, D., Kirsch, M., Kretz, O., Haas, C. A., De Angelis, M. H.&Herzig, S. (2012). Functional inactivation of the genome-wide association study obesity gene neuronal growth regulator 1 in mice causes a body mass phenotype. *PLoS One,* 7**,** e41537.

Lewis, C. M.&Knight, J. (2012). Introduction to genetic association studies. *Cold Spring Harbor protocols,* 2012**,** pdb. top068163.

Ley, R. E. (2010). Obesity and the human microbiome. *Current opinion in gastroenterology,* 26**,** 5-11.

Li, E. N. (2002). Chromatin modification and epigenetic reprogramming in mammalian development. *Nature Reviews Genetics,* 3**,** 662-673.

Li, S., Schlebusch, C.&Jakobsson, M. (2014). Genetic variation reveals large-scale population expansion and migration during the expansion of Bantu-speaking peoples. *Proceedings of the Royal Society of London B: Biological Sciences,* 281**,** 20141448.

Lim, S. S., Vos, T., Flaxman, A. D., Danaei, G., Shibuya, K., Adair-Rohani, H., Almazroa, M. A., Amann, M., Anderson, H. R.&Andrews, K. G. (2013). A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. *The Lancet,* 380**,** 2224-2260.

Lindblad-Toh, K., Garber, M., Zuk, O., Lin, M. F., Parker, B. J., Washietl, S., Kheradpour, P., Ernst, J., Jordan, G.&Mauceli, E. (2011). A high-resolution map of human evolutionary constraint using 29 mammals. *Nature,* 478**,** 476-482.

Lindgren, C. M., Heid, I. M., Randall, J. C., Lamina, C., Steinthorsdottir, V., Qi, L., Speliotes, E. K., Thorleifsson, G., Willer, C. J.&Herrera, B. M. (2009). Genome-wide association scan meta-analysis identifies three Loci influencing adiposity and fat distribution. *PLoS Genet,* 5**,** e1000508.

Liu, C.-T., Buchkovich, M. L., Winkler, T. W., Heid, I. M., Borecki, I., Fox, C. S., Mohlke, K. L., North, K. E., Cupples, L. A.&African Ancestry Anthropometry Genetics, C. (2014). Multi-ethnic fine-mapping of 14 central adiposity loci. *Hum Mol Genet***,** ddu183.

Liu, C. T., Monda, K. L., Taylor, K. C., Lange, L., Demerath, E. W., Palmas, W., Wojczynski, M. K., Ellis, J. C., Vitolins, M. Z., Liu, S., et al. (2013). Genome-wide association of body fat distribution in African ancestry populations suggests new loci. *PLoS Genet,* 9**,** e1003681.

Liu, D., Zhu, Z.&Tepel, M. (2008a). The role of transient receptor potential channels in metabolic syndrome. *Hypertens Res,* 31**,** 1989-1995.

Liu, Y.-J., Liu, X.-G., Wang, L., Dina, C., Yan, H., Liu, J.-F., Levy, S., Papasian, C. J., Drees, B. M.&Hamilton, J. J. (2008b). Genome-wide association scans identified CTNNBL1 as a novel gene for obesity. *Hum Mol Genet,* 17**,** 1803-1813.

Locke, A. E., Kahali, B., Berndt, S. I., Justice, A. E., Pers, T. H., Day, F. R., Powell, C., Vedantam, S., Buchkovich, M. L.&Yang, J. (2015). Genetic studies of body mass index yield new insights for obesity biology. *Nature,* 518**,** 197-206.

Lohman, T.&Milliken, L. (2003). Body composition assessment in the obese. *Obesity: Etiology, Assessment, Treatment, and Prevention. Champaign, IL: Human Kinetics***,** 73-84.

Lombard, Z. (2008). *Computational Computational Prediction Prediction Prediction of Gene Targets for Fetal Alcohol for Fetal Alcohol for Fetal Alcohol Spectrum Spectrum Disorders Disorders.* Faculty of Health Science, University of the Witwatersrand, Johannesburg.

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.

Loos, R. J. (2012). Genetic determinants of common obesity and their value in prediction. *Best Pract Res Clin Endocrinol Metab,* 26**,** 211-26.

Loos, R. J. 2013. Obesity: Genetic Factors. *In:* B., C. (ed.) *Encyclopedia of Human Nutrition.* Waltham, MA: Academic Press.

Loos, R. J. F. (2009). Recent progress in the genetics of common obesity. *British journal of clinical pharmacology,* 68**,** 811-829.

Loos, R. J. F., Lindgren, C. M., Li, S., Wheeler, E., Zhao, J. H., Prokopenko, I., Inouye, M., Freathy, R. M., Attwood, A. P.&Beckmann, J. S. (2008). Common variants near MC4R are associated with fat mass, weight and risk of obesity. *Nat Genet,* 40**,** 768-775.

Lu, Y., Day, F. R., Gustafsson, S., Buchkovich, M. L., Na, J., Bataille, V., Cousminer, D. L., Dastani, Z., Drong, A. W., Esko, T., et al. (2016). New loci for body fat percentage reveal link between adiposity and cardiometabolic disease risk. *Nat Commun,* 7**,** 10495.

Lu, Y.&Loos, R. J. (2013). Obesity genomics: assessing the transferability of susceptibility loci across diverse populations. *Genome Med,* 5**,** 55.

Lukaski, H. C. (1987). Methods for the assessment of human body composition: traditional and new. *Am J Clin Nutr,* 46**,** 537-556.

Lukaski, H. C., Johnson, P. E., Bolonchuk, W. W.&Lykken, G. I. (1985). Assessment of fat-free mass using bioelectrical impedance measurements of the human body. *Am J Clin Nutr,* 41**,** 810-817.

Luke, A., Guo, X., Adeyemo, A. A., Wilks, R., Forrester, T., Lowe Jr, W., Comuzzie, A. G., Martin, L. J., Zhu, X.&Rotimi, C. N. (2001). Heritability of obesity-related traits among Nigerians, Jamaicans and US black people. *International journal of obesity and related metabolic disorders: journal of the International Association for the Study of Obesity,* 25**,** 1034-1041.

Lundeen, E. A., Norris, S. A., Adair, L. S., Richter, L. M.&Stein, A. D. (2015). Sex differences in obesity incidence: 20‐year prospective cohort in South Africa. *Pediatric obesity*.

Mackay, T. F., Stone, E. A.&Ayroles, J. F. (2009). The genetics of quantitative traits: challenges and prospects. *Nature Reviews Genetics,* 10**,** 565-577.

Maher, B. (2008). Personal genomes: The case of the missing heritability. *Nature News,* 456**,** 18-21.

Malhotra, A., Kobes, S., Knowler, W. C., Baier, L. J., Bogardus, C.&Hanson, R. L. (2011). A Genome‐Wide Association Study of BMI in American Indians. *Obesity (Silver Spring),* 19**,** 2102-2106.

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. (2009). Finding the missing heritability of complex diseases. *Nature,* 461**,** 747-753.

Marian, A. J. (2012). Elements of 'missing heritability'. *Current opinion in cardiology,* 27**,** 197-201.

May, A., Hazelhurst, S., Li, Y., Norris, S. A., Govind, N., Tikly, M., Hon, C., Johnson, K. J., Hartmann, N.&Staedtler, F. (2013). Genetic diversity in black South Africans from Soweto. *BMC genomics,* 14**,** 644.

Mayosi, B. M., Flisher, A. J., Lalloo, U. G., Sitas, F., Tollman, S. M.&Bradshaw, D. (2009). The burden of non-communicable diseases in South Africa. *Lancet,* 374**,** 934-947.

Mboowa, G. (2014). Genetics of sub-Saharan African human population: implications for HIV/AIDS, tuberculosis, and malaria. *International journal of evolutionary biology,* 2014.

Mcpherson, J. D., Marra, M., Hillier, L., Waterston, R. H., Chinwalla, A., Wallis, J., Sekhon, M., Wylie, K., Mardis, E. R.&Wilson, R. K. (2001). A physical map of the human genome. *Nature,* 409**,** 934-941.

Melka, M. G., Bernard, M., Mahboubi, A., Abrahamowicz, M., Paterson, A. D., Syme, C., Lourdusamy, A., Schumann, G., Leonard, G. T.&Perron, M. (2011). Genome-wide scan for loci of adolescent obesity and their relationship with blood pressure. *The Journal of Clinical Endocrinology & Metabolism,* 97**,** E145-E150.

Menke, A., Muntner, P., Wildman, R. P., Reynolds, K.&He, J. (2007). Measures of adiposity and cardiovascular disease risk factors. *Obesity (Silver Spring),* 15**,** 785-795.

Mercer, T. R., Dinger, M. E.&Mattick, J. S. (2009). Long non-coding RNAs: insights into functions. *Nature Reviews Genetics,* 10**,** 155-159.

Meyre, D., Delplanque, J., Chèvre, J.-C., Lecoeur, C., Lobbens, S., Gallina, S., Durand, E., Vatin, V., Degraeve, F.&Proença, C. (2009). Genome-wide association study for early-onset and morbid adult obesity identifies three new risk loci in European populations. *Nat Genet,* 41**,** 157-159.

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**,** 1215.

Monda, K. L., Chen, G. K., Taylor, K. C., Palmer, C., Edwards, T. L., Lange, L. A., Ng, M. C., Adeyemo, A. A., Allison, M. A., Bielak, L. F., et al. (2013). A meta-analysis identifies new loci associated with body mass index in individuals of African ancestry. *Nat Genet,* 45**,** 690-6.

Mou, Z., Hyde, T. M., Lipska, B. K., Martinowich, K., Wei, P., Ong, C.-J., Hunter, L. A., Palaguachi, G. I., Morgun, E.&Teng, R. (2015). Human Obesity Associated with an Intronic SNP in the Brain-Derived Neurotrophic Factor Locus. *Cell reports,* 13**,** 1073-1080.

Müller, M., Bosy‐Westphal, A.&Krawczak, M. (2010). Genetic studies of common types of obesity: a critique of the current use of phenotypes. *Obesity Reviews,* 11**,** 612-618.

Murphy, G. A., Asiki, G., Nsubuga, R. N., Young, E. H., Seeley, J., Sandhu, M. S.&Kamali, A. (2014). The use of anthropometric measures for cardiometabolic risk identification in a rural African population. *Diabetes Care,* 37**,** e64-5.

Murrin, C. M., Kelly, G. E., Tremblay, R. E.&Kelleher, C. C. (2012). Body mass index and height over three generations: evidence from the Lifeways cross-generational cohort study. *BMC Public Health,* 12**,** 1.

Musunuru, K., Romaine, S. P., Lettre, G., Wilson, J. G., Volcik, K. A., Tsai, M. Y., Taylor Jr, H. A., Schreiner, P. J., Rotter, J. I.&Rich, S. S. (2012). Multi-ethnic analysis of lipid-associated loci: the NHLBI CARe project. *PLoS One,* 7**,** e36473.

N.I.H. (2007). *Policy for sharing of data obtained in NIH supported or conducted genome-wide association studies (GWAS)*, National Institutes of Health.

Nan, C., Guo, B., Warner, C., Fowler, T., Barrett, T., Boomsma, D., Nelson, T., Whitfield, K., Beunen, G.&Thomis, M. (2012). Heritability of body mass index in pre-adolescence, young adulthood and late adulthood. *European journal of epidemiology,* 27**,** 247-253.

Neel, J. V. (1962). Diabetes mellitus: a “thrifty” genotype rendered detrimental by “progress”? *Am J Hum Genet,* 14**,** 353.

Nettle, D.&Harriss, L. (2003). Genetic and linguistic affinities between human populations in Eurasia and West Africa. *Human Biology***,** 331-344.

Ng, M., Fleming, T., Robinson, M., Thomson, B., Graetz, N., Margono, C., Mullany, E. C., Biryukov, S., Abbafati, C., Abera, S. F., et al. (2014). Global, regional, and national prevalence of overweight and obesity in children and adults during 1980–2013: a systematic analysis for the Global Burden of Disease Study 2013. *The Lancet*.

Ng, M. C., Hester, J. M., Wing, M. R., Li, J., Xu, J., Hicks, P. J., Roh, B. H., Lu, L., Divers, J., Langefeld, C. D., et al. (2012). Genome-wide association of BMI in African Americans. *Obesity (Silver Spring),* 20**,** 622-7.

Ng, M. C. Y., Tam, C. H. T., So, W. Y., Ho, J. S. K., Chan, A. W., Lee, H. M., Wang, Y., Lam, V. K. L., Chan, J. C. N.&Ma, R. C. W. (2010). Implication of Genetic Variants Near NEGR1, SEC16B, TMEM18, ETV5/DGKG, GNPDA2, LIN7C/BDNF, MTCH2, BCDIN3D/FAIM2, SH2B1, FTO, MC4R, and KCTD15 with Obesity and Type 2 Diabetes in 7705 Chinese. *The Journal of Clinical Endocrinology & Metabolism,* 95**,** 2418-2425.

Ngcungcu, T. (2013). *Exploring the role of genetic variation at the leptin and leptin receptor genes (LEP and LEPR) in obesity and hypertension in a black South African cohort.* Faculty of Health Sciences, University of the Witwatersrand, Johannesburg.

Nishida, C., Ko, G. T.&Kumanyika, S. (2010). Body fat distribution and noncommunicable diseases in populations: overview of the 2008 WHO Expert Consultation on Waist Circumference and Waist–Hip Ratio. *Eur J Clin Nutr,* 64**,** 2-5.

Norris, S. A.&Richter, L. M. (2005). Usefulness and Reliability of Tanner Pubertal Self‐Rating to Urban Black Adolescents in South Africa. *Journal of Research on Adolescence,* 15**,** 609-624.

Obesity Task Force, W. H. O. (2005). Global Strategy on Diet, Physical Activity and Health: Obesity and Overweight.

Okada, Y., Kubo, M., Ohmiya, H., Takahashi, A., Kumasaka, N., Hosono, N., Maeda, S., Wen, W., Dorajoo, R.&Go, M. J. (2012). Common variants at CDKAL1 and KLF9 are associated with body mass index in east Asian populations. *Nat Genet,* 44**,** 302-306.

Okorodudu, D., Jumean, M., Montori, V. M., Romero-Corral, A., Somers, V., Erwin, P.&Lopez-Jimenez, F. (2010). Diagnostic performance of body mass index to identify obesity as defined by body adiposity: a systematic review and meta-analysis. *Int J Obes (Lond),* 34**,** 791-799.

Omran, A. R. (2001). The epidemiologic transition. A theory of the Epidemiology of population change. 1971. *Bulletin of the World Health Organization,* 79**,** 161.

Parekh, P. J., Arusi, E., Vinik, A. I.&Johnson, D. A. (2014). The role and influence of gut microbiota in pathogenesis and management of obesity and metabolic syndrome. *Frontiers in endocrinology,* 5.

Park, H. S., Hong, C., Kim, B. J.&So, I. (2014). The pathophysiologic roles of TRPM7 channel. *The Korean Journal of Physiology & Pharmacology,* 18**,** 15-23.

Pearson, T. A.&Manolio, T. A. (2008). How to interpret a genome-wide association study. *JAMA: the journal of the American Medical Association,* 299**,** 1335-1344.

Peden, J. F.&Farrall, M. (2011). Thirty-five common variants for coronary artery disease: the fruits of much collaborative labour. *Hum Mol Genet***,** ddr384.

Perneger, T. V. (1998). What's wrong with Bonferroni adjustments. *BMJ: British Medical Journal,* 316**,** 1236.

Peters, U., North, K. E., Sethupathy, P., Buyske, S., Haessler, J., Jiao, S., Fesinmeyer, M. D., Jackson, R. D., Kuller, L. H.&Rajkovic, A. (2013). A systematic mapping approach of 16q12. 2/FTO and BMI in more than 20,000 African Americans narrows in on the underlying functional variation: results from the Population Architecture using Genomics and Epidemiology (PAGE) study. *PLoS Genet,* 9**,** e1003171.

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

Polderman, T. J., Benyamin, B., De Leeuw, C. A., Sullivan, P. F., Van Bochoven, A., Visscher, P. M.&Posthuma, D. (2015). Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nat Genet*.

Popkin, B. M. (2006). Global nutrition dynamics: the world is shifting rapidly toward a diet linked with noncommunicable diseases. *Am J Clin Nutr,* 84**,** 289-298.

Popkin, B. M., Adair, L. S.&Ng, S. W. (2012). Global nutrition transition and the pandemic of obesity in developing countries. *Nutr Rev,* 70**,** 3-21.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A.&Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet,* 38**,** 904-909.

Pruim, R. J., Welch, R. P., Sanna, S., Teslovich, T. M., Chines, P. S., Gliedt, T. P., Boehnke, M., Abecasis, G. R.&Willer, C. J. (2010). LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics,* 26**,** 2336-7.

Puoane, T., Steyn, K., Bradshaw, D., Laubscher, R., Fourie, J., Lambert, V.&Mbananga, N. (2002). Obesity in South Africa: the South African demographic and health survey. *Obesity Research,* 10**,** 1038-1048.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., M.A.R, F., Bender, D., Maller, J., Sklar, P., De Bakker, P., Daly, M., et al. (2007). PLINK: a toolset for whole-genome association and population-based linkage analysis. *Am J Hum Genet,* 81.

Ramsay, M., Tiemessen, C. T., Choudhury, A.&Soodyall, H. (2011). Africa: the next frontier for human disease gene discovery? *Hum Mol Genet,* 20**,** R214-R220.

Randall, J. C., Winkler, T. W., Kutalik, Z., Berndt, S. I., Jackson, A. U., Monda, K. L., Kilpeläinen, T. O., Esko, T., Mägi, R.&Li, S. (2013). Sex-stratified genome-wide association studies including 270,000 individuals show sexual dimorphism in genetic loci for anthropometric traits.

Ravelli, G.-P., Stein, Z. A.&Susser, M. W. (1976). Obesity in young men after famine exposure in utero and early infancy. *New England Journal of Medicine,* 295**,** 349-353.

Reich, D. E.&Lander, E. S. (2001). On the allelic spectrum of human disease. *TRENDS in Genetics,* 17**,** 502-510.

Reilly, J. J.&Kelly, J. (2010). Long-term impact of overweight and obesity in childhood and adolescence on morbidity and premature mortality in adulthood: systematic review. *Int J Obes (Lond),* 35**,** 891-898.

Ren, D., Zhou, Y., Morris, D., Li, M., Li, Z.&Rui, L. (2007). Neuronal SH2B1 is essential for controlling energy and glucose homeostasis. *The Journal of clinical investigation,* 117**,** 397-406.

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**,** 504-511.

Richter, L. M., Panday, S.&Norris, S. A. (2009). Factors influencing enrollment: a case study from Birth to Twenty, the 1990 birth cohort in Soweto-Johannesburg. *Eval Program Plann,* 32**,** 197-203.

Rose, K. M., Newman, B., Mayer‐Davis, E. J.&Selby, J. V. (1998). Genetic and Behavioral Determinants of Waist‐Hip Ratio and Waist Circumference in Women Twins. *Obesity Research,* 6**,** 383-392.

Rosenberg, N. A., Huang, L., Jewett, E. M., Szpiech, Z. A., Jankovic, I.&Boehnke, M. (2010). Genome-wide association studies in diverse populations. *Nature Reviews Genetics,* 11**,** 356-366.

Sandholt, C. H., Grarup, N., Pedersen, O.&Hansen, T. (2015). Genome-wide association studies of human adiposity: Zooming in on synapses. *Molecular and cellular endocrinology*.

Santos, D. A., Dawson, J. A., Matias, C. N., Rocha, P. M., Minderico, C. S., Allison, D. B., Sardinha, L. B.&Silva, A. M. (2014). Reference values for body composition and anthropometric measurements in athletes.

Santos, D. A., Silva, A. M., Matias, C. N., Fields, D. A., Heymsfield, S. B.&Sardinha, L. B. (2010). Accuracy of DXA in estimating body composition changes in elite athletes using a four compartment model as the reference method. *Nutr Metab (Lond),* 7**,** 7075-7.

Schäfer, M., Bräuer, A. U., Savaskan, N. E., Rathjen, F. G.&Brümmendorf, T. (2005). Neurotractin/kilon promotes neurite outgrowth and is expressed on reactive astrocytes after entorhinal cortex lesion. *Molecular and Cellular Neuroscience,* 29**,** 580-590.

Scherag, A., Dina, C., Hinney, A., Vatin, V., Scherag, S., Vogel, C. I., Muller, T. D., Grallert, H., Wichmann, H. E.&Balkau, B. (2010). Two new Loci for body-weight regulation identified in a joint analysis of genome-wide association studies for early-onset extreme obesity in French and german study groups. *PLoS Genet,* 6**,** e1000916.

Schlebusch, C. M., Skoglund, P., Sjödin, P., Gattepaille, L. M., Hernandez, D., Jay, F., Li, S., De Jongh, M., Singleton, A.&Blum, M. G. B. (2012). Genomic variation in seven

Khoe-San groups reveals adaptation and complex African history. *Science,* 338**,** 374-379.

Schleinitz, D., Klöting, N., Lindgren, C., Breitfeld, J., Dietrich, A., Schön, M., Lohmann, T., Dressler, M., Stumvoll, M.&Mccarthy, M. (2014). Fat depot-specific mRNA expression of novel loci associated with waist–hip ratio. *Int J Obes (Lond),* 38**,** 120-125.

Schmid, P. M., Heid, I., Buechler, C., Steege, A., Resch, M., Birner, C., Endemann, D. H., Riegger, G. A.&Luchner, A. (2012). Expression of fourteen novel obesity-related genes in zucker diabetic fatty rats. *Cardiovasc Diabetol,* 11**,** 48.

Shisana, O., Labadarios, D., Rehle, T., Simbayi, L., Zuma, K., Dhansay, A., Reddy, P., Parker, W., Hoosain, E.&Naidoo, P. (2013). SANHANES-1 Team (2013) South African National Health and Nutrition Examination Survey. *Cape Town: HSRC Press.*

Shungin, D., Winkler, T. W., Croteau-Chonka, D. C., Ferreira, T., Locke, A. E., Mägi, R., Strawbridge, R. J., Pers, T. H., Fischer, K.&Justice, A. E. (2015). New genetic loci link adipose and insulin biology to body fat distribution. *Nature,* 518**,** 187-196.

Silveira, P. P., Portella, A. K., Goldani, M. Z.&Barbieri, M. A. (2007). Developmental origins of health and disease (DOHaD). *Jornal de pediatria,* 83**,** 494-504.

Sim, X., Ong, R. T.-H., Suo, C., Tay, W.-T., Liu, J., Ng, D. P.-K., Boehnke, M., Chia, K.-S., Wong, T.-Y.&Seielstad, M. (2011). Transferability of type 2 diabetes implicated loci in multi-ethnic cohorts from Southeast Asia. *PLoS Genet,* 7**,** e1001363.

Singh, A. S., Mulder, C., Twisk, J. W. R., Van Mechelen, W.&Chinapaw, M. J. M. (2008). Tracking of childhood overweight into adulthood: a systematic review of the literature. *Obesity Reviews,* 9**,** 474-488.

Smemo, S., Tena, J. J., Kim, K.-H., Gamazon, E. R., Sakabe, N. J., Gómez-Marín, C., Aneas, I., Credidio, F. L., Sobreira, D. R.&Wasserman, N. F. (2014). Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature*.

Smith, C. J.&Ryckman, K. K. (2015). Epigenetic and developmental influences on the risk of obesity, diabetes, and metabolic syndrome. *Diabetes, metabolic syndrome and obesity: targets and therapy,* 8**,** 295.

Song, Y., Hsu, Y. H., Niu, T., Manson, J. E., Buring, J. E.&Liu, S. (2009). Common genetic variants of the ion channel transient receptor potential membrane melastatin 6 and 7 (TRPM6 and TRPM7), magnesium intake, and risk of type 2 diabetes in women. *BMC Med Genet,* 10**,** 4.

Speakman, J. R. (2013). Functional Analysis of Seven Genes Linked to Body Mass Index and Adiposity by Genome-Wide Association Studies: A Review. *Human heredity,* 75**,** 57-79.

Speakman, J. R.&O'rahilly, S. (2012). Fat: an evolving issue. *Disease models & mechanisms,* 5**,** 569-573.

Speliotes, E. K., Willer, C. J., Berndt, S. I., Monda, K. L., Thorleifsson, G., Jackson, A. U., Lango Allen, H., Lindgren, C. M., Luan, J., Magi, R., et al. (2010). Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat Genet,* 42**,** 937-48.

Speliotes, E. K., Yerges-Armstrong, L. M., Wu, J., Hernaez, R., Kim, L. J., Palmer, C. D., Gudnason, V., Eiriksdottir, G., Garcia, M. E.&Launer, L. J. (2011). Genome-wide association analysis identifies variants associated with nonalcoholic fatty liver disease that have distinct effects on metabolic traits. *PLoS Genet,* 7**,** e1001324.

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

Spencer, K. L., Malinowski, J., Carty, C. L., Franceschini, N., Fernandez-Rhodes, L., Young, A., Cheng, I., Ritchie, M. D., Haiman, C. A., Wilkens, L., et al. (2013). Genetic variation and reproductive timing: African American women from the Population Architecture using Genomics and Epidemiology (PAGE) Study. *PLoS One,* 8**,** e55258.

Sung, Y. J., Pérusse, L., Sarzynski, M. A., Fornage, M., Sidney, S., Sternfeld, B., Rice, T., Terry, J. G., Jacobs, D. R.&Katzmarzyk, P. (2015). Genome-wide association studies

suggest sex-specific loci associated with abdominal and visceral fat. *Int J Obes (Lond)*.
Tan, L. J., Zhu, H., He, H., Wu, K. H., Li, J., Chen, X. D., Zhang, J. G., Shen, H., Tian, Q., Krousel-Wood, M., et al. (2014). Replication of 6 obesity genes in a meta-analysis of genome-wide association studies from diverse ancestries. *PLoS One,* 9**,** e96149.
Tan, S. C.&Yiap, B. C. (2009). DNA, RNA, and protein extraction: the past and the present. *BioMed research international,* 2009.
Tekola-Ayele, F., Adeyemo, A., Chen, G., Hailu, E., Aseffa, A., Davey, G., Newport, M. J.&Rotimi, C. N. (2015). Novel genomic signals of recent selection in an Ethiopian population. *European Journal of Human Genetics,* 23**,** 1085-1092.
Tekola-Ayele, F., Adeyemo, A. A.&Rotimi, C. N. (2013). Genetic epidemiology of type 2 diabetes and cardiovascular diseases in Africa. *Progress in cardiovascular diseases,* 56**,** 251-260.
Teo, Y. Y., Small, K. S.&Kwiatkowski, D. P. (2010). Methodological challenges of genome-wide association analysis in Africa. *Nat Rev Genet,* 11**,** 149-60.
Thandrayen, K., Norris, S., Micklesfield, L.&Pettifor, J. (2014). Fracture patterns and bone mass in South African adolescent–mother pairs: the Birth to Twenty cohort. *Osteoporosis International,* 25**,** 693-700.
Thompson, E. A. (2013). Identity by descent: variation in meiosis, across genomes, and in populations. *Genetics,* 194**,** 301-326.
Thorleifsson, G., Walters, G. B., Gudbjartsson, D. F., Steinthorsdottir, V., Sulem, P., Helgadottir, A., Styrkarsdottir, U., Gretarsdottir, S., Thorlacius, S.&Jonsdottir, I. (2009). Genome-wide association yields new sequence variants at seven loci that associate with measures of obesity. *Nat Genet,* 41**,** 18-24.
Tollman, S. M., Kahn, K., Sartorius, B., Collinson, M. A., Clark, S. J.&Garenne, M. L. (2008). Implications of mortality transition for primary health care in rural South Africa: a population-based surveillance study. *Lancet,,* 372**,** 893-901.
Trégouët, D.-A., König, I. R., Erdmann, J., Munteanu, A., Braund, P. S., Hall, A. S., Großhennig, A., Linsel-Nitschke, P., Perret, C.&Desuremain, M. (2009). Genome-wide haplotype association study identifies the SLC22A3-LPAL2-LPA gene cluster as a risk locus for coronary artery disease. *Nat Genet,* 41**,** 283-285.
Tsai, F.&Coyle, W. J. (2009). The microbiome and obesity: is obesity linked to our gut flora? *Current gastroenterology reports,* 11**,** 307-313.
Van Asselt, K. M., Kok, H. S., Van Der Schouw, Y. T., Peeters, P. H., Pearson, P. L.&Grobbee, D. E. (2006). Role of genetic analyses in cardiology part II: heritability estimation for gene searching in multifactorial diseases. *Circulation,* 113**,** 1136-1139.
Van Dijk, S. J., Molloy, P. L., Varinli, H., Morrison, J. L., Muhlhausler, B. S., Buckley, M., Clark, S. J., Mcmillen, I. C., Noakes, M.&Samaras, K. (2015). Epigenetics and human obesity. *Int J Obes (Lond),* 39**,** 85-97.
Vasan, S. K., Fall, T., Job, V., Gu, H. F., Ingelsson, E., Brismar, K., Karpe, F.&Thomas, N. (2013). A common variant in the FTO locus is associated with waist–hip ratio in Indian adolescents. *Pediatric obesity,* 8**,** e45-e49.
Vattikuti, S., Guo, J.&Chow, C. C. (2012). Heritability and genetic correlations explained by common SNPs for metabolic syndrome traits. *PLoS Genet,* 8**,** e1002637.
Visscher, P. M. (2008). Sizing up human height variation. *Nat Genet,* 40**,** 489-490.
Visscher, P. M., Brown, M. A., Mccarthy, M. I.&Yang, J. (2012). Five years of GWAS discovery. *Am J Hum Genet,* 90**,** 7-24.
Visscher, P. M., Hill, W. G.&Wray, N. R. (2008). Heritability in the genomics era—concepts and misconceptions. *Nature Reviews Genetics,* 9**,** 255-266.
Visscher, P. M., Yang, J.&Goddard, M. E. (2010). A commentary on 'common SNPs explain a large proportion of the heritability for human height'by Yang et al.(2010). *Twin Research and Human Genetics,* 13**,** 517-524.
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. (2012). The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. *PLoS Genet,* 8**,** e1002793.

Voight, B. F., Scott, L. J., Steinthorsdottir, V., Morris, A. P., Dina, C., Welch, R. P., Zeggini, E., Huth, C., Aulchenko, Y. S.&Thorleifsson, G. (2010). Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nat Genet,* 42**,** 579-589.

Volckmar, A.-L., Song, J.-Y., Jarick, I., Pütter, C., Göbel, M., Horn, L., Struve, C., Haas, K., Knoll, N.&Grallert, H. (2015). Fine Mapping of a GWAS-Derived Obesity Candidate Region on Chromosome 16p11. 2. *PLoS One,* 10.

Vuillaume, M. L., Naudion, S., Banneau, G., Diene, G., Cartault, A., Cailley, D., Bouron, J., Toutain, J., Bourrouillou, G.&Vigouroux, A. (2014). New candidate loci identified by array‐CGH in a cohort of 100 children presenting with syndromic obesity. *American Journal of Medical Genetics Part A,* 164**,** 1965-1975.

Walley, A. J., Asher, J. E.&Froguel, P. (2009). The genetic contribution to non-syndromic human obesity. *Nat Rev Genet,* 10**,** 431-42.

Wang, Z.-M., Pierson, R. N.&Heymsfield, S. B. (1992). The five-level model: a new approach to organizing body-composition research. *Am J Clin Nutr,* 56**,** 19-28.

Wardle, J., Carnell, S., Haworth, C., M.&Plomin, R. (2008). Evidence for a strong genetic ingluence on childhood adiposity depsite the force of the obesogenic environment. *Am. J. Clin Nutr.,* 87**,** 398-404.

Wells, J. C. (2014). Commentary: The paradox of body mass index in obesity assessment: not a good index of adiposity, but not a bad index of cardio-metabolic risk. *International journal of epidemiology,* 43**,** 672-674.

Wen, W., Cho, Y.-S., Zheng, W., Dorajoo, R., Kato, N., Qi, L., Chen, C.-H., Delahanty, R. J., Okada, Y.&Tabara, Y. (2012). Meta-analysis identifies common variants associated with body mass index in east Asians. *Nat Genet,* 44**,** 307-311.

Wheeler, E., Huang, N., Bochukova, E. G., Keogh, J. M., Lindsay, S., Garg, S., Henning, E., Blackburn, H., Loos, R. J., Wareham, N. J., et al. (2013). Genome-wide SNP and CNV analysis identifies common and low-frequency variants associated with severe early-onset obesity. *Nat Genet,* 45**,** 513-7.

White, J., Sofat, R., Hemani, G., Shah, T., Engmann, J., Dale, C., Shah, S., Kruger, F. A., Giambartolomei, C.&Swerdlow, D. I. (2016). Plasma urate concentration and risk of coronary heart disease: a Mendelian randomisation analysis. *The Lancet Diabetes & Endocrinology*.

WHO (1995). Physical Status: The Use and Interpretation of Anthropometry. *World Health Organisation Technical Report Series.,* 854.

WHO (2008 (b)). WHO STEPwise approach to surveillance. *STEPS*.

WHO (2011). Waist circumference and waist-hip ratio. *Report of a WHO Expert Consultation.*

Willer, C. J., Speliotes, E. K., Loos, R. J. F., Li, S., Lindgren, C. M., Heid, I. M., Berndt, S. I., Elliott, A. L., Jackson, A. U.&Lamina, C. (2009). Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nat Genet,* 41**,** 25-34.

Winkler, T. W., Justice, A. E., Graff, M., Barata, L., Feitosa, M. F., Chu, S., Czajkowski, J., Esko, T., Fall, T., Kilpelainen, T. O., et al. (2015). The Influence of Age and Sex on Genetic Associations with Adult Body Size and Shape: A Large-Scale Genome-Wide Interaction Study. *PLoS Genet,* 11**,** e1005378.

Wood, E. T., Stover, D. A., Ehret, C., Destro-Bisol, G., Spedini, G., Mcleod, H., Louie, L., Bamshad, M., Strassmann, B. I.&Soodyall, H. (2005). Contrasting patterns of Y chromosome and mtDNA variation in Africa: evidence for sex-biased demographic processes. *European Journal of Human Genetics,* 13**,** 867-876.

Wu, Y., Waite, L. L., Jackson, A. U., Sheu, W. H., Buyske, S., Absher, D., Arnett, D. K., Boerwinkle, E., Bonnycastle, L. L., Carty, C. L., et al. (2013). Trans-ethnic fine-mapping of lipid loci identifies population-specific signals and allelic heterogeneity that increases the trait variance explained. *PLoS Genet,* 9**,** e1003379.

Xia, Q.&Grant, S. F. A. (2013). The genetics of human obesity. *Ann N Y Acad Sci,* 1281**,** 178-190.

Yako, Y. Y., Echouffo‐Tcheugui, J. B., Balti, E. V., Matsha, T. E., Sobngwi, E., Erasmus, R. T.&Kengne, A. P. (2015). Genetic association studies of obesity in Africa: a systematic review. *Obesity Reviews,* 16**,** 259-272.

Yang, J., Benyamin, B., Mcevoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G.&Montgomery, G. W. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nat Genet,* 42**,** 565-569.
Yang, J., Lee, S. H., Goddard, M. E.&Visscher, P. M. (2011a). GCTA: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics,* 88**,** 76-82.
Yang, J., Manolio, T. A., Pasquale, L. R., Boerwinkle, E., Caporaso, N., Cunningham, J. M., De Andrade, M., Feenstra, B., Feingold, E.&Hayes, M. G. (2011b). Genome partitioning of genetic variation for complex traits using common SNPs. *Nat Genet,* 43**,** 519-525.
Yang, J., Weedon, M. N., Purcell, S., Lettre, G., Estrada, K., Willer, C. J., Smith, A. V., Ingelsson, E., O'connell, J. R.&Mangino, M. (2011c). Genomic inflation factors under polygenic inheritance. *European Journal of Human Genetics,* 19**,** 807-812.
Yazdi, F. T., Clee, S. M.&Meyre, D. (2015). Obesity genetics in mouse and human: back and forth, and back again. *PeerJ,* 3**,** e856.
Zaitlen, N., Kraft, P., Patterson, N., Pasaniuc, B., Bhatia, G., Pollack, S.&Price, A. L. (2013). Using extended genealogy to estimate components of heritability for 23 quantitative and dichotomous traits. *PLoS Genet,* 9**,** e1003520.
Zegers, D., Van Hul, W., Van Gaal, L. F.&Beckers, S. (2012). Monogenic and complex forms of obesity: insights from genetics reveal the leptin-melanocortin signaling pathway as a common player. *Crit Rev Eukaryot Gene Expr,* 22**,** 325-43.
Zhao, J., Bradfield, J. P., Li, M., Wang, K., Zhang, H., Kim, C. E., Annaiah, K., Glessner, J. T., Thomas, K.& Garris, M. (2009). The Role of Obesity‐associated Loci Identified in Genome‐wide Association Studies in the Determination of Pediatric BMI. *Obesity (Silver Spring),* 17**,** 2254-2257.
Zhou, X. & Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nat Genet,* 44**,** 821-4.

Electronic references

E-Resource2 https://en.wikipedia.org/wiki/Epistasis.

E-Resource for Figure 1.3
https://www.measureup.com.au/media/docs/female_body_composition_example

Genesis software for drawing PC plots: http://www.bioinf.wits.ac.za/software/genesis/

GWAS/Power plot software: (http://www.mybiosoftware.com/gwaspowerqt-1-0-statistical-power-calculation-software-designed-gwas.html)

Appendices

Appendix A: Power analysis

The power plots were drawn using the programs PowerGWAS/QT (http://www.mybiosoftware.com/gwaspowerqt-1-0-statistical-power-calculation-software-designed-gwas.html) (Feng et al., 2011). This program is commonly used with GWAS when using quantitative traits. All power plots were drawn by Ananyo Choudury (Bioinformatician at SBIMB).

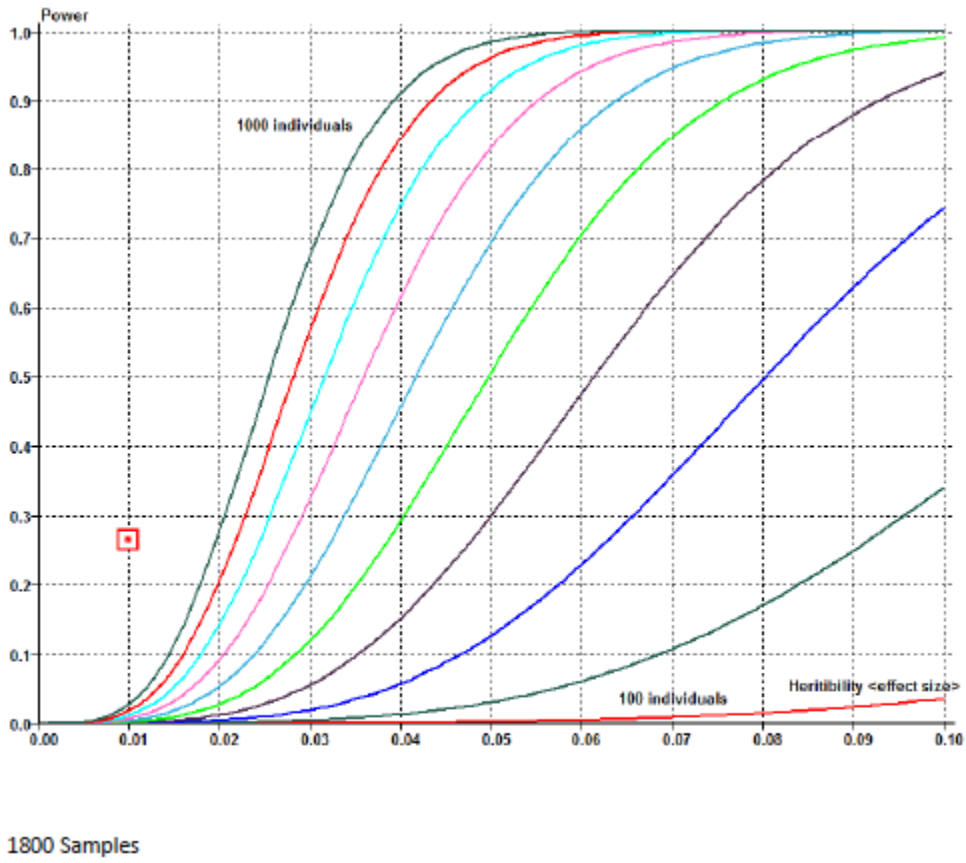


Figure 1: Power plot drawn with 1800 individuals shows that in 1800 individuals we have 25% power to capture an effect size of Beta = 0.01.

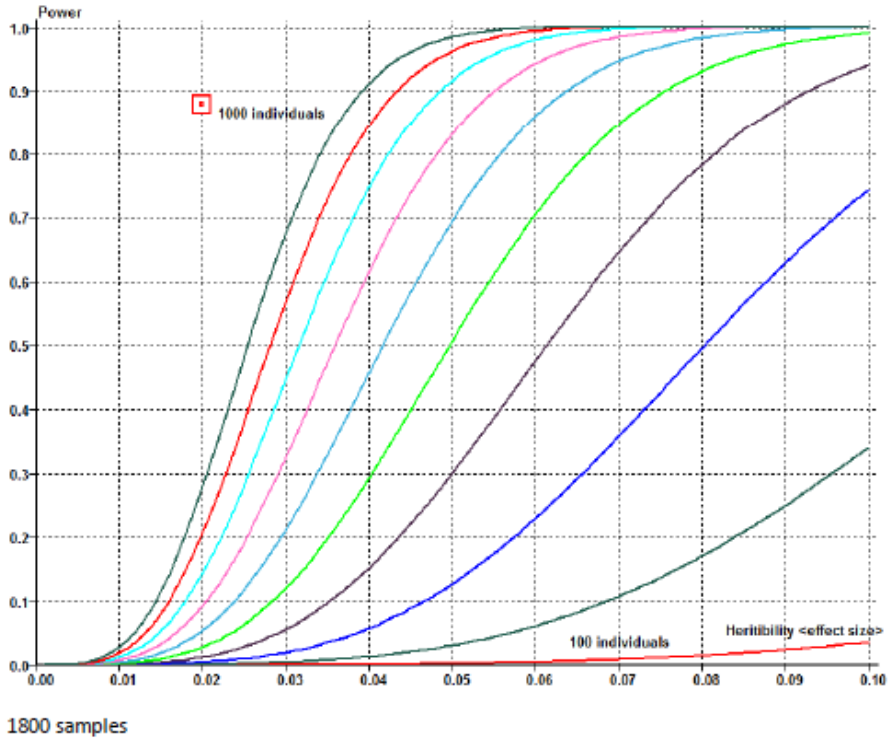


Figure 2: Power plot drawn using 1800 individuals shows that in 1800 individuals we have ~90% power to capture an effect size of Beta = 0.02

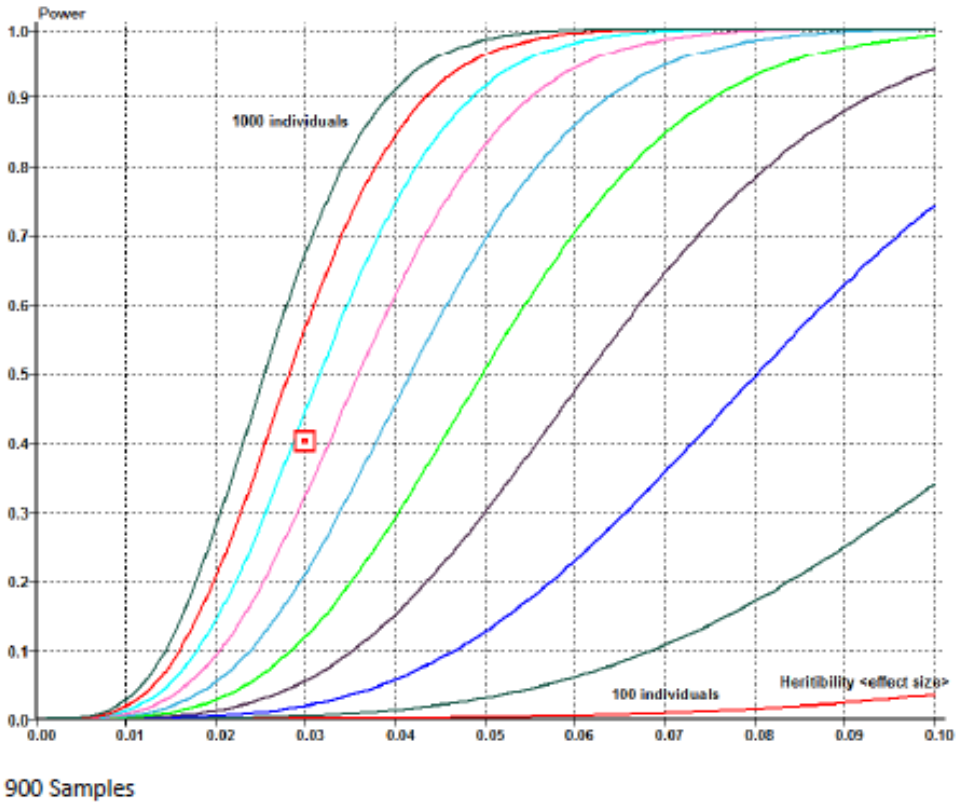


Figure 3: Power plot drawn using 900 individuals shows that ~55% power to capture an effect size of Beta = 0.03 in 900 individuals

APPENDIX B: PLINK commands used during quality control process

All QC was done together with Liesl Hendry for quality assurance purposes.

Pre-QC step: Conversion of final reports in forward format into tped/tfam format, check and then conversion to bed files

1. Convert all the forward report files into one tped and tfam format files for each batch
    a. Converting raw genotypes to tped format (batch* - refers to either batch 1 or 2)

Convert to tped.py

```
#!/usr/local/bin/python
import os

# Arrays
snpdata = { }
csvfiles = [ ]
tfamarray = [ ]
tpedarray = { }
tfamfilename = "batch*.tfam"
tpedfilename = "batch*.tped"

# Create a list of CSV files
files=os.listdir("./")
for eachfile in files:
   if eachfile[-3:] == "csv":
         csvfiles.append(eachfile)
# csvfiles =
["Metabo_Batch2_Plate_11_12_FinalReport_Forward.csv","Metabo_Batc
h2_Plate_3_4_FinalReport_Forward.csv","Metabo_Batch2_Plate_9_10_F
inalReport_Forward.csv","Metabo_Batch2_Plate_1_2_FinalReport_Forw
ard.csv","Metabo_Batch2_Plate_5_6_FinalReport_Forward.csv","Metab
o_Batch2_Plate13_FinalReport_Forward.csv","MetaboBatch2_Plate_7_8
_FinalReport_Forward.csv"]

# 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
```

2. Check a few genotypes against samples in the report file to see that it worked.

3. `plink --tfile batch* --extract snplist.txt --recode --out extracted_1`

4. And then check manually in excel

5. Repeat 1 and 2 for both batches

6. Convert tped/tfam to bed file

```
plink --tfile batch* --make-bed --out batch* --noweb
```

Removal of “NaN” SNPs

1. Made a text file with the ID’s of samples containing the ‘NaN’ SNPs

Removed SNPs with the command

```
plink --bfile batch* --exclude NaNsnps.txt --make-bed --out
NoNanbatch*
```

QC OF BATCHES SEPARATELY

Initial removal of poorly genotyped samples

* - refers to either batch 1 or 2.

```
plink --noweb --bfile  NoNanbatch* --mind 0.20 --make-bed --out
NoNanbatch*_mind_0.2 --allow-no-sex
```

SNP QC

- Missingness

A range of 'geno' thresholds were tested

```
plink --bfile  NoNanbatch*_mind_0.2  --geno 0.10 --out
NoNanbatch*_mind_0.2_snp90
plink --bfile  NoNanbatch*_mind_0.2  --geno 0.08 --out
NoNanbatch*_mind_0.2_snp92
plink --bfile  NoNanbatch*_mind_0.2  --geno 0.06 --out
NoNanbatch*_mind_0.2_snp94
plink --bfile  NoNanbatch*_mind_0.2  --geno 0.05 --out
NoNanbatch*_mind_0.2_snp95
plink --bfile  NoNanbatch*_mind_0.2  --geno 0.04 --out
NoNanbatch*_mind_0.2_snp96
plink --bfile  NoNanbatch*_mind_0.2  --geno 0.03 --out
NoNanbatch*_mind_0.2_snp97
plink --bfile  NoNanbatch*_mind_0.2  --geno 0.02 --out
NoNanbatch*_mind_0.2_snp98
plink --bfile  NoNanbatch*_mind_0.2  --geno 0.01 --out
NoNanbatch*_mind_0.2_snp99
plink --bfile  NoNanbatch*_mind_0.2  --geno 0.009 --out
NoNanbatch*_mind_0.2_snp99.1
plink --bfile  NoNanbatch*_mind_0.2  --geno 0.008 --out
NoNanbatch*_mind_0.2_snp99.2
plink --bfile  NoNanbatch*_mind_0.2  --geno 0.007 --out
NoNanbatch*_mind_0.2_snp99.3
```

Batch 1: Based on our plot, we decide on a 0.02 (2%) threshold for SNP

- MAF

A range of MAFs were tested to decide on the threshold.

```
plink --bfile  NoNanbatch*_mind_0.2 --maf 0.05 --out
NoNanbatch*_mind_0.2_0.05
plink --bfile  NoNanbatch*_mind_0.2 --maf 0.04 --out
NoNanbatch*_mind_0.2_0.04
plink --bfile  NoNanbatch*_mind_0.2 -- maf 0.03 --out
NoNanbatch*_mind_0.2_0.03
plink --bfile  NoNanbatch*_mind_0.2 -- maf 0.02 --out
NoNanbatch*_mind_0.2_0.02
plink --bfile  NoNanbatch*_mind_0.2 -- maf 0.01 --out
NoNanbatch*_mind_0.2_0.01
plink --bfile  NoNanbatch*_mind_0.2 -- maf 0.009 --out
NoNanbatch*_mind_0.2_0.009
plink --bfile  NoNanbatch*_mind_0.2 -- maf 0.008 --out
NoNanbatch*_mind_0.2_0.008
plink --bfile  NoNanbatch*_mind_0.2 -- maf 0.006 --out
NoNanbatch*_mind_0.2_0.006
plink --bfile  NoNanbatch*_mind_0.2 -- maf 0.004 --out
NoNanbatch*_mind_0.2_0.004
plink --bfile  NoNanbatch*_mind_0.2 -- maf 0.003 --out
NoNanbatch*_mind_0.2_0.003
plink --bfile  NoNanbatch*_mind_0.2 -- maf 0.002 --out
NoNanbatch*_mind_0.2_0.002
plink --bfile  NoNanbatch*_mind_0.2 -- maf 0.001 --out
NoNanbatch*_mind_0.2_0.001
```

- HWE

Tested a range of theholds to decide on the the level that didn't loose the greatest amount of SNPs

```
plink --bfile  NoNanbatch*_mind_0.2 --hwe 0.00000001 -
outNoNanbatch*_mind_0.2_10_8
```

```
plink --bfile  NoNanbatch*_mind_0.2 --hwe 0.0000001 --out
NoNanbatch*_mind_0.2_10_7
plink --bfile  NoNanbatch*_mind_0.2 --hwe 0.000001 --out
NoNanbatch*_mind_0.2_10_6
plink --bfile  NoNanbatch*_mind_0.2 --hwe 0.00001 --out
NoNanbatch*_mind_0.2_10_5
plink --bfile  NoNanbatch*_mind_0.2 --hwe 0.0001 --out
NoNanbatch*_mind_0.2_10_4
plink --bfile  NoNanbatch*_mind_0.2 --hwe 0.001 --out
NoNanbatch*_mind_0.2_10_3
```

Once the thresholds were selected the final datasets could then be filtered accordingly:

```
plink --noweb --bfile  NoNanbatch1_mind_0.2  --geno 0.02 --maf 0.01-
-hwe 0.00001 --make-bed --out batch1_snp_qc2 --allow-no-sex

plink --noweb --bfile  NoNanbatch2_mind_0.2  --geno 0.02 --maf 0.01-
-hwe 0.00001 --make-bed --out batch2_snp_qc2 --allow-no-sex
```

SAMPLE QC

- Missingness

Tested a range of “mind” thresholds to plot graph to decide on our threshold for removal of individuals (did this on the SNP QC filtered data/bed files).

```
plink --noweb --bfile  batch*_snp_qc2 --mind 0.10 --out
batch*_snp_qc2_ind90
plink --noweb --bfile  batch*_snp_qc2 --mind 0.08 --out
batch*_snp_qc2_ind92
plink --noweb --bfile  batch*_snp_qc2 --mind 0.06 --out
batch*_snp_qc2_ind94
plink --noweb --bfile  batch*_snp_qc2 --mind 0.05 --out
batch*_snp_qc2_ind95
plink --noweb --bfile  batch*_snp_qc2 --mind 0.04 --out
batch*_snp_qc2_ind96
plink --noweb --bfile  batch*_snp_qc2 --mind 0.03 --out
batch*_snp_qc2_ind97
```

```
plink --noweb --bfile  batch*_snp_qc2 --mind 0.02 --out
batch*_snp_qc2_ind98
plink --noweb --bfile  batch*_snp_qc2 --mind 0.01 --out
batch*_snp_qc2_ind99
plink --noweb --bfile  batch*_snp_qc2 --mind 0.009 --out
batch*_snp_qc2_ind99.1
plink --noweb --bfile  batch*_snp_qc2 --mind 0.008 --out
batch*_snp_qc2 _ind99.2
```

Filtering based on chosen thresholds:

```
plink --noweb --bfile  batch1_snp_qc2 --mind 0.02 --make-bed --out
batch1_clean --allow-no-sex

plink --noweb --bfile  batch2_snp_qc2 --mind 0.03 --make-bed --out
batch2_clean --allow-no-sex
```

- Sex check

Run on raw data

```
plink --bfile batch* --check-sex --out batch *
```

Ouput file = batch1.sexcheck

In excel look at last column (homozygocity rate: <0.2/0.35=female; >0.8=male; 0.2/0.35-0.8=inconclusive/unspecified)
Column 3 – ascertained sex
Column 4 – sex according to genotype data (1-male; 2-female)

Batch 1:
Females (according to genotyping) = 999
25 problem individuals (0.35 or greater). Of the 25, 23 are definite males and 2 are undetermined. Of these, 24 individuals have not yet been removed in previous QC steps and are thus removed at this point.

Save 24 individuals (as Family ID/Individual ID pairs) in text file called sexremoved_batch1

```
plink --bfile batch1_clean --remove sexremoved_batch1.txt --noweb --
make-bed --out batch1_clean2
```

Individuals remaining (batch 1): 1000 (999 plus 1 duplicate)

Batch 2:
Females (according to genotyping) = 629
Unspecified (according to genotyping) =31
Male (according to genotyping) = 588

6 missing phenotype sexes – apply sex values according to genotypes (as values are definite males and females)
1 possible male (already removed in previous QC step)
4 undetermined sexes according to genotype (already removed in previous QC step)
91 discrepancies (of these 65 have already been removed in previous QC steps and 26 have not yet been removed in previous QC steps and are thus removed at this point)

Save 26 individuals (as Family ID/Individual ID pairs) in text file called sexremoved_batch2

```
plink --bfile batch2_clean --remove sexremoved_batch2.txt --noweb --
make-bed --out batch2_clean2
```

- PCA analysis

PCA no. 1

Prune dataset to remove highly correlated variants

```
plink --bfile batch*_clean2 --indep-pairwise 50 5 0.2 --noweb --out
ldpruned2_batch*
plink --bfile batch*_clean2 --extract ldpruned2_batch*.prune.in --
make-bed --out batch*_ldpruned2
```

Batch 1: 59601 SNPs remaining after pruning (these will be used to run PCA and IBD)
Batch 2: 55508 SNPs remaining after pruning (these will be used to run PCA and IBD)

Running PCA

(1) Make a “phenotype” file from the fam file: column 1=FamID, column 2=IndID, column 3=population (in our case Bt20), column 4=batch* [Batch*_clean2_phenotype.txt]

(2) Make a *queuerunpca.sh* script (Batch 1 – queuerunpca1_batch1.sh; Batch 2 – queuerunpca1_batch2.sh)

```
#!/bin/bash
#PBS -N Batch*_pca
#PBS -q WitsLong
#PBS -l walltime=24:00:00,mem=10GB

cd /home/venesa/Metaboanalysis_30.01/Batch*_pca1

./runpca.sh batch*_ldpruned2
```

Save it in the working directory (venesa)
Put the following into the Batch*pca1 directory:
bed, bim and fam files (`batch*_ldpruned2.bed, batch*_ ldpruned2.bim, batch*_ldpruned2.fam`)
Phenotype file (Batch*_clean2_phenotype.txt)
runpca.sh script

```
#! /bin/bash

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
```

(3) Make the runpca.sh and queuerunpca.sh scripts executable.

```
chmod 733 runpca.sh
chmod 733 queuerunpca1_batch1.sh
chmod 733 queuerunpca1_batch2.sh
```

(4) Run the jobs

```
qsub queuerunpca1_batch1.sh
qsub queuerunpca1_batch2.sh
```

Defined new cut-offs for PCA analysis

- After removing individuals for heterozygosity we had 973 (B1) and 987 in B2
- We then removed 9 samples from PCA analysis, B1 and 4 samples from PCA analysis B2, after PCA
- Left with 974-B1 and 983 in B2

Admixture – PCA plots

PCA plot with ldpruned4 (i.e. no other populations included)

Make new ldpruned files

```
plink --bfile batch*_clean4 --indep-pairwise 50 5 0.2 --noweb --out
ldpruned4_batch*
plink --bfile batch*_clean4 --extract ldpruned4_batch*.prune.in --
make-bed --noweb --out batch*_ldpruned4
```

Batch 1: 59605 SNPs remaining after pruning
Batch 2: 55567 SNPs remaining after pruning

Folders:
Batch1_clean4 and Batch2_clean4

Running PCA

(1) Make a "phenotype" file from the fam file: column 1=FamID, column 2=IndID, column 3=population (in our case Bt20B*) [batch*_ldpruned4_phenotype.txt]
(2) Make a *queuerunpca.sh script* (Batch * – queuerunpca_batch*_ldpruned4.sh)

```
#!/bin/bash
#PBS -N Batch*_pca_ldpruned4
#PBS -q WitsLong
#PBS -l walltime=24:00:00,mem=10GB

cd /home/venesa/Metaboanalysis_30.01/Batch*_clean4

./runpca.sh batch*_ldpruned4
```

Save it in the working directory

(3) Put the following into the Batch*_clean4 directory:
bed, bim and fam files (batch*_ldpruned4.bed, batch*_ ldpruned4.bim, batch*_ldpruned4.fam)
Phenotype file (batch*_ldpruned4_phenotype.txt)
runpca.sh script

Make the scripts executable and run

Removing samples

To decide which samples to remove, we chose the following cut-offs:
Horizontal: -0.075<=x<=0.075
Vertical: -0.05<=x<=0.05

PCA plot with Africans (excluding coloureds and Khoisan)

Make new ldpruned files

```
plink --bfile batch*_clean5 --indep-pairwise 50 5 0.2 --noweb --out
ldpruned5_batch*
plink --bfile batch*_clean5 --extract ldpruned5_batch*.prune.in --
make-bed --noweb --out batch*_ldpruned5
```

Batch 1: 59601 SNPs remaining after pruning
Batch 2: 55585 SNPs remaining after pruning
Use the batch1_allafrican2 files

Make a list of Bt20 individuals from batch1_allafrican2 (batch1.txt) and remove them from these files.

```
plink --bfile batch*_allafrican2 --remove batch1.txt --noweb --make-
bed --out african_nokca
```

Now merge the new ldpruned5 files for each batch to this african_nokca file

```
plink --bfile batch*_ldpruned5 --bmerge african_nokca.bed
african_nokca.bim african_nokca.fam --noweb--make-bed --out batch*_
african_nokca
```

The first attempt at merging failed due to strand errors, therefore run the following command:

```
plink --bfile batch*_ldpruned5 --flip batch*_african_nokca.missnp --
make-bed --out batch*_flipped --noweb
```

Then try merging again:

```
plink --bfile batch*_flipped --bmerge african_nokca.bed
african_nokca.bim african_nokca.fam --make-bed --out batch*_
african_nokca2
```

Once data are merged, prune away those SNPs that do not overlap between datasets

plink --bfile batch*_african_nokca2 --geno 0.05 --make-bed --out batch*_ african_nokca3

Batch 1: 13532 SNPs remaining after merging and pruning
Batch 2: 12126 SNPs remaining after merging and pruning

Make a new phenotype file

Make from the fam files: column 1=FamID, column 2=IndID, column 3=population (in our case Bt20B1 or Bt20B2 and whatever the other populations are)

```
awk '{print $1,$2}' batch*_african_nokca3.fam >
batch*_african_nokca3_phenotype.txt
```

(This maintains the order of the study participants in the fam file)

Then add the phenotype column in excel

*Make new folders: Batch*_african_nokca3*

Running PCA

(1) Make a *queuerunpca.sh* script (queuerunpca_batch*_african_nokca3.sh)

```
#!/bin/bash
#PBS -N Batch*_pca_african_nokca3
#PBS -q WitsLong
#PBS -l walltime=24:00:00,mem=10GB

cd /home/venesa/Metaboanalysis_30.01/Batch*_african_nokca3

./runpca.sh batch*_ african_nokca3
```

Save it in the working directory (venesa or wherever)

(2) Put the following into the Batch*_ african_nokca3 directory:
bed, bim and fam files (batch*_african_nokca3.bed, batch*_african_nokca3.bim, batch*_african_nokca3.fam)
Phenotype file (batch*_ african_nokca3_phenotype.txt)
runpca.sh script

Make the scripts executable and run the jobs

Batch1: 0 further individuals to remove
Batch 1: 0 further individuals to remove)

<u>Global PCA (all populations incl. Khoisan and coloureds)</u>

In folder called batch*_pca_global

Use batch*_ldpruned5 files to merge with Andrew's pruned data files (allthin)

```
plink --bfile batch*_ldpruned5 --bmerge allthin.bed allthin.bim
allthin.fam --make-bed --out batch*_allthin**
```

The first attempt at merging failed due to strand errors, therefore run the following command:

```
plink --bfile batch*_ldpruned5 --flip batch*_allthin**.missnp --
make-bed --out batch*_flipped --noweb
```

Then try merging again:

```
plink --bfile batch*_flipped --bmerge allthin.bed allthin.bim
allthin.fam --make-bed --out batch*_allthin**
```

Once data are merged, prune away those SNPs that do not overlap between datasets

```
plink --bfile batch*_allthin** --geno 0.05 --make-bed --out
batch*_allthin4
```

Batch 1: 14010 SNPs remaining after merging and pruning
Batch 2: 13287 SNPs remaining after merging and pruning

Make a new phenotype file

Make from the fam files: column 1=FamID, column 2=IndID, column 3=population (in our case Bt20B1 or Bt20B2 and whatever the other populations are)

```
awk '{print $1,$2}'batch*_allthin4.fam >
batch*_allthin4_phenotype.txt
```

(This maintains the order of the study participants in the fam file)

Then add the phenotype column in excel
Running PCA

(1) Make a *queuerunpca.sh script* (queuerunpca_allthin4_batch*.sh)

```
#!/bin/bash
#PBS -N Batch*_pca_allthin4
#PBS -q WitsLong
#PBS -l walltime=24:00:00,mem=10GB

cd /home/venesa/Metaboanalysis_30.01/batch*_pca_global

./runpca.sh batch*_allthin4
```

Save it in the working directory (venesa or wherever)

(2) Put the following into the Batch*_pca_global directory:

bed, bim and fam files (batch*_allthin4.bed, batch*_ allthin4.bim, batch*_ allthin4.fam)
Phenotype file (batch*_ allthin4_phenotype.txt)
runpca.sh script

Make the scripts executable and run the jobs

Draw PCA plots.

Clustering generally good and what we expected (although coloured and Khoisan individuals make the plot quite messy). Didn't exclude any more individuals although some were slightly out of cluster so should be watched carefully.

Batch 1:
9412207

Batch 2:
3076321
9412207
3385402
7888045
3341858

- IBD

Run on pruned data set (Anderson et al.)

Four folders in main IBD folder (Batch1_pruned; Batch2_pruned)
In each of the four folders put the bed, bim and fam files (batch*_ldpruned2)

<u>Removing related individuals step 1</u>

```
plink --bfile batch*_ldpruned2 --genome --min 0.05 --noweb
```

[Change names of .genome file to name that appears in runibd.pl script and change name in script. Save script and make it executable.

Run the runibd.pl scripts for all for IBD runs (pruned and clean for batch 1 and batch 2) directly in the cluster.

```
perl runibd.pl > ibsdist_related_batch1_ldpruned2
perl runibd.pl > ibsdist_related_batch2_ldpruned2

runibd.pl script

#!/usr/bin/perl -w


open(IN,"batch*_ldpruned2.genome");
$fl=<IN>;$fl="";

while(<IN>){
$line=$_;
chomp($line);

@tabs=split(/\s+/,$line);

if($tabs[10] < 0.05 or $tabs[12]>0.95){

print $line,"\n";

}
}


################################
sub log10{
my $n=shift;
return log($n)/log(10);
}
####################################    ]
```

*Not necessary to run perl script. Alternatively assess genome file manually.

Sort by PI_HAT column and pick out pairs with PI_HAT value of >0.18 (Anderson et al.)

Batch1: 7 pairs (incl. 1 duplicate) with PI_HAT>0.18
Batch2: 13 pairs (incl. 6 pairs with PI_HAT of 1 that aren't duplicates) with PI_HAT>0.18

For Batch 2, the remaining 7 duplicates weren't picked up as being related. Therefore all of the remaining duplicates must be removed.

Deciding which individual of related pairs to remove

Draw the PCA plot (ldpruned2) for batch* in Genesis (using generated evec file and phenotype file)
We discovered the following:

A number of the samples from the IBD pairs weren't in the PCA plot. Looking at the PCA log file, we saw that a number of samples were removed during running the PCA as they were outliers.

For Batch 1: 13 outliers.
Of these 13, only 7 (6 in 3 pairs and one separate) overlapped with the IBD pairs (coloured in red). For the other 3 pairs, we looked to see which was the best individual to exclude based on the position on the PCA plot (coloured in green). There were 6 extra outliers (coloured in blue).
There are a total of 16 individuals that need to be removed at this point.

plink --bfile batch1_clean2 --remove ibd_excludedind_batch1.txt --noweb --make-bed --out batch1_clean3

For Batch 2: 52 outliers.
Of these 52, only 16 (16 in 8 pairs) overlapped with the IBD pairs (coloured in red). For the other 5 pairs, we looked to see which was the best individual to exclude based on the position on the PCA plot (coloured in green). There were 36 extra outliers (coloured in blue).
There are a total of 57 individuals that need to be removed at this point.

plink --bfile batch2_clean2 --remove ibd_excludedind_batch2.txt --noweb --make-bed --out batch2_clean3

Individuals remaining after IBD (Batch 1): 984
Individuals remaining after IBD (Batch 2): 1002

Re-prune dataset (clean 3) before running IBD again

```
plink --bfile batch*_clean3 --indep-pairwise 50 5 0.2 --noweb --out
ldpruned3_batch*
plink --bfile batch*_clean3 --extract ldpruned3_batch*.prune.in --
make-bed --out batch*_ldpruned3
```

Batch 1: 59607 SNPs remaining after pruning (these will be used to run IBD 2)
Batch 2: 55557 SNPs remaining after pruning (these will be used to run IBD 2)

Rerun of IBD (IBD 2)

Run on pruned data set (Anderson et al.)

Four folders in main IBD folder (Batch1_pruned; Batch2_pruned)
In each of the four folders put the bed, bim and fam files (batch*_ldpruned3)

```
plink --bfile batch*_ldpruned3 --genome --min 0.05 --noweb
```

Sort by PI_HAT column in .genome file and check for relatedness

Batch 1: PI_HAT<= 0.1619
Batch 2: PI_HAT<= 0.1687

Therefore no further samples need to be removed due to relatedness.

Now working in Batch*_clean4 folder.

- Duplicates

Removal of duplicates (7 pairs and 1 remaining of pair 8) for batch 2

Make list of all 8 duplicate Ids (twice for seven of them to remove both sets=15 individuals) – batch2_removedup.txt

- Outlying heterozygosity rates

Heterozygocity rate vs Proportion of missing SNPs plots (Identification of individuals with outlying heterozygosity rate)

Calculated missingness using NoNanBatch* files

```
plink --bfile NoNanbatch*--missing --out NoNanbatch*
```

The sixth column in the imiss file (F_MISS) gives the proportion of missing SNPs per individual.

```
plink --bfile NoNanbatch* --het --out NoNanbatch*
```

In the .het file, the third column gives the observed number of homozygous genotypes [O(Hom)] and the fifth column gives the number of non-missing genotypes [N(NM)], per individual.

Calclate the observed heterozygosity rate per individual using the formula (N(NM) - O(Hom))/N(NM).

Create a graph where the observed heterozygosity rate per individual is plotted on the x-axis and the proportion of missing SNPs per individuals is plotted on the y-axis.
Examine the plot to decide reasonable thresholds at which to exclude individuals based on extreme heterozygosity. We chose to exclude all individuals with a heterozygosity rate ± 3 standard deviations from the mean.
Batch 1: 8 individuals (7 excluded in previous steps)

0.237286028 Mean+ 3SD
0.181778524 Mean- 3SD

Batch 2: 52 individuals (all excluded in previous steps)

0.295526 Mean+3SD
0.15677 Mean- 3SD

Add the family ID and individual ID of the 1 sample in batch 1 to batch1_removedup.txt

Run remove commands (for duplicates and heterozygosity):

```
plink --bfile batch*_clean3 --remove batch*_removedup.txt --noweb --
make-bed --out batch*_clean4
```

Make GRM_2.sh

```
!/bin/bash

#PBS -N Heritability
#PBS -q WitsLong
#PBS -l walltime=20:00:00,mem=3GB
#PBS -l nodes=1:ppn=1

WORKING_DIR='/home/venesa/heritability'

#Change to the working directory where the files are.
cd $WORKING_DIR

# name of binary plink filename (excluding .bed/.bim/.fam suffix)
plinkfile="prunedbmerge"

# name of phenotype file in plink format (i.e. col1=fid, col2=iid, col3=phenotype
phenfile="sorted_combined_batches.phen"

# name of covariate file in plink format (i.e col1=fid, col2=iid, col3=covariate)
covarfile="combine_batches.covar"

# make up a name for the grm for all individuals
allfile="GRM_all"

# make up a name for the grm for related individuals
relatedsfile="GRM_relateds"

# maximum relatedness threshold
threshold=0.05

# make up a name for output file (it will have .hsq appended to the end by gcta)
outfile="bmi_snps_heritability"

## Alternative way to have multiple phenotypes
```

phenotype_col="8"

```
gcta64 \
      --bfile ${plinkfile} \
      --make-grm \
      --maf 0.01 \
      --out ${allfile}

gcta64 \
      --grm ${allfile} \
      --pca 10 \
      --out ${allfile}
```

MLMA run in GCTA: When using GCTA; the mpheno command on its own takes the 3rd column automatically as the column (in the phenotype file) being analysed. So running the command mpheno 2 takes the 4th column, mpheno 3 takes the 5th column and so on

Column number	**mpheno command (function)**
1- FID	
2-IID	
3-SEX (1-M; 2-FM)	
4-AGE (yrs.)	
5-HEIGHT (mm)	
6-HEIGHT_1 (m)	
7-WEIGHT (kg)	
8-BMI ($kg.m^{-2}$)	6
9-logBMI	7
10-WC (mm)	8
11-HC (mm)	9
12-WHR (index)	10
13-free fat mass (kg)	11
14-lean mass (kg)	12
15-percentage body fat	13

```
Model1a.covar: sex+age only
Model2a.covar: sex+age+height
Model3a.covar: sex+age+height+leanmass
```

```
Model4a.covar: sex+age+height +fatmass
```

mlma_2.sh

```
#!/bin/bash

#PBS -N MLMA
#PBS -q WitsLong
#PBS -l walltime=20:00:00,mem=3GB
#PBS -l nodes=1:ppn=1

WORKING_DIR='/home/venesa/heritability'

#Change to the working directory where the files are.
cd $WORKING_DIR

# name of binary plink filename (excluding .bed/.bim/.fam suffix)
plinkfile="prunedbmerge2"

# name of phenotype file in plink format (i.e. col1=fid, col2=iid, col3=phenotype
phenfile="sorted_combined_batches.phen"

# make up a name for the grm for related individuals
allfile="combinedall"

# covariate file (including the suffix.qcovar if quantitative, if sex then .covar)
covarfile1="model1a.qcovar"
covarfile2="model2a.qcovar"
covarfile3="model3a.qcovar"
covarfile4="model4a.qcovar"

# make up a name for output file (it will have .mmla appended to the end by gcta)
#outfile="test.mma"

## Alternative way to have multiple phenotypes (mpheno-normally column 4 but if you give a
column number after that its usually the position of the column
#say BMI is in column 8 (-2) will be mpheno 6))
#column="7"
```

```
#gcta64 \
#        --mlma
#        --bfile ${plinkfile} \
#        --grm ${allfile} \
#        --pheno ${phenfile} \
#        --mpheno ${column} \
#        --qcovar ${covarfile} \
#        --out ${outfile}


function the_function ()
{
    gcta64 \
        --mlma \
     --bfile ${plinkfile} \
        --grm ${allfile} \
        --pheno ${phenfile} \
        --mpheno "$1" \
        --qcovar $2 \
        --out $1_$2
}

the_function 7 $covarfile1
the_function 8 $covarfile1
the_function 8 $covarfile2
the_function 9 $covarfile1
the_function 9 $covarfile2
the_function 10 $covarfile1
the_function 11 $covarfile1
the_function 11 $covarfile2
the_function 11 $covarfile3
the_function 12 $covarfile1
the_function 12 $covarfile2
the_function 12 $covarfile4
the_function 13 $covarfile1
the_function 13 $covarfile2
```

Modified mlma script for combined females:

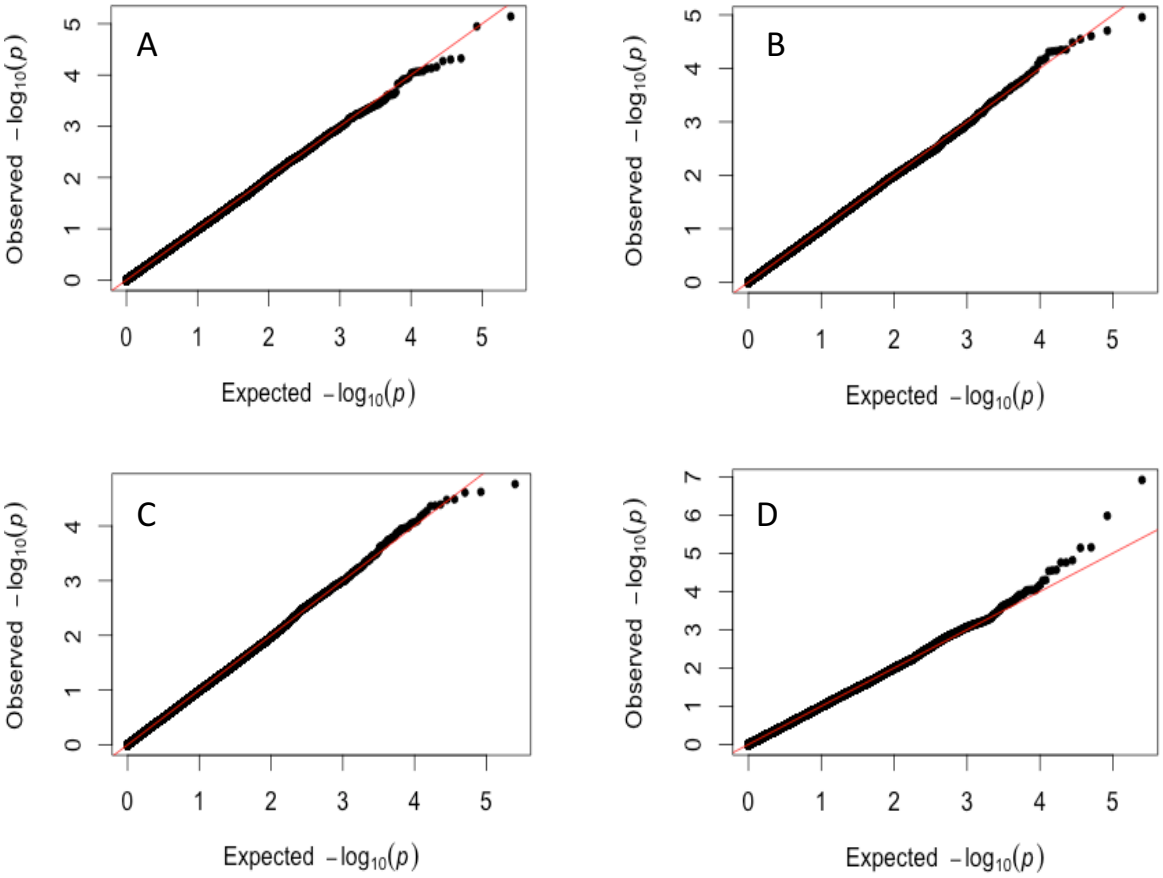
mlma_combinedfemales.sh

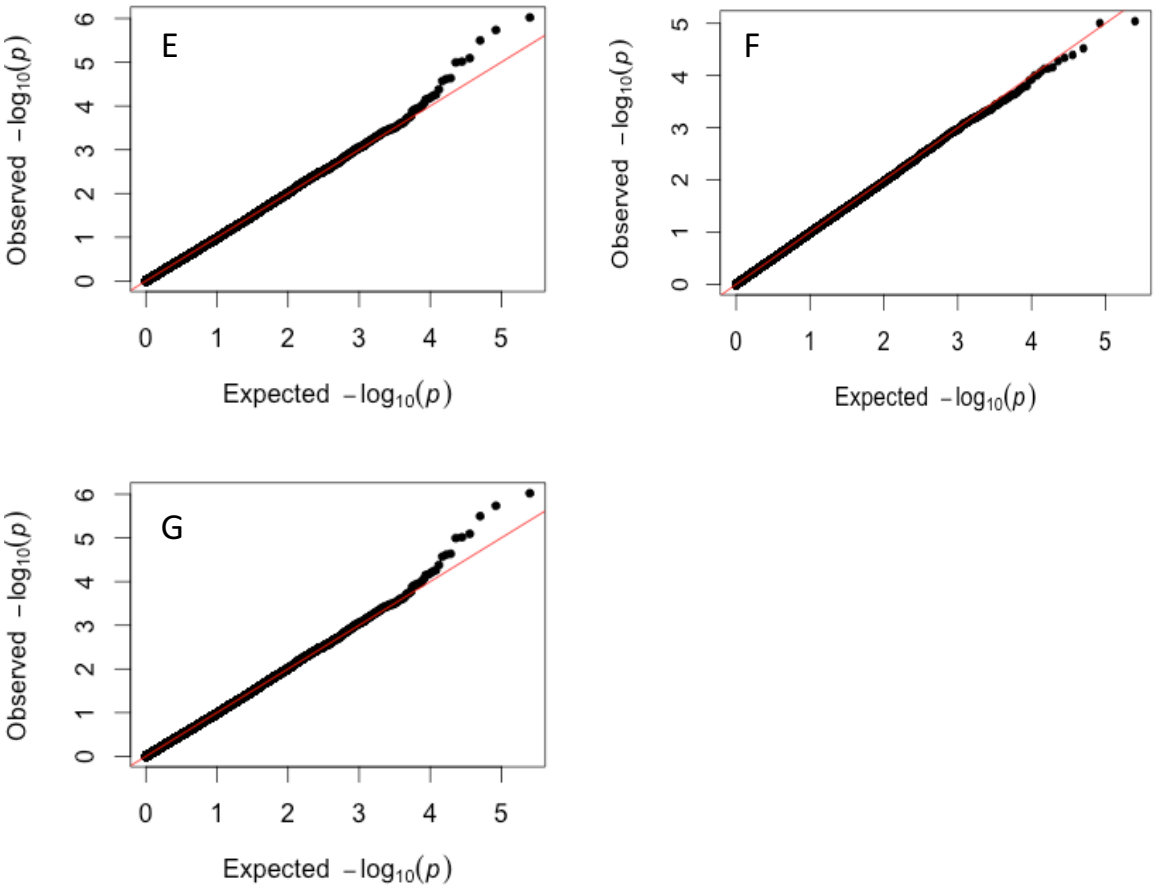
```
#!/bin/bash

#PBS -N MLMA
#PBS -q WitsLong
#PBS -l walltime=20:00:00,mem=3GB
#PBS -l nodes=1:ppn=1

WORKING_DIR='/home/venesa/heritability'

#Change to the working directory where the files are.
cd $WORKING_DIR

# name of binary plink filename (excluding .bed/.bim/.fam suffix)
plinkfile="combinedfemales"

# name of phenotype file in plink format (i.e. col1=fid, col2=iid, col3=phenotype
phenfile="sorted_combined_batches.phen"

# make up a name for the grm for related females
allfile="GRM_female"

# covariate file (including the suffix.qcovar if quantitative, if sex then .covar)
covarfile1="PCA_age_females.qcovar"
covarfile2="PCA_age_height.qcovar"
covarfile3="PCA_age_height_leanmass.qcovar"
covarfile4="PCA_age_height_fatmass.qcovar"

# make up a name for output file (it will have .mmla appended to the end by gcta)
#outfile="test.mma"

## Alternative way to have multiple phenotypes (mpheno-normally column 4 but if you give a
column number after that its usually the position of the column
#say BMI is in column 8 (-2) will be mpheno 6))
#column="7"

#gcta64 \
#        --mlma
#        --bfile ${plinkfile} \
#        --grm ${allfile} \
```

```
#         --pheno ${phenfile} \
#         --mpheno ${column} \
#         --qcovar ${covarfile} \
#         --out ${outfile}


function the_function ()
{
    gcta64 \
        --mlma \
     --bfile ${plinkfile} \
        --grm ${allfile} \
        --pheno ${phenfile} \
        --mpheno "$1" \
        --qcovar $2 \
        --out $1_$2
}

the_function 7 $covarfile1
the_function 8 $covarfile1
the_function 8 $covarfile2
the_function 9 $covarfile1
the_function 9 $covarfile2
the_function 10 $covarfile1
the_function 11 $covarfile1
the_function 11 $covarfile2
the_function 11 $covarfile3
the_function 12 $covarfile1
the_function 12 $covarfile2
the_function 12 $covarfile4
the_function 13 $covarfile1
the_function 13 $covarfile2
```

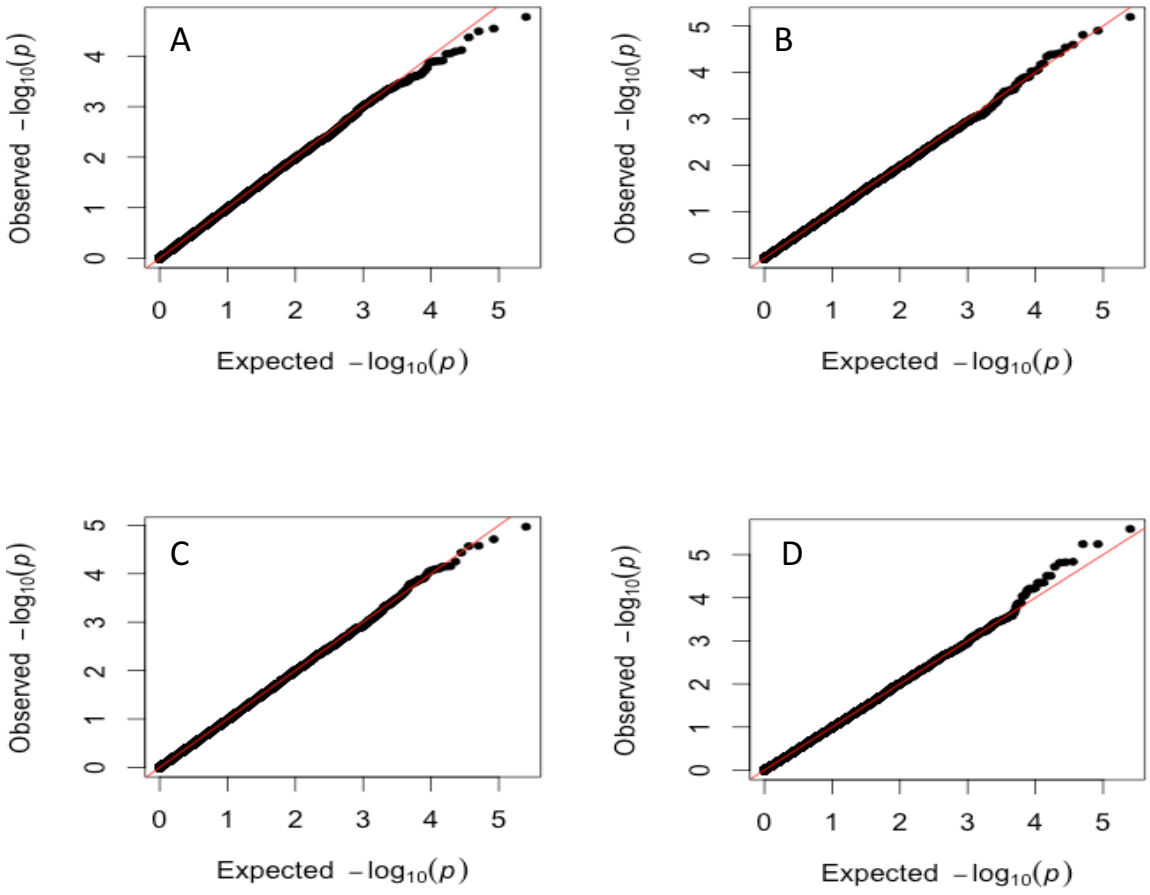
APPENDIX D: QQ Plots to assess the distribution of the test statistic for phenotypes in the study

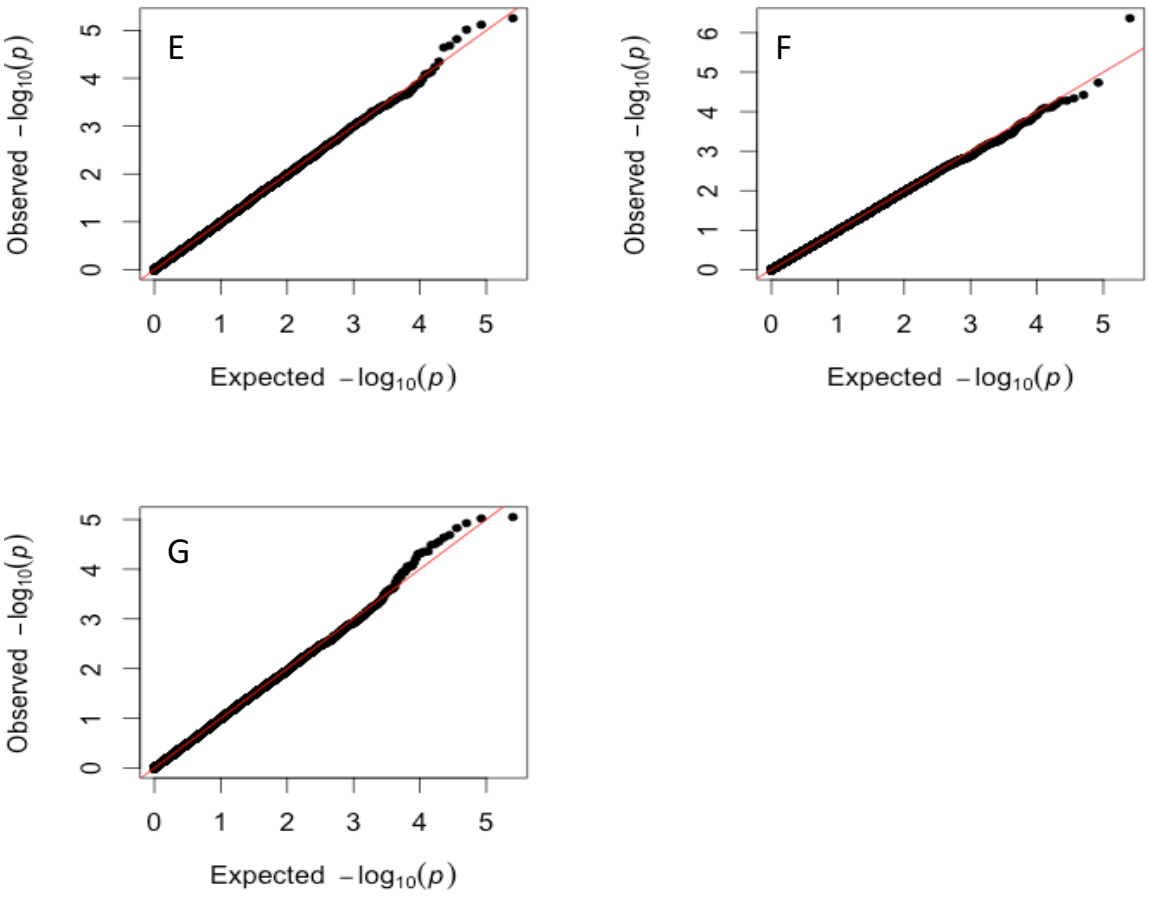
QQ plots were drawn in R vs.3.2.2 using the package qqman to visualise the distribution of the test-statistic for each phenotype analysis, and showed no evidence of population stratification.



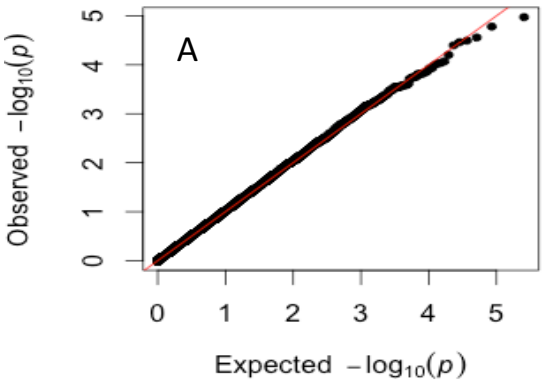


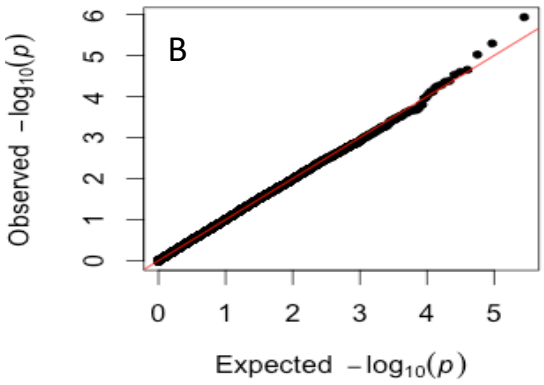
QQ Plot for all seven phenotypes when datasets are combined. A- logBMI; B-WC; C-HC, D-WHR, E-Fat mass, F-Lean mass, G-PFM

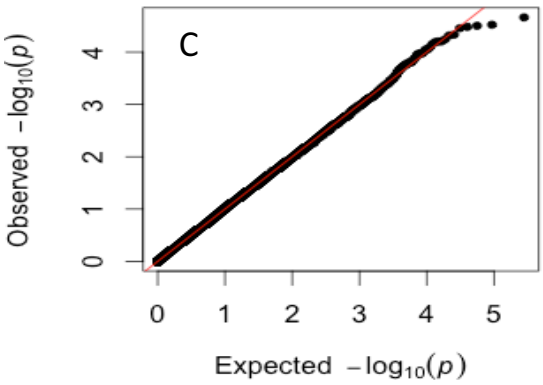

A
B
C
D
Observed $-\log_{10}(p)$
Expected $-\log_{10}(p)$
0
1
2
3
4
5

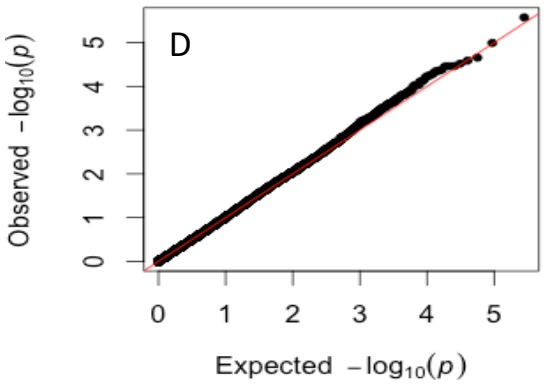


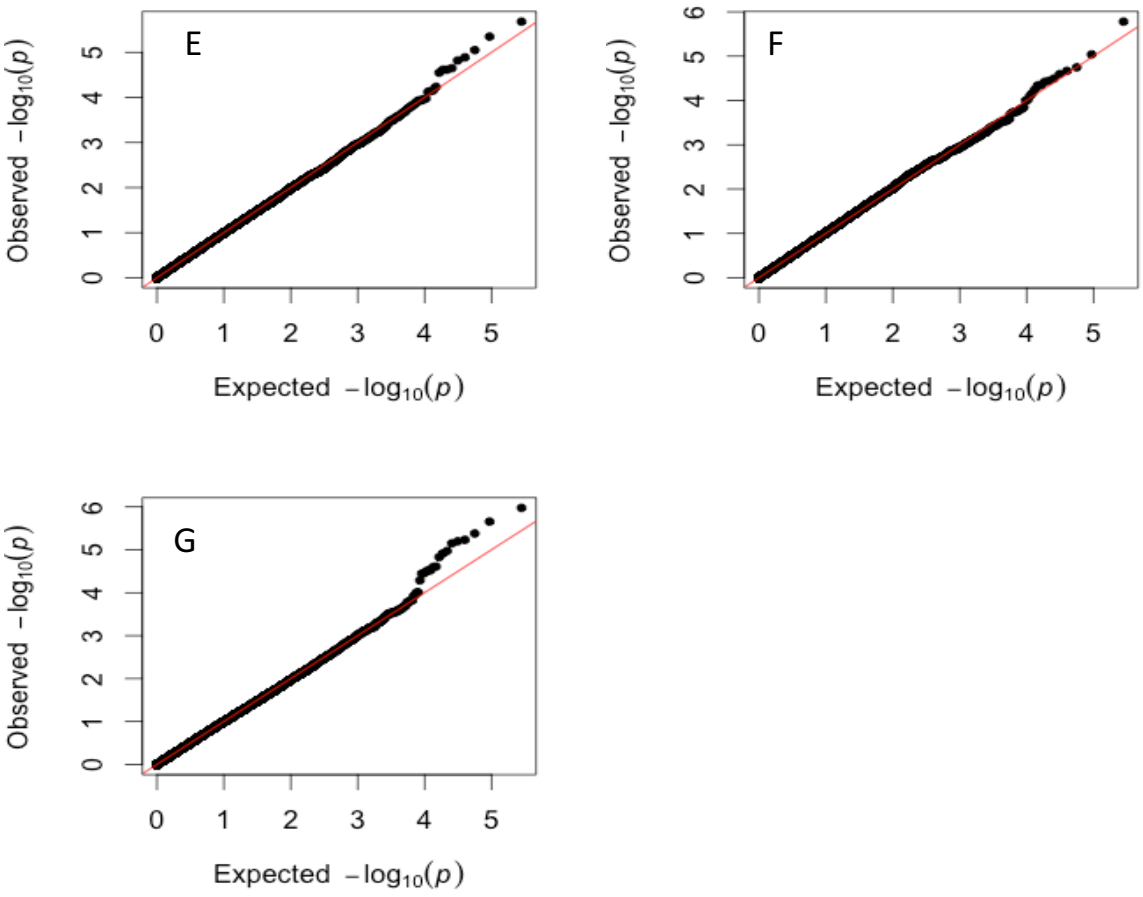
QQ Plot for all seven phenotypes when females are combined. A- logBMI; B-WC; C-HC, D-WHR, E-Fat mass, F-Lean mass, G-PFM


A
Observed $-\log_{10}(p)$
Expected $-\log_{10}(p)$
0 1 2 3 4 5
0 1 2 3 4 5

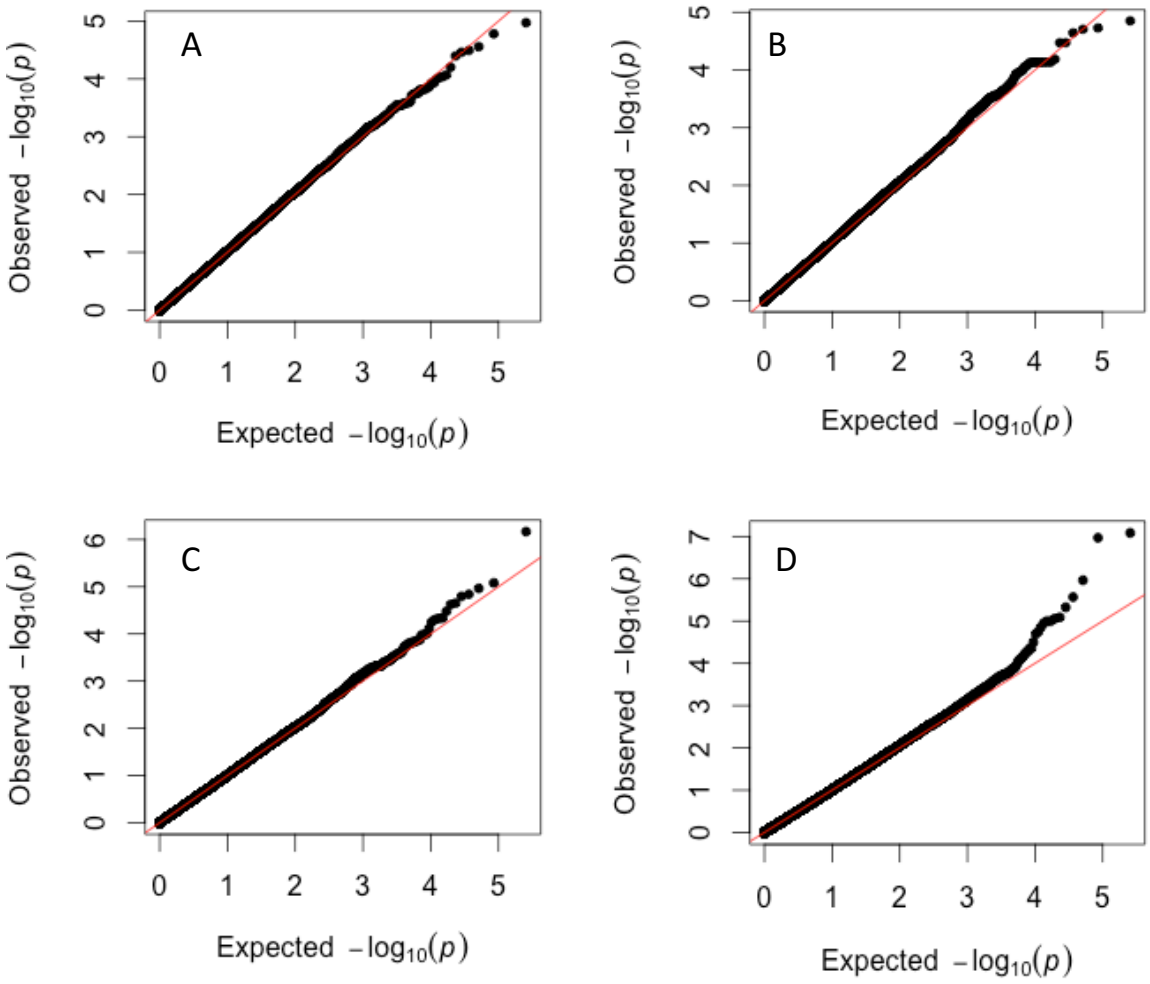

B
Observed $-\log_{10}(p)$
Expected $-\log_{10}(p)$
0 1 2 3 4 5 6
0 1 2 3 4 5

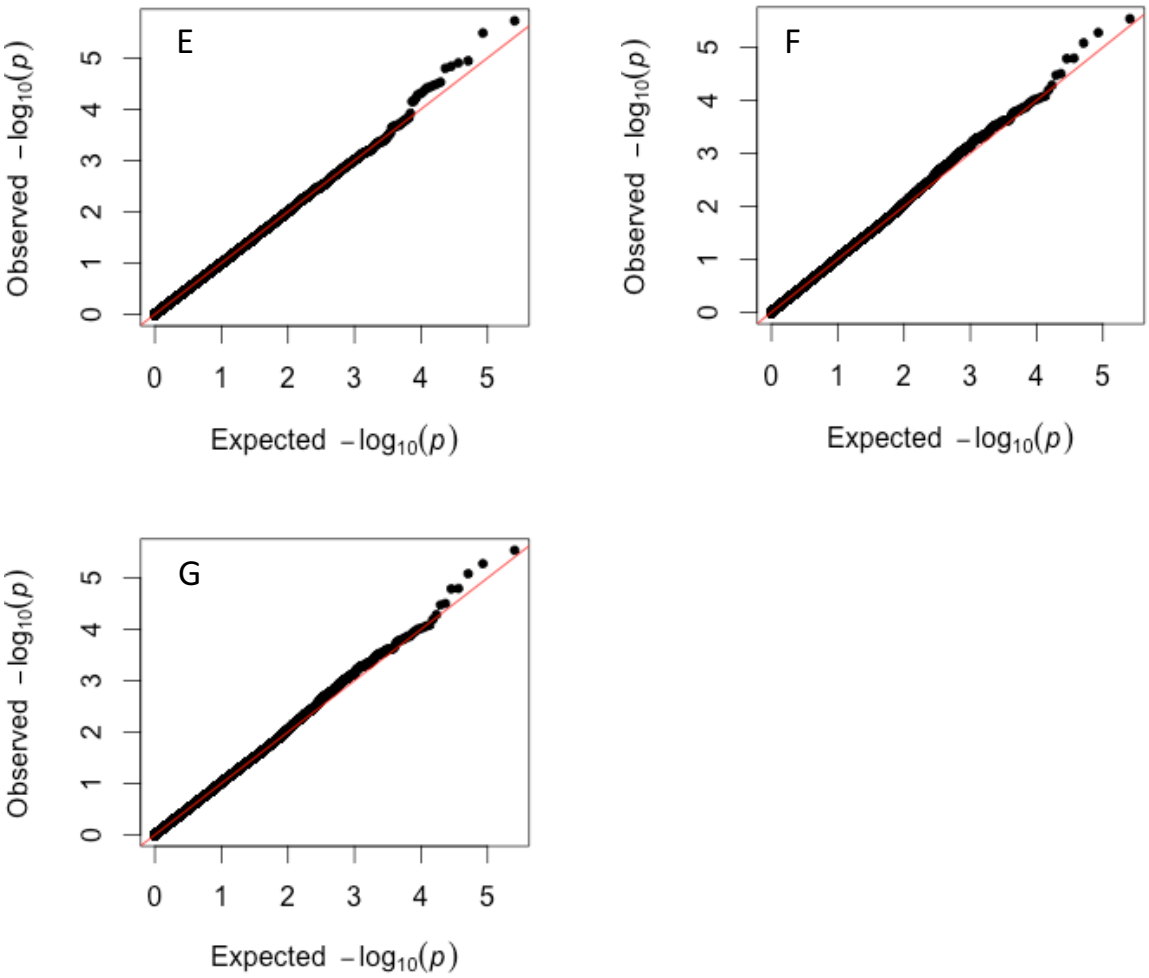

C
Observed $-\log_{10}(p)$
Expected $-\log_{10}(p)$
0 1 2 3 4
0 1 2 3 4 5


D
Observed $-\log_{10}(p)$
Expected $-\log_{10}(p)$
0 1 2 3 4 5
0 1 2 3 4 5

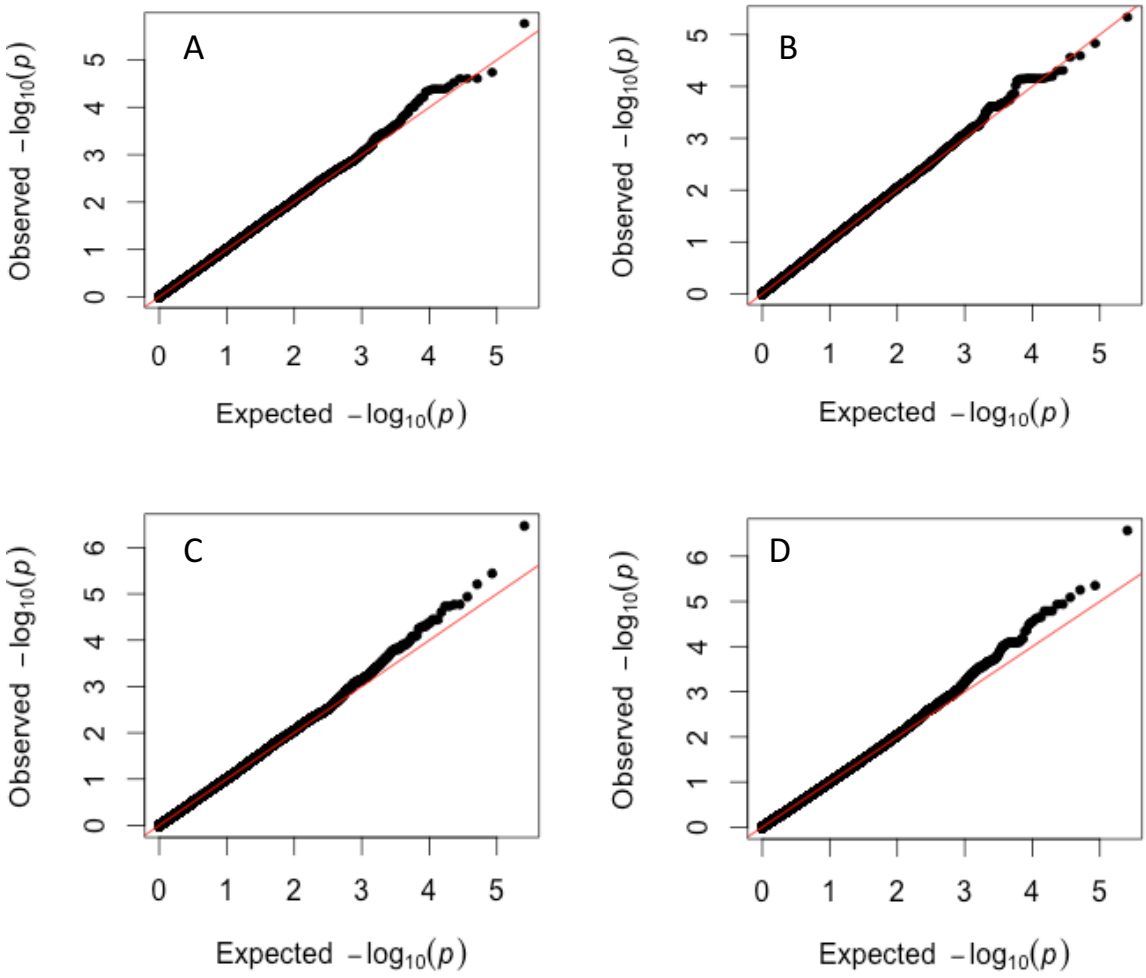


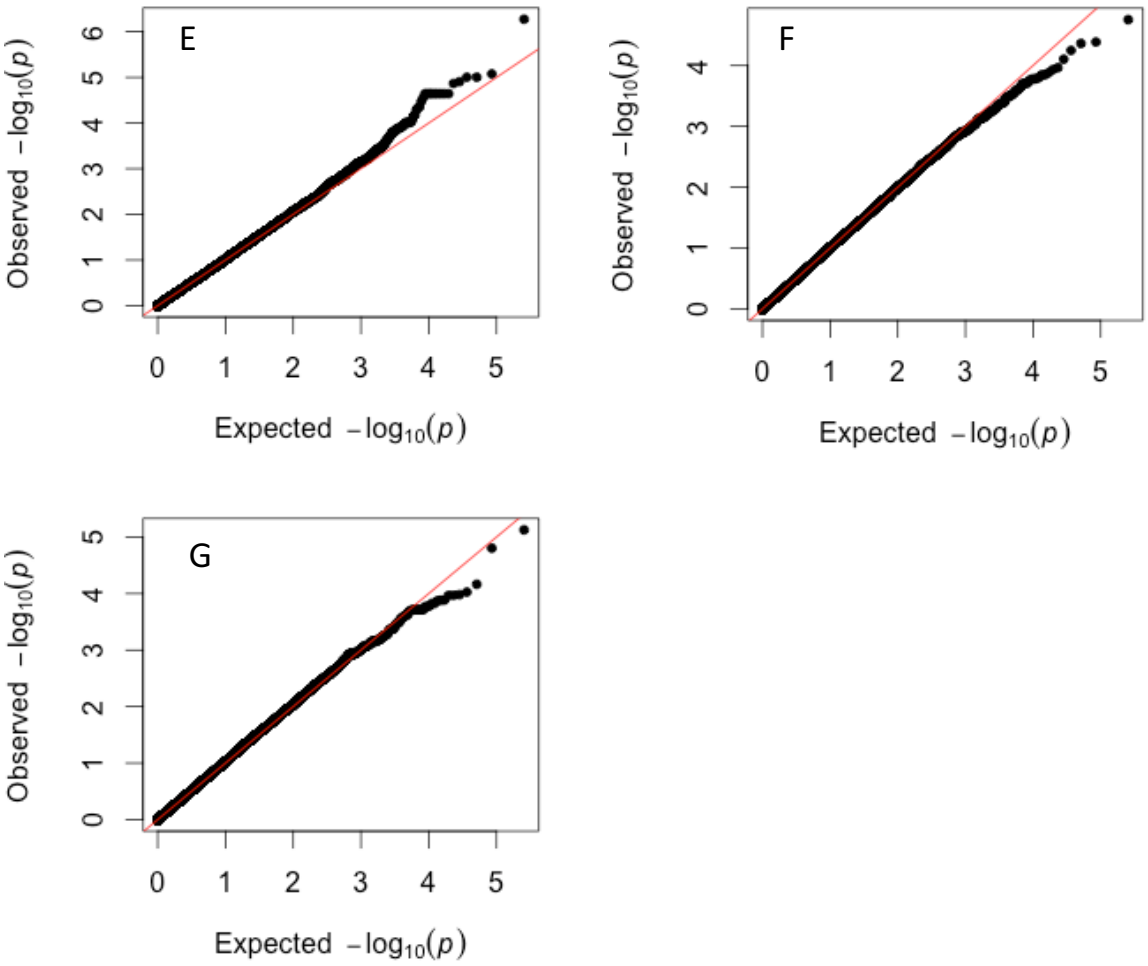
QQ Plot for all seven phenotypes with female caregivers. A- logBMI; B-WC; C-HC, D-WHR, E-Fat mass, F-Lean mass, G-PFM


A
B
C
D
Observed $-\log_{10}(p)$
Expected $-\log_{10}(p)$

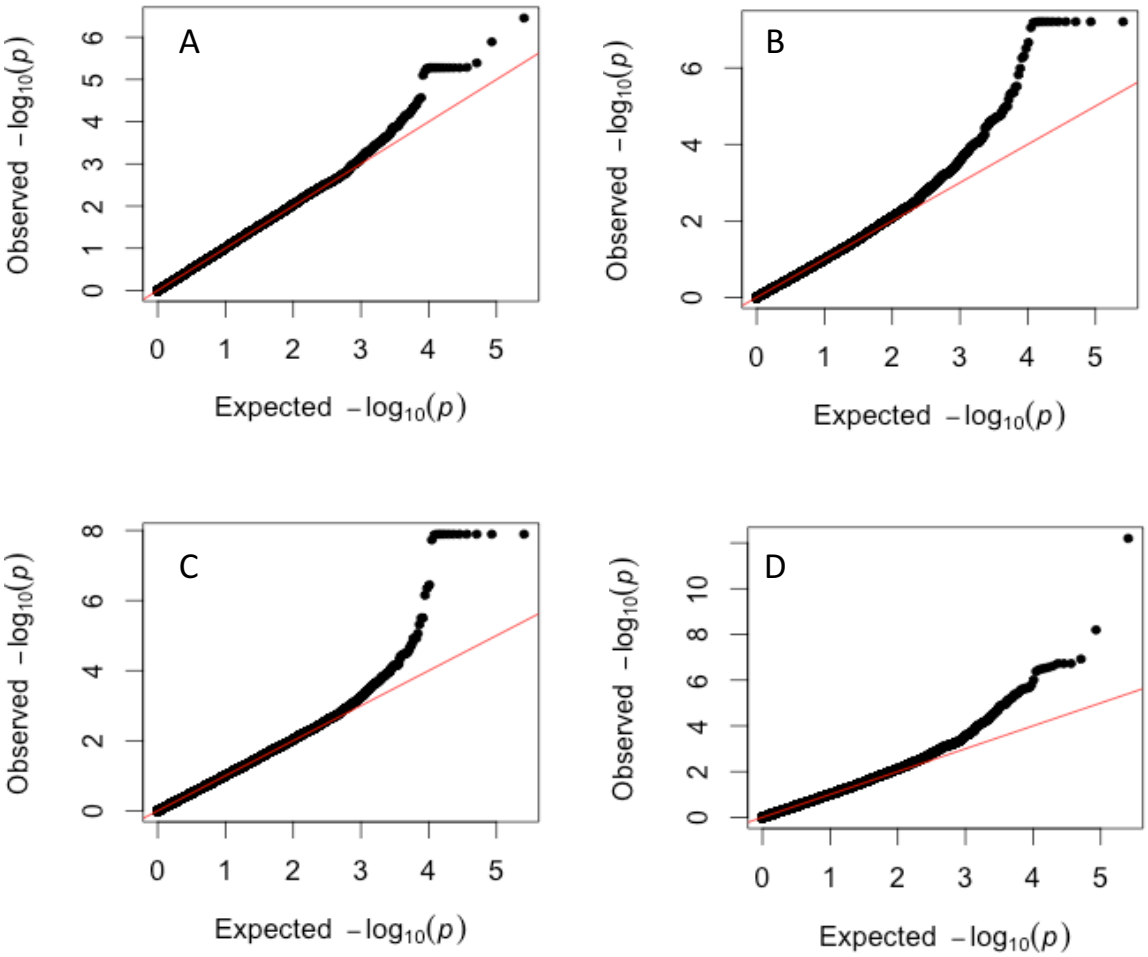


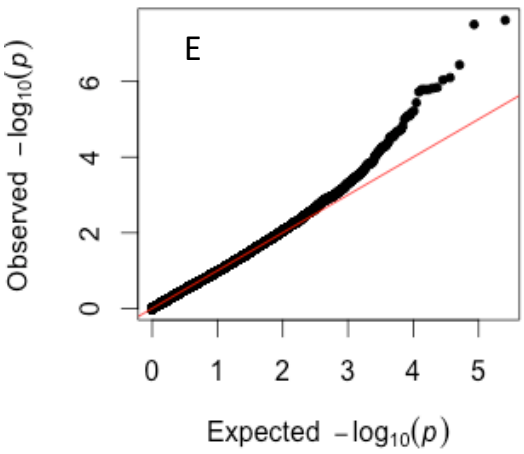
QQ Plot for all seven phenotypes with young adults. A- logBMI; B-WC; C-HC, D-WHR, E-Fat mass, F-Lean mass, G-PFM

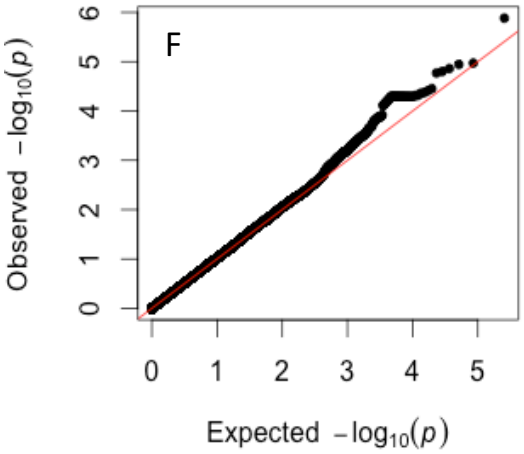

A
B
C
D
Observed $-\log_{10}(p)$
Expected $-\log_{10}(p)$

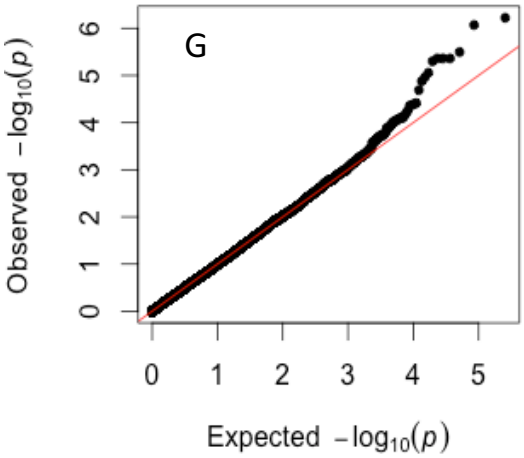


QQ Plot for all seven phenotypes with young female adults. A- logBMI; B-WC; C-HC, D-WHR, E-Fat mass, F-Lean mass, G-PFM


A
B
C
D
Observed $-\log_{10}(p)$
Expected $-\log_{10}(p)$







QQ Plot for all seven phenotypes with young male adults. A- logBMI; B-WC; C-HC, D-WHR, E-Fat mass, F-Lean mass, G-PFM

Appendix E: All results from association testing (basic, linear regression and mixed model association analysis)

Section A: Unadjusted association analysis

Table E1. The top associated SNPs with body composition traits in the female caregiver dataset

CHR-refers to chromosome; BP-base position, NMISS-number of samples used in the analysis, SE-standard error, Beta refers to per allele change in the phenotype

TRAIT	CHR	SNP	*GENE SYMBOL*	GENE LOCATION	BP[4]	NMISS	BETA	SE	P- value
BMI	8	rs76966674	*LPL | SLC18A1*	intergenic	19987694	972	0.03	0.01	3.08×10^{-5}
	2	rs6541885	*LOC728241 | CNTNAP5*	intergenic	123668890	972	-0.05	0.01	3.68×10^{-5}
	7	rs7808316	*DGKB | TMEM195*	intergenic	15112007	972	-0.02	0.00	3.82×10^{-5}
	1	rs9442211	*FMN2 | GREM2*	intergenic	238709807	972	-0.02	0.00	3.87×10^{-5}
	1	rs6425446	*LOC400796 | SEC16B*	intergenic	176031200	972	0.02	0.00	6.33×10^{-5}
	6	rs11967047	*CDKAL1*	intron	20949459	972	-0.02	0.01	7.33×10^{-5}
	6	rs7769223	*KHDRBS2 | LOC100128610*	intergenic	63215802	972	-0.02	0.00	7.46×10^{-5}
	7	rs75802768	*JAZF1 | LOC100128081*	intergenic	28088471	972	-0.02	0.01	9.58×10^{-5}
	2	rs4411698	*LOC728241 | CNTNAP5*	intergenic	123651577	972	-0.04	0.01	9.79×10^{-5}
WC	3	rs12493688	*HES1 | LOC100131551*	intergenic	195401542	960	136.60	31.36	1.46×10^{-5}
	6	rs9364558	*LPAL2*	intron	160849934	956	28.18	6.53	1.76×10^{-5}
	2	rs2114591	*SP110*	intron	230758813	959	25.93	6.04	1.96×10^{-5}
	4	rs7661253	*LOC728191 | ODZ3*	intergenic	183276086	958	49.57	12.00	3.93×10^{-5}
	17	rs11867979	*LOC728073 | RPL38*	intergenic	69148107	960	-33.78	8.19	4.08×10^{-5}
	16	rs1861554	*FTO*	intron	52607268	960	48.77	12.03	5.41×10^{-5}
	9	rs7021554	*ANRIL | LOC729983*	intergenic	22142884	960	45.32	11.21	5.67×10^{-5}
	13	rs7990860	*OLFM4 | LOC387930*	intergenic	52887195	960	-24.62	6.15	6.78×10^{-5}
	14	rs2282277	*LOC729165*	intron	44017413	960	-57.42	14.66	9.60×10^{-5}
WHR	15	rs12438098	*LOC728292 | MCTP2*	intergenic	92188493	960	0.03	0.01	2.29×10^{-6}

[4] Note that all BP positions are reported using NCBI Build 36 (hg18)

WHR	1	rs10926800	*PLD5* | *LOC400723*	intergenic	240867832	960	-0.02	0.00	3.11×10^{-6}
	11	rs481855	*PGR*	intron	100474570	960	-0.02	0.00	1.30×10^{-5}
	6	rs195852	*POU3F2* | *FBL4*	intergenic	99397313	960	0.02	0.00	4.96×10^{-5}
	2	rs11679288	*ALK*	intron	29494717	960	0.02	0.00	5.16×10^{-5}
	16	rs2037912	*PPL*	coding-1055/2;	4873940	960	0.06	0.01	5.50×10^{-5}
	12	rs115649230	*R3HDM2* | *INHBC*	intergenic	56001394	960	0.07	0.02	5.74×10^{-5}
	12	rs145249220	*R3HDM2* | *INHBC*	intergenic	56064088	960	0.07	0.02	5.74×10^{-5}
	6	rs9456538	*SLC22A3*	intron	160773383	960	0.02	0.00	6.06×10^{-5}
	18	rs10164068	*LOC100132894*	intron	72381591	959	-0.02	0.00	6.40×10^{-5}
	16	rs1861554	*FTO*	intron	52607268	960	0.03	0.01	6.65×10^{-5}
HC	1	rs6425446	*LOC400796* | *SEC16B*	intergenic	176031200	960	26.34	6.07	1.58×10^{-5}
	1	rs113015095	*ATG4C*	intron	63067732	951	39.73	9.55	3.49×10^{-5}
	12	rs3138139	*RDH5*	intron	54402150	960	-26.59	6.40	3.58×10^{-5}
	2	rs6541885	*LOC728241* | *CNTNAP5*	intergenic	123668890	960	-66.24	16.17	4.54×10^{-5}
	15	rs624191	*LOC100128434* | *GPR176*	intergenic	37880127	960	-33.28	8.13	4.59×10^{-5}
	2	rs4411698	*LOC728241* | *CNTNAP5*	intergenic	123651577	960	-60.63	14.82	4.65×10^{-5}
	8	rs78933755	*PPP1R3B* | *LOC100129150*	intergenic	9245811	960	104.90	25.93	5.63×10^{-5}
	11	rs12574325	*BDNFOS*	intron	27569624	960	51.73	12.89	6.42×10^{-5}
	11	rs16917135	*BDNFOS*	intron	27571281	960	51.73	12.89	6.42×10^{-5}
	8	rs35584813	*PPP1R3B* | *LOC100129150*	intergenic	9245040	960	102.20	25.48	6.53×10^{-5}
Fat mass	1	rs4311843	*DOCK7* | *ATG4C*	intergenic	176035387	927	2366	514.70	4.87×10^{-6}
	1	rs4072161	*LOC400796* | *SEC16B*	intergenic	176036310	927	2218	488.70	6.42×10^{-6}
	1	rs4348685	*LOC400796* | *SEC16B*	intergenic	176034251	927	-2371	545.10	$1,51 \times 10^{-5}$
	1	rs4075235	*LOC400796* | *SEC16B*	intergenic	176034945	927	-2371	545.10	1.51×10^{-5}
	2	rs4411698	*LOC728241* | *CNTNAP5*	intergenic	123651577	927	-5142	1202.00	2.09×10^{-5}
	1	rs2094122	*ANKRD38* | *USP1*	intergenic	62627179	923	2863	678.90	2.72×10^{-5}
	7	rs75802768	*JAZF1* | *LOC100128081*	intergenic	28088471	927	-2975	716.20	3.58×10^{-5}
	22	rs738671	*LOC100128818* | *TBC1D22A*	intergenic	45927613	927	2045	495.70	4.04×10^{-5}
	2	rs2114591	*SP110*	intron	230758813	926	1976	483.90	4.81×10^{-5}
	9	rs10781091	*GDA* | *ZFAND5*	intergenic	74064112	927	3451	860.30	6.52×10^{-5}
Lean mass	9	rs10781091	*GDA* | *ZFAND5*	intergenic	74064112	927	2352	480.40	1.15×10^{-6}
	14	rs34664345	*PNN* | *MIA2*	intergenic	38761479	927	1245	282.20	1.14×10^{-5}
	7	rs17443228	*LRRN3* | *IMMP2L*	intergenic	110774133	927	-2039	497.00	4.44×10^{-5}
	8	rs6601538	*KR6*	intron	10812864	927	-1336	329.30	5.39×10^{-5}
	5	rs17554781	*RGNEF* | *ENC1*	intergenic	73344284	927	1193	296.00	6.05×10^{-5}
	1	rs2375278	*RUN3* | *SYF2*	intergenic	25401625	927	5146	1294.00	7.51×10^{-5}

	3	rs9872002	*LOC344787 | GPD1L*	intergenic	32065107	927	1151	292.20	8.76×10^{-5}
	1	rs6664016	*FAM5B | LOC400796*	intergenic	175587093	927	1198	304.30	8.83×10^{-5}
	11	rs1263149	*BUD13*	intron	116145151	927	1044	266.80	9.71×10^{-5}
	11	rs1263151	*BUD13*	intron	116146238	927	1054	269.70	9.97×10^{-5}
PFM	1	rs6425446	*LOC400796 | SEC16B*	intergenic	176031200	927	1.51	0.32	2.99×10^{-6}
	1	rs4072161	*LOC400796 | SEC16B*	intergenic	176036310	927	1.47	0.32	6.66×10^{-6}
	1	rs4311843	*LOC400796 | SEC16B*	intergenic	176035387	927	1.48	0.34	1.55×10^{-5}
	1	rs2068973	*LOC400796 | SEC16B*	intergenic	176059338	927	1.40	0.32	1.61×10^{-5}
	11	rs116083352	*FADS2 | FADS3*	intergenic	61396917	927	3.22	0.74	1.65×10^{-5}
	1	rs6425453	*LOC400796 | SEC16B*	intergenic	176085789	927	1.36	0.32	2.39×10^{-5}
	1	rs4348685	*LOC400796 | SEC16B*	intergenic	176034251	927	-1.53	0.36	2.51×10^{-5}
	1	rs4075235	*LOC400796 | SEC16B*	intergenic	176034945	927	-1.53	0.36	2.51×10^{-5}
	1	rs10913437	*LOC400796 | SEC16B*	intergenic	176062763	927	1.35	0.32	3.32×10^{-5}

Table E2. Top associated SNPs with body composition traits in the young adult dataset

TRAIT	CHR	SNP	*GENE SYMBOL*	GENE LOCATION	BP[5]	NMI SS	BETA	SE	P- value
BMI	1	rs2753399	*ZFYVE9*	intron	52546945	953	0.09	0.02	9.98×10^{-7}
	3	rs157538	*EDEM1 | GRM7*	intergenic	6312837	951	0.02	0.01	1.56×10^{-5}
	10	rs7072408	*PCBD1 | UNC5B*	intergenic	72419228	949	-0.02	0.00	1.67×10^{-5}
	3	rs2370840	*ACVR2B*	utr	38504829	953	0.12	0.03	5.26×10^{-5}
	15	rs62017164	*TRPM7*	intron	48655616	953	0.07	0.02	5.84×10^{-5}
	15	rs62017165	*TRPM7*	intron	48655833	953	0.07	0.02	5.84×10^{-5}
	15	rs1060599	*TRPM7*	intron	48661924	953	0.07	0.02	5.84×10^{-5}
	15	rs62017202	*TRPM7*	intron	48736535	953	0.07	0.02	5.84×10^{-5}
	15	rs62017207	*TRPM7*	intron	48745151	953	0.07	0.02	5.84×10^{-5}
	15	rs62017208	*TRPM7*	intron	48746646	953	0.07	0.02	5.84×10^{-5}
	15	rs62017209	*TRPM7*	intron	48747259	953	0.07	0.02	5.84×10^{-5}

[5] Note that all BP positions are reported using NCBI Build 36 (hg18)

WC	1	rs2753399	*ZFYVE9*	intron	52546945	944	9.73	2.14	6.15×10^{-6}
	20	rs6059958	*HM13*	intron	29606939	942	-2.09	0.47	8.57×10^{-6}
	5	rs4607330	*RAB9P1 | EFNA5*	intergenic	105035114	939	3.64	0.82	1.08×10^{-5}
	12	rs2373459	*SPIC*	intron	100398087	944	2.24	0.52	1.64×10^{-5}
	15	rs4143844	*VPS13C*	intron	60100132	944	7.24	1.72	2.64×10^{-5}
	20	rs6120815	*HM13 | ID1*	intergenic	29648527	944	-1.83	0.44	2.81×10^{-5}
	15	rs34173668	*VPS13C*	intron	60092876	944	6.77	1.62	3.06×10^{-5}
	15	rs34958705	*VPS13C*	intron	60104049	944	6.77	1.62	3.06×10^{-5}
	15	rs34311301	*VPS13C*	intron	60129847	944	6.77	1.62	3.06×10^{-5}
	15	rs17271312	*VPS13C*	intron	60123571	939	6.75	1.62	3.29×10^{-5}
HC	1	rs2753399	*ZFYVE9*	intron	52546945	944	11.82	2.33	4.56×10^{-7}
	10	rs7072408	*PCBD1 | UNC5B*	intergenic	72419228	940	-2.40	0.54	1.01×10^{-5}
	4	rs7683492	*LOC644624 | ANKRD50*	intergenic	125737729	944	2.09	0.48	1.31×10^{-5}
	1	rs1329817	*POU3F1 | LOC400750*	intergenic	38709564	943	-1.93	0.46	2.72×10^{-5}
	12	rs17374759	*LOC644976 | LOC100131677*	intergenic	20228731	944	4.41	1.05	2.94×10^{-5}
	3	rs17186340	*STAC*	intron	36407617	944	4.24	1.04	4.63×10^{-5}
	1	rs11688	*JUN*	coding	59020581	941	-2.42	0.59	4.84×10^{-5}
	3	rs720679	*LRIG1 | KBTBD8*	intergenic	66916749	944	10.71	2.63	4.89×10^{-5}
	17	rs9893680	*KIAA1303*	intron	76407353	942	2.00	0.49	4.92×10^{-5}
	5	rs4607330	*RAB9P1 | EFNA5*	intergenic	105035114	939	3.67	0.90	5.17×10^{-5}
WHR	4	rs13113267	*MGC48628*	intron	91825412	943	0.02	0.00	6.19×10^{-8}
	1	rs12095241	*WARS2*	intron	119431520	941	0.02	0.00	9.89×10^{-7}
	3	rs2400349	*EPHB1 | LOC645218*	intergenic	136041127	943	0.05	0.01	1.78×10^{-6}
	1	rs12088290	*WARS2*	intron	119385449	943	0.02	0.00	4.34×10^{-6}
	3	rs17016150	*LOC644990 | LOC100130354*	intergenic	25313965	943	0.05	0.01	4.35×10^{-6}
	7	rs6972785	*JAZF1 | LOC100128081*	intergenic	28002776	942	0.02	0.01	4.57×10^{-6}
	9	rs4382560	*LOC646609 | LOC402360*	intergenic	22854667	940	0.02	0.00	6.28×10^{-6}
	9	rs28617580	*C9orf96*	intron	135242773	943	0.03	0.01	6.55×10^{-6}
	6	rs3923725	*LOC100129474 | SLC17A4*	intergenic	25842899	942	0.09	0.02	8.22×10^{-6}
	9	rs34399743	*C9orf96*	intron	135241179	943	0.03	0.01	9.13×10^{-6}
Fat mass	4	rs17668731	*AFF1*	intron	88218507	782	6601.00	1303.00	5.02×10^{-7}
	4	rs17604937	*AFF1*	intron	88204139	778	6040.00	1225.00	9.99×10^{-7}
	4	rs35658582	*AFF1 | KLHL8*	intergenic	88284946	787	6208.00	1277.00	1.41×10^{-6}
	1	rs2753399	*ZFYVE9*	intron	52546945	787	9871.00	2148.00	5.03×10^{-6}
	4	rs116385428	*AFF1*	intron	88257323	787	5223.00	1184.00	1.17×10^{-5}

Fat mass	1	chr1:56916879	*USP24* | *PPAP2B*	intergenic	56689467	785	3290.00	750.00	1.31×10^{-5}
	3	rs16831219	*PE5L* | *LOC131054*	intergenic	181389359	786	9087.00	2084.00	1.47×10^{-5}
	15	rs11071662	*FAM148B* | *FLJ38723*	intergenic	60287838	787	2466.00	567.60	1.59×10^{-5}
	21	rs2838727	*ITGB2*	intron	45140335	787	2426.00	569.50	2.30×10^{-5}
	16	rs1862748	*CDH1*	intron	67390444	785	2187.00	528.70	3.89×10^{-5}
	17	rs2955245	*CSH1* | *CSHL1*	intergenic	59333577	786	9848.00	2386.00	4.05×10^{-5}
Lean mass	6	rs115578077	*PHACTR1*	intron	13026552	785	-4091.00	895.80	5.75×10^{-6}
	6	rs115593358	*PHACTR1*	intron	13026798	787	-4056.00	906.90	8.88×10^{-6}
	19	rs8113016	*LOC284417*	intron	60525997	787	7851.00	1805.00	1.54×10^{-5}
	3	rs6441886	*CDCP1*	intron	45103101	786	-1613.00	371.50	1.59×10^{-5}
	13	rs116170577	*COL4A1*	intron	109621762	787	-2866.00	676.40	2.54×10^{-5}
	4	rs11734664	*STK32B*	intron	5276947	787	2849.00	686.50	3.70×10^{-5}
	16	rs111357538	*FTO*	intron	52400213	787	3500.00	859.90	5.18×10^{-5}
	19	rs12611418	*THEG* | *FAM148C*	intergenic	339413	787	8054.00	1988.00	5.59×10^{-5}
	16	rs7190220	*AKTIP* | *RPGRIP1L*	intergenic	52109683	787	-1595.00	394.50	5.81×10^{-5}
	16	rs117696251	*FTO*	intron	52358082	787	4666.00	1183.00	8.72×10^{-5}
PFM	4	rs17668731	*AFF1*	intron	88218507	782	9.21	1.83	6.04×10^{-7}
	4	rs35658582	*AFF1* | *KLHL8*	intergenic	88284946	787	8.62	1.80	1.87×10^{-6}
	4	rs17604937	*AFF1*	intron	88204139	778	8.26	1.73	2.08×10^{-6}
	3	rs4504165	*LOC730057*	intron	64676930	787	-5.18	1.11	3.46×10^{-6}
	3	rs6772129	*LOC730057*	intron	64675465	787	-5.01	1.09	4.87×10^{-6}
	3	rs9860730	*LOC730057*	intron	64676186	787	-5.02	1.10	5.71×10^{-6}
	3	rs4422297	*LOC730057*	intron	64679900	787	-4.94	1.09	7.08×10^{-6}
	4	rs116385428	*AFF1*	intron	88257323	787	7.34	1.66	1.17×10^{-5}
	1	chr1:56916879	*USP24* | *PPAP2B*	intergenic	56689467	785	4.56	1.05	1.74×10^{-5}
	21	rs2838727	*ITGB2*	intron	45140335	787	3.43	0.80	2.04×10^{-5}
	2	rs529963	*HJURP*	coding/nonsyn stop|R	234423239	787	-2.64	0.62	2.31×10^{-5}

Table E3. - Top associated SNPs with body composition traits when the dataset is stratified into young male adults

TRAIT	CHR	SNP	*GENE SYMBOL*	GENE LOCATION	BP[6]	NMISS	BETA	SE	P- value
BMI	19	rs11670765	*LOC339344*	intron	51093838	505	0.02	0.00	2.64×10^{-7}
	19	rs8113016	*LOC284417*	intron	60525997	505	0.07	0.01	1.54×10^{-6}
	12	rs7961916	*TB3 | LOC100129020*	intergenic	113839509	505	0.02	0.00	3.21×10^{-6}
	15	rs17598819	*TRPM7*	intron	48701919	503	0.12	0.03	5.03×10^{-6}
	15	rs17598264	*USP50 | TRPM7*	intergenic	48632003	505	0.12	0.03	5.13×10^{-6}
	15	rs62021060	*TRPM7*	utr	48640931	505	0.12	0.03	5.13×10^{-6}
	15	rs62017164	*TRPM7*	intron	48655616	505	0.12	0.03	5.13×10^{-6}
	15	rs62017165	*TRPM7*	intron	48655833	505	0.12	0.03	5.13×10^{-6}
	15	rs1060599	*TRPM7*	intron	48661924	505	0.12	0.03	5.13×10^{-6}
	15	rs62017202	*TRPM7*	intron	48736535	505	0.12	0.03	5.13×10^{-6}
WC	8	rs2013640	*DLC1*	intron	13133461	502	3.13	0.58	9.64×10^{-8}
	15	rs17598264	*TRPM7*	intron	48632003	502	15.76	3.08	4.36×10^{-7}
	15	rs62021060	*TRPM7*	intron	48640931	502	15.76	3.08	4.36×10^{-7}
	15	rs62017164	*TRPM7*	intron	48655616	502	15.76	3.08	4.36×10^{-7}
	15	rs62017165	*TRPM7*	intron	48655833	502	15.76	3.08	4.36×10^{-7}
	15	rs1060599	*TRPM7*	intron	48661924	502	15.76	3.08	4.36×10^{-7}
	15	rs62017202	*TRPM7*	intron	48736535	502	15.76	3.08	4.36×10^{-7}
	15	rs62017207	*TRPM7*	intron	48745151	502	15.76	3.08	4.36×10^{-7}
	15	rs62017208	*TRPM7*	intron	48746646	502	15.76	3.08	4.36×10^{-7}
	15	rs62017209	*TRPM7*	intron	48747259	502	15.76	3.08	4.36×10^{-7}
HC	15	rs17598264	*TRPM7*	intron	48632003	502	17.35	3.20	9.13×10^{-8}
	15	rs62021060	*TRPM7*	intron	48640931	502	17.35	3.20	9.13×10^{-8}
	15	rs62017164	*TRPM7*	intron	48655616	502	17.35	3.20	9.13×10^{-8}
	15	rs62017165	*TRPM7*	intron	48655833	502	17.35	3.20	9.13×10^{-8}
	15	rs1060599	*TRPM7*	intron	48661924	502	17.35	3.20	9.13×10^{-8}
	15	rs62017202	*TRPM7*	intron	48736535	502	17.35	3.20	9.13×10^{-8}

[6] Note that all BP positions are reported using NCBI Build 36 (hg18)

HC	15	rs62017207	*TRPM7*	intron	48745151	502	17.35	3.20	9.13×10^{-8}
	15	rs62017208	*TRPM7*	intron	48746646	502	17.35	3.20	9.13×10^{-8}
	15	rs62017209	*TRPM7*	intron	48747259	502	17.35	3.20	9.13×10^{-8}
	15	rs17598819	*TRPM7*	intron	48701919	500	17.35	3.20	9.27×10^{-8}
WHR	2	rs115743734	*COBLL1*	intron	165255601	502	0.09	0.02	5.26×10^{-9}
	7	rs114209770	*DGKB*	intron	14547297	502	0.08	0.02	6.04×10^{-8}
	7	rs76358458	*DGKB*	intron	14548446	502	0.08	0.02	6.04×10^{-8}
	12	rs76712056	*FLJ40142*	intron	108989096	502	0.08	0.01	1.15×10^{-7}
	3	rs1543143	*SRGAP3*	intron	9087876	502	0.06	0.01	1.24×10^{-7}
	3	rs17651358	*CNTN4*	intron	3039513	502	0.08	0.02	2.52×10^{-7}
	9	rs7020313	*UBQLN1 | GKAP1*	intergenic	85537074	502	0.08	0.02	3.04×10^{-7}
	12	rs12308957	*FLJ40142*	intron	22242796	502	0.06	0.01	7.13×10^{-7}
Fat mass	1	rs9970334	*ICMT | C1orf211*	intergenic	6218825	414	22930.00	4000.00	1.92×10^{-8}
	1	rs16852018	*LOC400796 | SEC16B*	intergenic	176098880	414	6397.00	1127.00	2.57×10^{-8}
	11	rs77275360	*SLC22A18*	intron	2894032	411	4767.00	924.20	3.91×10^{-7}
	1	rs6670797	*GP7 | MGC52498*	intergenic	52870169	414	9192.00	1812.00	5.94×10^{-7}
	1	rs2753399	*ZFYVE9*	intron	52546945	414	9002.00	1814.00	1.03×10^{-6}
	12	rs2116677	*OCC-1*	intron	104232358	414	6904.00	1442.00	2.37×10^{-6}
	3	rs115473751	*EIF5A2*	intron	172106758	414	4897.00	1026.00	2.51×10^{-6}
	1	rs72941254	*NEGR1 | LOC100132353*	intergenic	72446505	414	3203.00	676.70	3.04×10^{-6}
	1	rs72941257	*NEGR1 | LOC100132353*	intergenic	72451438	414	3203.00	676.70	3.04×10^{-6}
	1	rs72941270	*NEGR1 | LOC100132353*	intergenic	72465674	414	3203.00	676.70	3.04×10^{-6}
Lean mass	19	rs8113016	*LOC284417*	intron	60525997	414	8690.00	1651.00	2.27×10^{-7}
	1	rs587271	*SSBP3*	intron	54515699	414	9424.00	2076.00	7.39×10^{-6}
	1	rs59391530	*TAF1A | MIA3*	intergenic	220847044	414	3415.00	773.40	1.29×10^{-5}
	1	rs7525548	*TNNI3K*	intron	74774062	414	-1858.00	423.10	1.43×10^{-5}
	8	rs2409658	*PIN1*	intron	10706375	414	-2123.00	497.10	2.42×10^{-5}
	6	rs6924854	*PHACTR1*	intron	13229646	414	-2167.00	510.60	2.70×10^{-5}
	8	rs9969626	*PIN1*	intron	10708211	413	-2123.00	503.70	3.09×10^{-5}
	3	rs6441886	*CDCP1*	intron	45103101	414	-1595.00	383.70	3.95×10^{-5}
	17	rs4459609	*CYB561 | LOC342541*	intergenic	58902680	414	-1637.00	394.60	4.09×10^{-5}
	4	rs1844994	*LOC100132227*	intron	108417809	414	-2716.00	655.90	4.19×10^{-5}
PFM	1	rs16852018	*LOC400796 | SEC16B*	intergenic	176098880	414	7.066	1.40	6.73×10^{-7}
	11	rs77275360	*SLC22A18*	intron	2894032	411	5.716	1.14	8.00×10^{-7}
	1	rs6670797	*GP7 | MGC52498*	intergenic	52870169	414	10.75	2.24	2.26×10^{-6}

PFM	1	rs72941254	*NEGR1 | LOC100132353*	intergenic	72446505	414	3.794	0.84	7.48×10^{-6}
	1	rs72941257	*NEGR1 | LOC100132353*	intergenic	72451438	414	3.794	0.84	7.48×10^{-6}
	1	rs72941270	*NEGR1 | LOC100132353*	intergenic	72465674	414	3.794	0.84	7.48×10^{-6}
	3	rs58574369	*NT5DC2*	intron	52536300	414	4.583	1.01	7.77×10^{-6}
	1	rs72941224	*NEGR1 | LOC100132353*	intergenic	72411598	411	3.784	0.84	8.49×10^{-6}
	1	rs2753399	*ZFYVE9*	intron	52546945	414	10.12	2.25	8.90×10^{-6}
	14	rs10134920	*LOC100132612 | C14orf177*	intergenic	97789730	414	3.532	0.81	1.84×10^{-5}

Table E4. Top associated SNPs with body composition traits when the dataset is stratified into young female adults

TRAIT	CHR	SNP	GENE SYMBOL	GENE LOCATION	BP	NMISS	BETA	SE	P
BMI	3	rs149090	*EDEM1 | GRM7*	intergenic	6486781	448	0.12	0.02	2.02×10^{-6}
	5	rs32479	*ST8SIA4*	intron	100200494	448	0.13	0.03	1.79×10^{-5}
	3	rs9836894	*PBRM1*	intron	52691165	448	0.07	0.02	2.40×10^{-5}
	3	rs76913378	*GLT8D1*	intron	52706513	448	0.07	0.02	2.40×10^{-5}
	3	rs13325325	*NEK4*	intron	52778629	448	0.07	0.02	2.40×10^{-5}
	16	rs113530185	*DHODH | HP*	intergenic	70638303	448	0.06	0.01	2.92×10^{-5}
	3	rs9824342	*NISCH*	intron	52486947	448	0.06	0.02	3.49×10^{-5}
	3	rs9813653	*DNAH1*	intron	52366874	448	0.05	0.01	4.07×10^{-5}
	3	rs9828432	*DNAH1*	intron	52375005	448	0.05	0.01	4.07×10^{-5}
	3	rs9843055	*PHF7 | SEMA3G*	intergenic	52434363	448	0.05	0.01	4.07×10^{-5}
WC	5	rs32479	*ST8SIA4*	intron	100200494	442	17.60	4.11	2.31×10^{-5}
	11	rs114431455	*SLC35C1 | CRY2*	intergenic	45795746	442	8.01	1.91	3.34×10^{-5}
	5	rs4607330	*RAB9P1 | EFNA5*	intergenic	105035114	440	5.26	1.26	3.54×10^{-5}
	17	rs2006827	*GAS7*	intron	9798897	440	-2.94	0.70	3.63×10^{-5}
	15	rs34173668	*VPS13C*	intron	60092876	442	9.54	2.32	4.51×10^{-5}
	15	rs34958705	*VPS13C*	intron	60104049	442	9.54	2.32	4.51×10^{-5}
	15	rs34311301	*VPS13C*	intron	60129847	442	9.54	2.32	4.51×10^{-5}
	15	rs17271312	*VPS13C*	intron	60123571	439	9.51	2.32	5.03×10^{-5}
	4	rs7434427	*LOC285501 | LOC728081*	intergenic	179754269	442	3.11	0.76	5.41×10^{-5}
	11	rs80081080	*SLC35C1 | CRY2*	intergenic	45796003	442	6.83	1.69	5.95×10^{-5}

HC	5	rs32479	*ST8SIA4*	intron	100200494	443	20.28	3.97	4.72×10^{-7}
	3	rs149090	*EDEM1* | *GRM7*	intergenic	6486781	443	14.88	3.19	4.22×10^{-6}
	10	rs11598794	*SGMS1*	intron	51766998	442	10.34	2.35	1.34×10^{-5}
	11	rs114431455	*SLC35C1* | *CRY2*	intergenic	45795746	443	8.13	1.86	1.45×10^{-5}
	16	rs3091402	*IGSF6*	intron	21571296	443	4.07	0.94	2.00×10^{-5}
	3	rs9836894	*PBRM1*	intron	52691165	443	8.77	2.05	2.32×10^{-5}
	3	rs76913378	*GLT8D1*	intron	52706513	443	8.77	2.05	2.32×10^{-5}
	3	rs13325325	*NEK4*	intron	52778629	443	8.77	2.05	2.32×10^{-5}
	12	rs4766988	*PTPN11* | *RPH3A*	intergenic	111589413	443	9.52	2.25	2.81×10^{-5}
WHR	9	rs1571578	*ZNF618*	intron	115754911	442	0.11	0.02	7.73×10^{-7}
	12	rs12578133	*LOC100128389* | *LOC100131905*	intergenic	41307952	442	0.07	0.02	1.19×10^{-5}
	2	rs4312532	*SMEK2*	intron	55665565	442	-0.02	0.01	1.34×10^{-5}
	11	rs116828695	*KCNQ1*	intron	2428628	441	0.11	0.03	1.35×10^{-5}
	3	rs118152687	*CADM2*	intron	85886949	442	0.06	0.01	1.54×10^{-5}
	11	rs78366803	*LOC100128354* | *MTNR1B*	intergenic	92331703	442	0.08	0.02	1.69×10^{-5}
	11	rs11607061	*LOC100128354* | *MTNR1B*	intergenic	86737757	442	0.12	0.03	1.90×10^{-5}
	2	rs73922219	*BRE*	intron	28165640	439	0.05	0.01	2.14×10^{-5}
	6	rs1247330	*PLG* | *MAP3K4*	intergenic	161246640	442	-0.02	0.00	2.20×10^{-5}
	6	rs2073724	*PLG* | *MAP3K4*	intergenic	31237686	442	0.05	0.01	3.12×10^{-5}
Fat mass	3	rs117195838	*BTD*	intron	15625332	373	9906.00	1972.00	7.88×10^{-7}
	19	rs12611418	*LOC389435* | *OPRM1*	intergenic	339413	373	25370.00	5498.00	5.45×10^{-6}
	2	rs3934784	*LOC100131211* | *NAB1*	intergenic	191156494	373	35470.00	7769.00	6.78×10^{-6}
	2	rs13005335	*LOC100131211* | *NAB1*	intergenic	191179265	373	35470.00	7769.00	6.78×10^{-6}
	11	rs11224449	*FLJ32810*	intron	100166537	373	10290.00	2278.00	8.39×10^{-6}
	10	rs17106320	*GRID1*	coding	87888709	373	-3065.00	706.10	1.84×10^{-5}
	2	rs114285212	*LOC727944* | *TMEM18*	intergenic	630159	373	12060.00	2780.00	1.86×10^{-5}
	2	rs78501377	*LOC727944* | *TMEM18*	intergenic	630682	373	12060.00	2780.00	1.86×10^{-5}
	2	rs76275602	*LOC727944* | *TMEM18*	intergenic	631877	373	12060.00	2780.00	1.86×10^{-5}
	2	rs78460669	*LOC727944* | *TMEM18*	intergenic	633005	373	12060.00	2780.00	1.86×10^{-5}
	2	rs78559588	*LOC727944* | *TMEM18*	intergenic	637656	373	12060.00	2780.00	1.86×10^{-5}
Lean mass	11	rs90192	*SIDT2*	intron	116564557	373	1481.00	351.20	3.13×10^{-5}
	11	rs6590779	*LOC729305*	intergenic-	134317533	372	-1613.00	385.00	3.49×10^{-5}
	2	rs1477514	*NCK2* | *LOC100132455*	intergenic	105896372	373	2810.00	674.40	3.85×10^{-5}

Lean mass	4	rs11734664	*STK32B*	intron	5276947	373	2773.00	690.70	7.23×10^{-5}
	12	rs74103533	*LOC729298 | HMGA2*	intergenic	64496995	372	2157.00	539.10	7.65×10^{-5}
	11	rs236916	*PCSK7*	intron	116594838	373	1537.00	385.20	7.97×10^{-5}
	8	rs6472822	*LY96 | JPH1*	intergenic	75213296	373	1909.00	482.90	9.23×10^{-5}
	4	rs7434427	*LOC285501 | LOC728081*	intergenic	179754269	373	1479.00	374.70	9.52×10^{-5}
	10	rs11598794	*SGMS1*	intron	51766998	372	4592.00	1165.00	9.71×10^{-5}
PFM	3	rs117195838	*BTD*	intron	15625332	373	7.70	1.76	1.62×10^{-5}
	15	rs78946279	*VPS33B*	intron	89348470	373	2.20	0.51	2.28×10^{-5}
	19	rs4802246	*SFRS16*	intron	50255028	373	-2.93	0.69	2.50×10^{-5}
	2	rs3843862	*C2orf3 | LRRTM4*	intergenic	76096624	373	-3.41	0.84	6.03×10^{-5}
	10	rs17106320	*GRID1*	coding	87888709	373	-2.54	0.63	6.54×10^{-5}
	16	rs6501109	*A2BP1 | LOC283953*	intergenic	8398889	371	1.97	0.50	9.29×10^{-5}

Section B: Results from linear regression analysis (with adjustment for covariates)

Table E5. Top associated SNPs with body composition traits in female caregivers with adjustment for covariates under the various models outlined in Table 3.1 (Chapter 3), also including Model 3 (where FM is adjusted for LM) and Model 4 where LM is adjusted for FM.

CHR-refers to chromosome, BP-base position, NMISS-number of samples used in the analysis, SE-standard error, Beta refers to per allele change in the phenotype, A1 refers to the minor allele (default allele in PLINK). L95 and U 95 refer to to the lower and upper confidence intervals and P-refers to the *P*-value

TRAIT	CHR	SNP	GENE SYMBOL	GENE LOCATION	BP	A1	NMISS	BETA	SE	STAT	P
BMI	8	rs76966674	*LPL | SLC18A1*	intergenic	19987694	C	970	0.03	0.01	4.33	1.63×10^{-5}
Model 1	1	rs9442211	*FMN2 | GREM2*	intergenic	238709807	A	970	-0.02	0.00	-4.15	3.55×10^{-5}
	11	chr11:100005545	*CNTN5 | LOC440063*	intergenic	100005545	A	970	-0.04	0.01	-4.08	4.81×10^{-5}
	3	rs2271077	*GALNTL2. syn*	coding	16236424	T	970	0.02	0.01	4.07	5.18×10^{-5}
	2	rs6541885	*LOC728241 | CNTNAP5*	intergenic	123668890	A	970	-0.04	0.01	-4.06	5.39×10^{-5}
	7	rs7808316	*DGKB | TMEM195*	intergenic	15112007	A	970	-0.02	0.00	-4.04	5.68×10^{-5}
	7	rs2037695	*UBE2H | ZC3HC1*	intergenic	129416799	T	970	-0.02	0.00	-4.04	5.80×10^{-5}
	2	rs4411698	*LOC728241 | CNTNAP5*	intergenic	123651577	C	970	-0.04	0.01	-4.04	5.89×10^{-5}

BMI	7	rs75802768	*JAZF1* | *LOC100128081*	intergenic	28088471	C	970	-0.02	0.01	-4.02	6.38×10^{-5}
	14	rs34664345	*PNN* | *MIA2*	intergenic	38761479	G	970	0.02	0.00	4.01	6.51×10^{-5}
WC	2	rs2114591	*SP110*	intron	30758813	T	957	27.01	5.83	4.63	4.10×10^{-6}
Model 1	3	rs12493688	*HES1* | *LOC100131551*	intergenic	195401542	G	958	137.90	30.22	4.56	5.72×10^{-6}
	6	rs9364558	*LPAL2*	intron	160849934	G	954	28.39	6.30	4.51	7.31×10^{-6}
	9	rs10781091	*GDA* | *ZFAND5*	intergenic	74064112	A	958	45.05	10.30	4.37	1.36×10^{-5}
	7	rs6583393	*CDC14C* | *VWC2*	intergenic	49734785	A	958	94.31	22.57	4.18	3.20×10^{-5}
	12	rs75772217	*CU2*	intron	1102 17165	A	958	-117.80	28.79	-4.09	4.67×10^{-5}
	14	rs11622292	*PRO1768* | *FON3*	intergenic	89136080	T	958	33.68	8.27	4.07	5.00×10^{-5}
	4	rs7661253	*LOC728191* | *ODZ3*	intergenic	183276086	C	956	46.35	11.60	4.00	6.90×10^{-5}
	9	rs7021554	*ANRIL* | *LOC729983*	intergenic	22142884	T	958	43.16	10.82	3.99	7.13×10^{-5}
	13	rs7990860	*OLFM4* | *LOC387930*	intergenic	52887195	A	958	-23.60	5.94	-3.97	7.74×10^{-5}
HC	1	rs6425446	*LOC400796* | *SEC16B*	intergenic	176031200	A	958	25.58	6.04	4.23	2.52×10^{-5}
Model 1	2	rs4411698	*LOC728241* | *CNTNAP5*	intergenic	123651577	C	958	-61.21	14.71	-4.16	3.47×10^{-5}
	12	rs3138139	*RDH5*	intron	54402150	G	958	-26.40	6.36	-4.15	3.62×10^{-5}
	15	rs624191	*LOC100128434* | *GPR176*	intergenic	37880127	T	958	-33.02	8.09	-4.08	4.83×10^{-5}
	8	rs78933755	*PPP1R3B* | *LOC100129150*	intergenic	9245811	G	958	105.00	25.75	4.08	4.89×10^{-5}
	1	rs113015095	*ATG4C*	intron	63067732	A	949	38.59	9.50	4.06	5.25×10^{-5}
	11	rs12574325	*BDNFOS*	intron	27569624	A	958	51.89	12.80	4.06	5.41×10^{-5}
	11	rs16917135	*BDNFOS*	intron	27571281	T	958	51.89	12.80	4.06	5.41×10^{-5}
	8	rs35584813	*PPP1R3B* | *LOC100129150*	intergenic	9245040	C	958	102.30	25.30	4.04	5.74×10^{-5}
HC	2	rs4411698	*LOC728241* | *CNTNAP5*	intergenic	123651577	C	958	-62.23	14.58	-4.26	2.17×10^{-5}
Model 2	2	rs6541885	*LOC728241* | *CNTNAP5*	intergenic	123668890	A	958	-66.78	15.92	-4.19	2.99×10^{-5}
	1	rs6425446	*LOC400796* | *SEC16B*	intergenic	176031200	A	958	25.07	5.992	4.18	3.13×10^{-5}
	8	rs78933755	*PPP1R3B* | *LOC100129150*	intergenic	9245811	G	958	106.4	25.51	4.17	3.29×10^{-5}
	1	rs113015095	*ATG4C*	intron	63067732	A	949	39.14	9.408	4.16	3.47×10^{-5}
	15	rs624191	*LOC100128434* | *GPR176*	intergenic	37880127	T	958	-32.81	8.015	-4.09	4.62×10^{-5}
	8	rs35584813	*PPP1R3B* | *LOC100129150*	intergenic	9245040	C	958	102.4	25.07	4.08	4.79×10^{-5}
	8	rs73535332	*PPP1R3B* | *LOC100129150*	intergenic	9255047	C	958	109.3	27.03	4.04	5.69×10^{-5}
	11	rs12574325	*BDNFOS*	intron	27569624	A	958	51.01	12.69	4.02	6.25×10^{-5}
	11	rs16917135	*BDNFOS*	intron	27571281	T	958	51.01	12.69	4.02	6.25×10^{-5}
WHR	1	rs10926800	*PLD5* | *LOC400723*	intergenic	240867832	T	958	-0.02	0.00	-4.73	2.65×10^{-6}
Model 1	2	rs11679288	*ALK*	intron	29494717	T	958	0.02	0.00	4.44	1.02×10^{-5}
	5	rs67960962	*HAVCR1*	intron	156417764	C	958	0.02	0.00	4.27	2.20×10^{-5}
	1	rs34996202	*REG4*	coding. 104/3	120142649	G	958	0.02	0.01	4.23	2.57×10^{-5}

WHR	6	rs76551104	*LRRC16*	intron	25547206	A	958	0.04	0.01	4.19	3.01×10^{-5}
	16	rs2037912	*PPL*	coding-1055/2;nonsyn c|s	4873940	G	958	0.06	0.01	4.16	3.44×10^{-5}
	1	rs12075846	*LRRC7 | PIN1L*	intergenic	70139675	C	958	0.02	0.00	4.16	3.46×10^{-5}
	18	rs10164068	*LOC100132894*	intron	72381591	C	957	-0.02	0.00	-4.16	3.47×10^{-5}
	5	rs17307165	*NDUFS4 | ARL15*	intergenic	53113382	G	956	0.03	0.01	4.12	4.16×10^{-5}
	8	rs6586919	*LZTS1 | GFRA2*	intergenic	20481302	A	958	0.02	0.00	4.10	4.41×10^{-5}
Fat mass Model 1	1	rs6425446	*LOC400796 | SEC16B*	intergenic	176031200	A	927	2320.00	475.90	4.88	1.28×10^{-6}
	2	rs6541885	*LOC728241 | CNTNAP5*	intergenic	123668890	A	927	-5919.00	1296.00	-4.57	5.59×10^{-6}
	2	rs4411698	*LOC728241 | CNTNAP5*	intergenic	123651577	C	927	-5264.00	1179.00	-4.46	9.04×10^{-6}
	1	rs4072161	*LOC400796 | SEC16B*	intergenic	176036310	G	927	2138.00	480.20	4.45	9.54×10^{-6}
	1	rs4311843	*DOCK7 | ATG4C*	intergenic	176035387	C	927	2237.00	506.20	4.42	1.11×10^{-5}
	9	rs10781091	*GDA | ZFAND5*	intergenic	74064112	A	927	3706.00	844.10	4.39	1.26×10^{-5}
	1	rs4348685	*LOC400796 | SEC16B*	intergenic	176034251	G	927	-2281.00	535.50	-4.26	2.25×10^{-5}
	1	rs4075235	*LOC400796 | SEC16B*	intergenic	176034945	T	927	-2281.00	535.50	-4.26	2.25×10^{-5}
	2	rs2114591	*SP110*	intron	230758813	T	926	2007.00	474.70	4.23	2.59×10^{-5}
Fat mass Model 2	1	rs6425446	*LOC400796 | SEC16B*	intergenic	176031200	A	927	2262.00	473.80	4.78	2.08×10^{-6}
	2	rs6541885	*LOC728241 | CNTNAP5*	intergenic	123668890	A	927	-5944.00	1288.00	-4.62	4.48×10^{-6}
	2	rs4411698	*LOC728241 | CNTNAP5*	intergenic	123651577	C	927	-5237.00	1172.00	-4.47	8.91×10^{-6}
	1	rs4072161	*LOC400796 | SEC16B*	intergenic	176036310	G	927	2095.00	477.70	4.39	1.29×10^{-5}
	1	rs4311843	*DOCK7 | ATG4C*	intergenic	176035387	C	927	2191.00	503.60	4.35	1.51×10^{-5}
	9	rs10781091	*GDA | ZFAND5*	intergenic	74064112	A	927	3582.00	840.60	4.26	2.24×10^{-5}
	1	rs4348685	*LOC400796 | SEC16B*	intergenic	176034251	G	927	-2259.00	532.50	-4.24	2.43×10^{-5}
	1	rs4075235	*LOC400796 | SEC16B*	intergenic	176034945	T	927	-2259.00	532.50	-4.24	2.43×10^{-5}
	2	rs2114591	*SP110*	intron	230758813	T	926	1987.00	472.20	4.21	2.83×10^{-5}
Fat mass Model 3	3	rs6802030	*LOC730168 | TBL1R1*	intergenic	177944881	C	927	1435.00	323.60	4.44	1.03×10^{-5}
	7	rs4729039	*ANKIB1 | GATAD1*	intergenic	91880904	C	927	2312.00	522.40	4.43	1.08×10^{-5}
	1	rs823114	*NUCKS1 | RAB7L1*	intergenic	203986155	A	927	1671.00	380.30	4.40	1.24×10^{-5}
	9	rs2066184	*OSTF1 | PCSK5*	intergenic	77680325	C	927	1727.00	396.10	4.36	1.46×10^{-5}
	7	rs2070971	*GCK*	intron	44164108	A	927	-1322.00	320.70	-4.12	4.10×10^{-5}
	17	rs191474766	*ZNF750*	intron	78387864	A	927	-1399.00	344.60	-4.06	5.32×10^{-5}
	17	rs114826858	*ZNF750*	intron	78387395	C	926	-1760.00	441.10	-3.99	7.12×10^{-5}
	1	rs6425446	*LOC400796 | SEC16B*	intergenic	176031200	A	927	1249.00	313.40	3.99	7.28×10^{-5}
	17	rs116685344	*ZNF750*	intron	78386414	T	927	-1737.00	439.50	-3.95	8.37×10^{-5}
	17	rs114391003	*ZNF750*	intron	78387085	G	927	-1737.00	439.50	-3.95	8.37×10^{-5}
	1	rs6425453	*LOC400796 | SEC16B*	intergenic	176085789	G	927	1221.00	310.00	3.94	8.79×10^{-5}

Lean mass	9	rs10781091	*GDA* | *ZFAND5*	intergenic	74064112	A	927	2389.00	480.50	4.97	7.89×10^{-7}
Model 1	14	rs34664345	*PNN* | *MIA2*	intergenic	38761479	G	927	1259.00	282.10	4.46	9.15×10^{-6}
	7	rs17443228	*LRRN3* | *IMMP2L*	intergenic	110774133	A	927	-2029.00	496.90	-4.08	4.84×10^{-5}
	8	rs6601538	*KR6*	intron	10812864	G	927	-1328.00	329.20	-4.03	5.95×10^{-5}
	5	rs17554781	*RGNEF* | *ENC1*	intergenic	73344284	T	927	1188.00	295.90	4.01	6.45×10^{-5}
	1	rs2375278	*RUN3* | *SYF2*	intergenic	25401625	A	927	5103.00	1294.00	3.94	8.65×10^{-5}
	1	rs6664016	*FAM5B* | *LOC400796*	intergenic	175587093	G	927	1198.00	304.10	3.94	8.76×10^{-5}
Lean mass	9	rs10781091	*GDA* | *ZFAND5*	intergenic	74064112	A	927	2116.00	438.70	4.82	1.66×10^{-6}
Model 2	14	rs34664345	*PNN* | *MIA2*	intergenic	38761479	G	927	1148.00	257.20	4.46	9.12×10^{-6}
	7	rs7808316	*DGKB* | *TMEM195*	intergenic	15112007	A	927	-1101.00	255.20	-4.31	1.79×10^{-5}
	8	rs73535324	*PPP1R3B* | *LOC100129150*	intergenic	9252643	A	927	4702.00	1101.00	4.27	2.15×10^{-5}
	8	rs73535332	*PPP1R3B* | *LOC100129150*	intergenic	9255047	C	927	4759.00	1124.00	4.23	2.53×10^{-5}
	15	rs624191	*LOC100128434* | *GPR176*	intergenic	37880127	T	927	-1377.00	329.50	-4.18	3.20×10^{-5}
	8	rs76966674	*LPL* | *SLC18A1*	intergenic	19987694	C	927	1876.00	451.80	4.15	3.61×10^{-5}
	8	rs11778774	*PPP1R3B* | *LOC100129150*	intergenic	9251517	G	927	4308.00	1040.00	4.14	3.77×10^{-5}
	8	rs35584813	*PPP1R3B* | *LOC100129150*	intergenic	9245040	C	927	4342.00	1060.00	4.10	4.53×10^{-5}
	8	rs36090863	*PPP1R3B* | *LOC100129150*	intergenic	9246378	T	927	4342.00	1060.00	4.10	4.53×10^{-5}
Lean mass	11	rs12364237	*COMMD9* | *FLJ14213*	intergenic	36318610	T	927	1899.00	449.80	4.22	2.66×10^{-5}
Model 4	9	rs116485958	*GLIS3*	intron	4257720	G	927	3350.00	797.00	4.20	2.89×10^{-5}
	9	rs2066184	*OSTF1* | *PCSK5*	intergenic	77680325	C	927	-857.30	207.70	-4.13	4.01×10^{-5}
	15	rs17270501	*RORA*	intron	58907943	T	927	2005.00	491.60	4.08	4.92×10^{-5}
	15	rs2677909	*SLCO3A1*	intron	90445178	T	927	-770.30	191.80	-4.02	6.41×10^{-5}
	2	rs1347684	*OC728597* | *LOC727982*	intergenic	4130803	T	927	680.30	171.10	3.98	7.54×10^{-5}
	3	rs2606736	*ATG7*	intron	11375249	A	927	693.50	176.20	3.94	8.88×10^{-5}
	15	rs192009379	*IQCH*	intron	65480944	G	927	-1215.00	309.40	-3.93	9.22×10^{-5}
	1	rs1146581	*ACADM*	intron	75985612	A	927	-631.30	161.30	-3.91	9.77×10^{-5}
PFM	2	rs4411698	*LOC728241* | *CNTNAP5*	intergenic	123651577	C	927	-3.76	0.77	-4.88	1.26×10^{-6}
Model 1	2	rs6541885	*LOC728241* | *CNTNAP5*	intergenic	123668890	A	927	-4.04	0.85	-4.76	2.21×10^{-6}
	1	rs2068973	*LOC400796* | *SEC16B*	intergenic	176059338	A	927	1.45	0.31	4.65	3.78×10^{-6}
	1	rs6425453	*LOC400796* | *SEC16B*	intergenic	176085789	G	927	1.40	0.31	4.50	7.74×10^{-6}
	1	rs10913437	*LOC400796* | *SEC16B*	intergenic	176062763	T	927	1.40	0.31	4.46	9.03×10^{-6}
	1	rs4072161	*LOC400796* | *SEC16B*	intergenic	176036310	G	927	1.40	0.31	4.45	9.64×10^{-6}
	1	rs1854288	*LOC400796* | *SEC16B*	intergenic	176069710	A	927	1.36	0.31	4.38	1.32×10^{-5}
	1	rs12092449	*LOC400796* | *SEC16B*	intergenic	176081551	C	927	1.35	0.31	4.34	1.59×10^{-5}
	1	rs4348685	*LOC400796* | *SEC16B*	intergenic	176034251	G	927	-1.46	0.35	-4.15	3.70×10^{-5}

PFM	2	rs4411698	*LOC728241 | CNTNAP5*	intergenic	123651577	C	927	-3.78	0.77	-4.91	1.06×10^{-6}
Model 2	2	rs6541885	*LOC728241 | CNTNAP5*	intergenic	123668890	A	927	-4.03	0.85	-4.76	2.20×10^{-6}
	1	rs2068973	*LOC400796 | SEC16B*	intergenic	176059338	A	927	1.44	0.31	4.63	4.17×10^{-6}
	1	rs6425446	*LOC400796 | SEC16B*	intergenic	176031200	A	927	1.42	0.31	4.56	5.86×10^{-6}
	1	rs4072161	*LOC400796 | SEC16B*	intergenic	176036310	G	927	1.43	0.31	4.54	6.32×10^{-6}
	1	rs6425453	*LOC400796 | SEC16B*	intergenic	176085789	G	927	1.40	0.31	4.52	7.09×10^{-6}
	1	rs10913437	*LOC400796 | SEC16B*	intergenic	176062763	T	927	1.38	0.31	4.43	1.07×10^{-5}
	1	rs1854288	*LOC400796 | SEC16B*	intergenic	176069710	A	927	1.36	0.31	4.40	1.21×10^{-5}
	1	rs12092449	*LOC400796 | SEC16B*	intergenic	176081551	C	927	1.35	0.31	4.36	1.48×10^{-5}

Table E6. Top associated SNPs with body composition traits in young adults following linear regression with adjustment for covariates

TRAIT	CHR	SNP	GENE SYMBOL	GENE LOCATION	BP	A1	NMI SS	BETA	SE	STAT	P
BMI	10	rs7072408	*PCBD1 | UNC5B*	intergenic	72419228	G	949	-0.02	0.00	-4.43	1.06×10^{-5}
Model 1	1	rs2753399	*ZFYVE9*	intron	52546945	A	953	0.07	0.02	4.33	1.66×10^{-5}
	3	rs157538	*EDEM1 | GRM7*	intergenic	6312837	T	951	0.02	0.00	4.21	2.77×10^{-5}
	19	rs8113016	*LOC284417*	intron	60525997	G	953	0.06	0.01	4.18	3.21×10^{-5}
	10	rs12570727	*SLC39A12 | CACNB2*	intergenic	18465525	A	953	0.01	0.00	4.16	3.45×10^{-5}
	1	rs7540766	*LOC643717*	intron	217154788	C	953	0.01	0.00	4.13	4.01×10^{-5}
	15	rs7168761	*C15orf29 | TMEM85*	intergenic	32293510	G	953	-0.01	0.00	-4.02	6.27×10^{-5}
	12	rs776037	*INHBC*	intron	56129281	C	953	-0.01	0.00	-3.95	8.52×10^{-5}
	12	rs35287743	*MVK | C12orf34*	intergenic	108541633	T	940	0.06	0.01	3.93	9.14×10^{-5}
	7	rs799214	*BAZ1B*	intron	72497728	A	953	0.03	0.01	3.92	9.45×10^{-5}
WC	1	rs2753399	*ZFYVE9*	intron	52546945	A	944	9.10	2.08	4.39	1.28×10^{-5}
Model 1	4	rs11735605	*EPHA5 | CENPC1*	intergenic	66366424	C	944	3.10	0.72	4.31	1.79×10^{-5}
	10	rs12570727	*SLC39A12 | CACNB2*	intergenic	18465525	A	944	1.89	0.45	4.17	3.36×10^{-5}
	8	rs13267357	*LOC646909 | TMEM66*	intergenic	29914768	A	944	3.53	0.85	4.14	3.85×10^{-5}
	5	rs4607330	*RAB9P1 | EFNA5*	intergenic	105035114	T	939	3.29	0.80	4.13	4.01×10^{-5}
	19	rs8113016	*LOC284417*	intron	60525997	G	944	7.55	1.85	4.08	4.80×10^{-5}
	11	rs114879235	*CENTD2*	intron	72091467	T	944	5.24	1.29	4.08	4.94×10^{-5}
WC	10	rs7094782	*KIAA1217*	intron	24665411	G	944	3.56	0.88	4.04	5.73×10^{-5}
Model 1	2	rs11679410	*PLCL1 | SATB2*	intergenic	199450476	G	944	9.07	2.25	4.03	6.15×10^{-5}
	20	rs6059958	*HM13*	intron	29606939	C	942	-1.82	0.45	-4.01	6.52×10^{-5}

WC	1	rs2753399	*ZFYVE9*	intron	52546945	A	944	8.93	2.05	4.37	1.41×10^{-5}
Model 2	10	rs12570727	*SLC39A12 | CACNB2*	intergenic	18465525	A	944	1.92	0.45	4.30	1.87×10^{-5}
		chr11:7209146									
	11	7	*CENTD2*	intron	72091467	T	944	5.43	1.27	4.29	1.97×10^{-5}
	8	rs13267357	*LOC646909 | TMEM66*	intergenic	29914768	A	944	3.58	0.84	4.26	2.27×10^{-5}
	4	rs11735605	*EPHA5 | CENPC1*	intergenic	66366424	C	944	2.96	0.71	4.17	3.37×10^{-5}
	5	rs4607330	*RAB9P1 | EFNA5*	intergenic	105035114	T	939	3.27	0.78	4.17	3.37×10^{-5}
	17	rs2006827	*GAS7*	intron	9798897	C	941	-1.60	0.40	-4.01	6.51×10^{-5}
	15	rs62017164	*TRPM7*	intron	48655616	T	944	8.37	2.10	3.98	7.36×10^{-5}
	15	rs62017165	*TRPM7*	intron	48655833	T	944	8.37	2.10	3.98	7.36×10^{-5}
	15	rs1060599	*TRPM7*	intron	48661924	T	944	8.37	2.10	3.98	7.36×10^{-5}
HC	10	rs12570727	*SLC39A12 | CACNB2*	intergenic	18465525	A	944	2.19	0.45	4.83	1.63×10^{-6}
Model 1	19	rs8113016	*LOC284417*	intron	60525997	G	944	8.70	1.86	4.68	3.25×10^{-6}
	10	rs7072408	*PCBD1 | UNC5B*	intergenic	72419228	G	940	-2.24	0.48	-4.66	3.70×10^{-6}
	1	rs2753399	*ZFYVE9*	intron	52546945	A	944	9.39	2.09	4.49	7.87×10^{-6}
	1	rs11688	*JUN*	coding	59020581	A	941	-2.28	0.53	-4.32	1.71×10^{-5}
	9	rs10812811	*LINGO2*	intron	28399229	T	944	3.05	0.71	4.29	1.98×10^{-5}
	12	rs74955017	*FON4 | MYO1H*	intergenic	108256738	T	943	6.50	1.55	4.19	3.07×10^{-5}
	10	rs7082582	*GPAM | TECTB*	intergenic	113981387	T	943	1.75	0.42	4.13	3.97×10^{-5}
	12	rs4766988	*PTPN11 | RPH3A*	intergenic	111589413	A	944	4.54	1.12	4.06	5.29×10^{-5}
	12	rs17824620	*PTPN11 | RPH3A*	intergenic	111585377	A	944	4.54	1.12	4.06	5.29×10^{-5}
HC	10	rs12570727	*SLC39A12 | CACNB2*	intergenic	18465525	A	944	2.23	0.45	5.00	6.80×10^{-7}
Model 2	1	rs2753399	*ZFYVE9*	intron	52546945	A	944	9.19	2.05	4.48	8.29×10^{-6}
	10	rs7072408	*PCBD1 | UNC5B*	intergenic	72419228	G	940	-2.10	0.47	-4.42	1.08×10^{-5}
	1	rs11688	*JUN*	coding	59020581	A	941	-2.26	0.52	-4.36	1.44×10^{-5}
	19	rs8113016	*LOC284417*	intron	60525997	G	944	7.94	1.83	4.34	1.60×10^{-5}
	9	rs10812811	*LINGO2*	intron	28399229	T	944	2.98	0.70	4.26	2.23×10^{-5}
	10	rs7082582	*GPAM | TECTB*	intergenic	113981387	T	943	1.77	0.42	4.25	2.39×10^{-5}
	10	rs9299702	*ITGB1 | LOC401640*	intergenic	33331688	G	943	-1.94	0.46	-4.17	3.31×10^{-5}
	21	rs2827546	*NCAM2 | LOC284821*	intergenic	22797455	A	938	-2.15	0.53	-4.10	4.60×10^{-5}
	12	rs776037	*INHBC*	intron	56129281	C	944	-1.68	0.41	-4.09	4.66×10^{-5}
WHR	4	rs13113267	*MGC48628*	intron	91825412	A	943	0.02	0.00	5.41	$8.24\ 10^{-8}$
Model 1	3	rs2400349	*EPHB1 | LOC645218*	intergenic	136041127	A	943	0.05	0.01	5.36	1.07×10^{-7}
	1	rs12095241	*WARS2*	intron	119431520	G	941	0.02	0.00	4.91	1.08×10^{-6}
	7	rs6972785	*JAZF1 | LOC100128081*	intergenic	28002776	C	942	0.02	0.00	4.72	2.71×10^{-6}

WHR	9	rs4382560	*LOC646609 | LOC402360*	intergenic	22854667	C	940	0.02	0.00	4.60	4.73×10^{-6}
	9	rs28617580	*C9orf96*	intron	135242773	T	943	0.03	0.01	4.49	8.20×10^{-6}
	2	rs2221999	*SCN2A | FAM130A2*	intergenic	166091596	G	943	0.02	0.00	4.47	8.68×10^{-6}
	9	rs34399743	*C9orf96*	intron	135241179	A	943	0.03	0.01	4.44	1.01×10^{-5}
	1	rs12088290	*WARS2*	intron	119385449	T	943	0.02	0.00	4.44	1.01×10^{-5}
	7	rs115602681	*DGKB*	intron	14563567	C	942	0.05	0.01	4.42	1.12×10^{-5}
Fat mass	4	rs4240248	*FAM114A1*	intron	38593309	A	787	2403.00	497.90	4.83	1.68×10^{-6}
Model 1	1	rs2753399	*ZFYVE9*	intron	52546945	A	787	7314.00	1567.00	4.67	3.58×10^{-6}
	22	rs11704615	*PDGFB | RPL3*	intergenic	38026780	G	787	5827.00	1309.00	4.45	9.74×10^{-6}
	10	rs17106320	*GRID1*	coding	87888709	A	787	-1647.00	382.00	-4.31	1.84×10^{-5}
	8	rs13267357	*LOC646909 | TMEM66*	intergenic	29914768	A	787	2706.00	635.50	4.26	2.31×10^{-5}
	12	rs74955017	*FON4 | MYO1H*	intergenic	108256738	T	787	4891.00	1154.00	4.24	2.50×10^{-5}
	3	rs16831219	*PE5L | LOC131054*	intergenic	181389359	A	786	6335.00	1518.00	4.17	3.36×10^{-5}
	5	rs740366	*CAST*	intron	136302695	A	786	2283.00	547.30	4.17	3.38×10^{-5}
	10	rs10827955	*PLDC2*	intron	20405001	G	787	-1310.00	314.30	-4.17	3.41×10^{-5}
	16	rs935753	*LOC644649 | LOC729159*	intergenic	58934916	A	781	2156.00	509.10	4.24	2.55×10^{-5}
	4	rs4240248	*FAM114A1*	intron	38593309	A	787	1772.00	421.10	4.21	2.88×10^{-5}
Fat mass	8	rs6421010	*ZFAT1*	intron	135582302	C	786	-2156.00	528.80	-4.08	5.03×10^{-5}
Model 2	8	rs13267357	*LOC646909 | TMEM66*	intergenic	29914768	A	787	2142.00	535.80	4.00	7.03×10^{-5}
	2	rs4672376	*BCL11A*	intergenic	60347849	G	786	-1026.00	260.50	-3.94	8.94×10^{-5}
	10	rs2487927	*KIAA1462*	intron	30351372	T	787	-1696.00	431.20	-3.93	9.11×10^{-5}
	10	rs2478835	*KIAA1462*	coding	30357955	T	787	-1695.00	432.00	-3.92	9.52×10^{-5}
Lean mass	19	rs8113016	*LOC284417*	intron	60525997	G	787	6568.00	1275.00	5.15	3.28×10^{-7}
Model 1	11	rs11226119	*PDGFD | DDI1*	intergenic	103 404484	T	787	-2663.00	577.60	-4.61	4.68×10^{-6}
	20	rs10485674	*DH35 | MAFB*	intergenic	38727246	G	787	5488.00	1211.00	4.53	6.83×10^{-6}
	1	rs587271	*SSBP3*	intron	54515699	C	787	6301.00	1460.00	4.32	1.79×10^{-5}
	11	rs11224449	*FLJ32810*	intron	100166537	T	787	4380.00	1022.00	4.28	2.06×10^{-5}
	4	rs4235133	*YIPF7 | GUF1*	intergenic	44350101	A	787	1198.00	281.60	4.26	2.35×10^{-5}
	4	rs11734664	*STK32B*	intron	5276947	T	787	2065.00	487.40	4.24	2.54×10^{-5}
	9	rs17251166	*FKTN | TAL2*	intergenic	107449435	C	787	-1222.00	297.40	-4.11	4.37×10^{-5}
	15	rs1044355	*ULK3 | SCAMP2*	intergenic	72924117	A	787	-4246.00	1042.00	-4.07	5.11×10^{-5}
	9	rs73644048	*LINGO2*	intron	28475113	G	786	3659.00	904.80	4.04	5.77×10^{-5}
Lean mass	20	rs10485674	*DH35 | MAFB*	intergenic	38727246	G	787	4649.00	986.40	4.71	2.89×10^{-6}
Model 2	15	rs17158168	*HOMER2*	intron	81356715	A	786	1402.00	305.80	4.59	5.26×10^{-6}
	1	rs587271	*SSBP3*	intron	54515699	C	787	5335.00	1189.00	4.49	8.25×10^{-6}

Lean mass	6	rs78936838	*LRRC16*	intron	25402494	A	787	-1808.00	416.60	-4.34	1.60×10^{-5}
	19	rs8113016	*LOC284417*	intron	60525997	G	787	4547.00	1048.00	4.34	1.63×10^{-5}
			MULTIPLE_GENES:2613								
	3	rs11705953	*7.100132655*	complex	115768609	A	787	4510.00	1077.00	4.19	3.14×10^{-5}
	6	rs115551443	*PHACTR1*	intron	13031785	G	787	-1413.00	338.60	-4.17	3.36×10^{-5}
	16	rs77685839	*NRN1L*	coding	66477488	G	782	-1395.00	342.70	-4.07	5.17×10^{-5}
	10	rs10509761	*AS3MT*	intron	104622759	C	786	1358.00	338.00	4.02	6.43×10^{-5}
	6	rs74829142	*LRRC16*	intron	25408508	A	787	-1575.00	398.20	-3.96	8.33×10^{-5}
Lean mass	20	rs10485674	*DH35 | MAFB*	intergenic	38727246	G	787	4092.00	829.60	4.93	9.91×10^{-7}
Model 4	7	rs62451130	*JAZF1*	intron	27943699	T	782	4431.00	932.10	4.75	2.39×10^{-6}
	17	rs1526187	*CA10*	intron	47169634	G	786	-999.20	220.50	-4.53	6.75×10^{-6}
	2	rs75276762	*KIAA1486 | IRS1*	intergenic	226755929	T	786	-875.50	207.90	-4.21	2.85×10^{-5}
	6	chr6:25402494	*LRRC16*	intron	25402494	A	787	-1477.00	351.30	-4.21	2.91×10^{-5}
	8	rs6989280	*TRIB1 | LOC100130231*	intergenic	126577928	G	787	-1253.00	298.80	-4.19	3.06×10^{-5}
	12	rs79224935	*ATP2B1*	utr	88506542	A	787	1456.00	348.90	4.17	3.32×10^{-5}
	12	rs74940559	*ATP2B1*	intron	88530271	C	787	1456.00	348.90	4.17	3.32×10^{-5}
	1	rs587271	*SSBP3*	intron	54515699	C	787	4170.00	1004.00	4.15	3.66×10^{-5}
	7	rs62451122	*JAZF1*	intron	27921261	T	787	3822.00	934.30	4.09	4.74×10^{-5}
PFM	4	rs4240248	*FAM114A1*	intron	38593309	A	787	2.21	0.49	4.56	5.97×10^{-6}
Model 1	17	rs10512513	*PRKCA*	intron	61965600	G	787	1.50	0.35	4.34	1.64×10^{-5}
	2	rs7576822	*NBEAL1*	intron	203782462	T	787	4.52	1.12	4.03	6.13×10^{-5}
	2	rs9678194	*NBEAL1*	intron	203760073	G	786	4.52	1.12	4.02	6.30×10^{-5}
	12	rs11838318	*DYRK2 | IFNG*	intergenic	66558171	C	787	1.95	0.49	4.02	6.32×10^{-5}
	1	rs16852018	*LOC400796 | SEC16B*	intergenic	176098880	A	786	-1.32	0.33	-4.00	6.84×10^{-5}
	2	rs137874153	*WDR12*	intron	203461410	C	787	5.45	1.37	3.98	7.40×10^{-5}
	2	rs114986742	*LOC100129743 | NBEAL1*	intergenic	203627965	T	787	5.01	1.28	3.92	9.58×10^{-5}
	2	rs116337069	*RAPH1 | CD28*	intergenic	204134869	T	787	5.01	1.28	3.92	9.58×10^{-5}
	8	rs13267357	*LOC646909 | TMEM66*	intergenic	29914768	A	787	5.01	1.28	3.92	9.58×10^{-5}
PFM	4	rs4240248	*FAM114A1*	intron	38593309	A	787	2.21	0.49	4.56	5.97×10^{-6}
Model 2	17	rs10512513	*PRKCA*	intron	61965600	G	787	1.50	0.35	4.34	1.64×10^{-5}
	2	rs7576822	*NBEAL1*	intron	203782462	T	787	4.52	1.12	4.03	6.13×10^{-5}
	2	rs9678194	*NBEAL1*	intron	203760073	G	786	4.52	1.12	4.02	6.30×10^{-5}
	2	rs4674359	*AO1 | AO2*	intergenic	201302574	C	787	1.95	0.49	4.02	6.32×10^{-5}
	12	rs11838318	*DYRK2 | IFNG*	intergenic	66558171	C	786	-1.32	0.33	-4.00	6.84×10^{-5}
	1	rs16852018	*LOC400796 | SEC16B*	intergenic	176098880	A	787	5.45	1.37	3.98	7.40×10^{-5}
	2	rs137874153	*WDR12*	intron	203461410	C	787	5.01	1.28	3.92	9.58×10^{-5}

PFM	2	rs114986742	*LOC100129743 | NBEAL1*	intergenic	203627965	T	787	5.01	1.28	3.92	9.58×10^{-5}
	2	rs116337069	*RAPH1 | CD28*	intergenic	204134869	T	787	5.01	1.28	3.92	9.58×10^{-5}

Table E7. Top associated SNPs with body composition traits in young female adults following linear regression with adjustment for covariates

TRAIT	CHR	SNP	GENE SYMBOL	GENE LOCATION	BP	A1	NMISS	BETA	SE	STAT	P
BMI	3	rs149090	*EDEM1 | GRM7*	intergenic	6486781	G	448	0.12	0.02	4.85	1.70×10^{-6}
Model 1	5	rs32479	*ST8SIA4*	intron	100200494	G	448	0.13	0.03	4.33	1.82×10^{-5}
	3	rs9836894	*PBRM1*	intron	52691165	G	448	0.07	0.02	4.26	2.46×10^{-5}
	3	rs76913378	*GLT8D1*	intron	52706513	C	448	0.07	0.02	4.26	2.46×10^{-5}
	3	rs13325325	*NEK4*	intron	52778629	G	448	0.07	0.02	4.26	2.46×10^{-5}
	16	rs113530185	*DHODH | HP*	intergenic	70638303	C	448	0.06	0.01	4.22	2.99×10^{-5}
	3	rs9824342	*NISCH*	intron	52486947	G	448	0.06	0.02	4.18	3.58×10^{-5}
	3	rs9813653	*DNAH1*	intron	52366874	T	448	0.05	0.01	4.14	4.11×10^{-5}
	3	rs9828432	*DNAH1*	intron	52375005	A	448	0.05	0.01	4.14	4.11×10^{-5}
	3	rs9843055	*PHF7 | SEMA3G*	intergenic	52434363	C	448	0.05	0.01	4.14	4.11×10^{-5}
	3	rs34540591	*SEMA3G*	coding	52449814	T	448	0.05	0.01	4.14	4.11×10^{-5}
WC	17	rs2006827	*GAS7*	intron	9798897	C	440	-3.12	0.69	-4.49	9.14×10^{-6}
Model 1	5	rs32479	*ST8SIA4*	intron	100200494	G	442	17.60	4.06	4.34	1.79×10^{-5}
	3	rs149090	*EDEM1 | GRM7*	intergenic	6486781	G	442	14.10	3.27	4.31	2.03×10^{-5}
	11	rs114431455	*SLC35C1 | CRY2*	intergenic	45795746	C	442	8.09	1.89	4.29	2.20×10^{-5}
	5	rs4607330	*RAB9P1 | EFNA5*	intergenic	105035114	T	440	5.29	1.24	4.25	2.60×10^{-5}
	15	rs34173668	*VPS13C*	intron	60092876	C	442	9.42	2.29	4.12	4.51×10^{-5}
	15	rs34958705	*VPS13C*	intron	60104049	G	442	9.42	2.29	4.12	4.51×10^{-5}
	15	rs34311301	*VPS13C*	intron	60129847	C	442	9.42	2.29	4.12	4.51×10^{-5}
	11	rs80081080	*SLC35C1 | CRY2*	intergenic	45796003	G	442	6.84	1.66	4.11	4.64×10^{-5}
	15	rs17271312	*VPS13C*	intron	60123571	C	439	9.40	2.29	4.10	4.97×10^{-5}
WC	17	rs2006827	*GAS7*	intron	9798897	C	440	-3.17	0.68	-4.64	4.62×10^{-6}
Model 2	5	rs32479	*ST8SIA4*	intron	100200494	G	442	17.54	4.00	4.38	1.48×10^{-5}
	5	rs4607330	*RAB9P1 | EFNA5*	intergenic	105035114	T	440	5.22	1.23	4.26	2.56×10^{-5}
	3	rs149090	*EDEM1 | GRM7*	intergenic	6486781	G	442	13.70	3.23	4.24	2.74×10^{-5}

WC	9	rs10810279	*FREM1*	intron	14868094	A	442	3.49	0.85	4.10	4.91×10^{-5}
	7	rs10269203	*POU6F2*	intron	39430634	T	441	2.93	0.71	4.09	5.06×10^{-5}
	3	rs117195838	*BTD*	intron	15625332	T	442	10.27	2.54	4.04	6.42×10^{-5}
	10	rs16930166	*LYZL1 | LOC387647*	intergenic	29647143	T	442	8.75	2.17	4.03	6.59×10^{-5}
	2	rs114285212	*LOC727944 | TMEM18*	intergenic	630159	A	442	12.39	3.09	4.01	7.02×10^{-5}
HC	5	rs32479	*ST8SIA4*	intron	100200494	G	443	20.28	3.97	5.11	4.80×10^{-7}
Model 1	3	rs149090	*EDEM1 | GRM7*	intergenic	6486781	G	443	15.27	3.21	4.76	2.68×10^{-6}
	10	rs11598794	*SGMS1*	intron	51766998	A	442	10.38	2.35	4.42	1.25×10^{-5}
	11	rs114431455	*SLC35C1 | CRY2*	intergenic	45795746	C	443	8.15	1.86	4.39	1.42×10^{-5}
	16	rs3091402	*IGSF6*	intron	21571296	A	443	4.09	0.95	4.33	1.85×10^{-5}
	6	rs761840	*TMEM200A | LOC285733*	intergenic	131155994	C	443	9.64	2.26	4.27	2.45×10^{-5}
	3	rs9836894	*PBRM1*	intron	52691165	G	443	8.73	2.05	4.25	2.56×10^{-5}
	3	rs76913378	*GLT8D1*	intron	52706513	C	443	8.73	2.05	4.25	2.56×10^{-5}
	3	rs13325325	*NEK4*	intron	52778629	G	443	8.73	2.05	4.25	2.56×10^{-5}
	11	rs7935037	*CHST1 | DKFZp779M0652*	intergenic	45679874	G	443	18.22	4.32	4.22	3.02×10^{-5}
HC	5	rs32479	*ST8SIA4*	intron	100200494	G	443	20.20	3.90	5.18	3.40×10^{-7}
Model 2	3	rs149090	*EDEM1 | GRM7*	intergenic	6486781	G	443	14.84	3.16	4.69	3.61×10^{-6}
	16	rs3091402	*IGSF6*	intron	21571296	A	443	4.25	0.93	4.58	6.17×10^{-6}
	6	rs761840	*TMEM200A | LOC285733*	intergenic	131155994	C	443	9.85	2.22	4.44	1.15×10^{-5}
	3	rs117195838	*BTD*	intron	15625332	T	443	10.85	2.49	4.35	1.68×10^{-5}
	12	rs797611	*PTPN11 | RPH3A*	intergenic	111533680	T	443	9.83	2.26	4.35	1.70×10^{-5}
	12	rs4766988	*PTPN11 | RPH3A*	intergenic	111589413	A	443	9.59	2.21	4.33	1.85×10^{-5}
	12	rs17824620	*PTPN11 | RPH3A*	intergenic	111585377	A	443	9.59	2.21	4.33	1.85×10^{-5}
	11	rs7935037	*CHST1 | DKFZp779M0652*	intergenic	45679874	G	443	18.11	4.25	4.26	2.48×10^{-5}
	3	rs9836894	*PBRM1*	intron	52691165	G	443	8.44	2.02	4.18	3.59×10^{-5}
	3	rs76913378	*GLT8D1*	intron	52706513	C	443	8.44	2.02	4.18	3.59×10^{-5}
WHR	9	rs1571578	*ZNF618*	intron	115754911	A	442	0.11	0.02	5.23	2.69×10^{-7}
Model 1	11	rs78366803	*LOC100128354 | MTNR1B*	intergenic	92331703	C	442	0.09	0.02	4.65	4.45×10^{-6}
	1	rs72929703	*ANKRD38 | USP1*	intergenic	62613996	A	442	0.04	0.01	4.60	5.56×10^{-6}
	11	rs11607061	*TMEM135 | RAB38*	intergenic	86737757	A	442	0.12	0.03	4.52	8.11×10^{-6}
	3	rs118152687	*CADM2*	intron	85886949	A	442	0.05	0.01	4.44	1.14×10^{-5}
	6	rs1247330	*PLG | MAP3K4*	intergenic	161246640	C	442	-0.02	0.00	-4.43	1.17×10^{-5}
	1	rs114779017	*LDLRAP1*	intron	25752966	T	436	0.03	0.01	4.36	1.63×10^{-5}
	1	rs114995236	*LDLRAP1*	intron	25751183	T	442	0.03	0.01	4.36	1.64×10^{-5}
	1	rs75401001	*LDLRAP1*	intron	25752594	G	442	0.03	0.01	4.36	1.64×10^{-5}

Fat mass	3	rs117195838	*BTD*	intron	15625332	T	373	1.01×10^{4}	1.98×10^{3}	5.07×10^{2}	6.36×10^{-7}
Model 1	19	rs12611418	*LOC389435 | OPRM1*	intergenic	339413	C	373	2.54×10^{4}	5.51×10^{3}	4.60×10^{2}	5.71×10^{-6}
	2	rs3934784	*LOC100131211 | NAB1*	intergenic	191156494	C	373	3.55×10^{4}	7.78×10^{3}	4.56×10^{2}	7.05×10^{-6}
	2	rs13005335	*LOC100131211 | NAB1*	intergenic	191179265	G	373	3.55×10^{4}	7.78×10^{3}	4.56×10^{2}	7.05×10^{-6}
	11	rs11224449	*FLJ32810*	intron	100166537	T	373	1.03×10^{4}	2.28×10^{3}	4.52×10^{2}	8.31×10^{-6}
	10	rs17106320	*GRID1*	coding	87888709	A	373	-3.07×10^{3}	7.07×10^{2}	-4.34×10^{2}	1.87×10^{-5}
	2	rs114285212	*LOC727944 | TMEM18*	intergenic	630159	A	373	1.21×10^{4}	2.78×10^{3}	4.33×10^{2}	1.91×10^{-5}
	2	rs78501377	*LOC727944 | TMEM18*	intergenic	630682	C	373	1.21×10^{4}	2.78×10^{3}	4.33×10^{2}	1.91×10^{-5}
	2	rs76275602	*LOC727944 | TMEM18*	intergenic	631877	A	373	1.21×10^{4}	2.78×10^{3}	4.33×10^{2}	1.91×10^{-5}
	2	rs78460669	*LOC727944 | TMEM18*	intergenic	633005	C	373	1.21×10^{4}	2.78×10^{3}	4.33×10^{2}	1.91×10^{-5}
Fat mass	3	rs117195838	*BTD*	intron	15625332	T	373	1.01×10^{4}	1.98×10^{3}	5.11×10^{2}	5.32×10^{-7}
Model 2	19	rs12611418	*LOC389435 | OPRM1*	intergenic	339413	C	373	2.50×10^{4}	5.53×10^{3}	4.52×10^{2}	8.36×10^{-6}
	2	rs3934784	*LOC100131211 | NAB1*	intergenic	191156494	C	373	3.50×10^{4}	7.80×10^{3}	4.48×10^{2}	9.92×10^{-6}
	2	rs13005335	*LOC100131211 | NAB1*	intergenic	191179265	G	373	3.50×10^{4}	7.80×10^{3}	4.48×10^{2}	9.92×10^{-6}
	11	rs11224449	*FLJ32810*	intron	100166537	T	373	1.02×10^{4}	2.29×10^{3}	4.43×10^{2}	1.24×10^{-5}
	10	rs17106320	*GRID1*	coding	87888709	A	373	-3.12×10^{3}	7.07×10^{2}	-4.41×10^{2}	1.36×10^{-5}
	2	rs114285212	*LOC727944 | TMEM18*	intergenic	630159	A	373	1.20×10^{4}	2.79×10^{3}	4.29×10^{2}	2.27×10^{-5}
	2	rs78501377	*LOC727944 | TMEM18*	intergenic	630682	C	373	1.20×10^{4}	2.79×10^{3}	4.29×10^{2}	2.27×10^{-5}
	2	rs76275602	*LOC727944 | TMEM18*	intergenic	631877	A	373	1.20×10^{4}	2.79×10^{3}	4.29×10^{2}	2.27×10^{-5}
	2	rs78460669	*LOC727944 | TMEM18*	intergenic	633005	C	373	1.20×10^{4}	2.79×10^{3}	4.29×10^{2}	2.27×10^{-5}
	2	rs78559588	*LOC727944 | TMEM18*	intergenic	637656	C	373	1.20×10^{4}	2.79×10^{3}	4.29×10^{2}	2.27×10^{-5}
Fat mass	16	rs935753	*LOC644649 | LOC729159*	intergenic	58934916	A	373	3284.00	754.40	4.35	1.75×10^{-5}
Model 3	12	rs10773511	*LOC644489 | TMEM132C*	intergenic	127295998	T	373	1812.00	425.60	4.26	2.64×10^{-5}
	3	rs7810655	*COLQ*	intron	15509721	A	373	3651.00	859.70	4.25	2.74×10^{-5}
	3	rs80245351	*COLQ*	intron	15510186	T	373	3651.00	859.70	4.25	2.74×10^{-5}
	3	rs79803475	*COLQ*	intron	15510923	C	373	3651.00	859.70	4.25	2.74×10^{-5}
	3	rs76214503	*COLQ*	intron	15511325	C	373	3651.00	859.70	4.25	2.74×10^{-5}
	3	rs77099202	*COLQ*	intron	15511710	A	373	3651.00	859.70	4.25	2.74×10^{-5}
	3	rs74519258	*COLQ*	intron	15526083	T	373	3651.00	859.70	4.25	2.74×10^{-5}
	3	rs74781229	*COLQ*	intron	15526219	T	373	3651.00	859.70	4.25	2.74×10^{-5}
	3	rs78249903	*COLQ*	intron	15528593	C	373	3651.00	859.70	4.25	2.74×10^{-5}
Lean mass	2	rs1477514	*NCK2 | LOC100132455*	intergenic	105896372	A	373	2866.00	679.00	4.22	3.06×10^{-5}
Model 1	11	rs90192	*SIDT2*	intron	116564557	G	373	1479.00	351.70	4.21	3.27×10^{-5}
	11	rs6590779	*LOC729305*	intergenic	134317533	T	372	-1614.00	385.40	-4.19	3.54×10^{-5}

Lean mass	12	rs74103533	*LOC729298 | HMGA2*	intergenic	64496995	T	372	2161.00	539.90	4.00	7.60×10^{-5}
	4	rs11734664	*STK32B*	intron	5276947	T	373	2770.00	692.30	4.00	7.64×10^{-5}
	11	rs236916	*PCSK7*	intron	116594838	A	373	1536.00	386.50	3.98	8.46×10^{-5}
	4	rs7434427	*LOC285501 | LOC728081*	intergenic	179754269	T	373	1488.00	375.60	3.96	8.92×10^{-5}
	8	rs6472822	*LY96 | JPH1*	intergenic	75213296	T	373	1907.00	483.60	3.94	9.59×10^{-5}
Lean mass	12	rs1373573	*LOC100131677 | PDE3A*	intergenic	20424861	T	373	1199.00	275.70	4.35	1.76×10^{-5}
Model 2	3	chr3:186006566	*DGKG*	coding	187489260	G	373	3317.00	799.00	4.15	4.10×10^{-5}
	4	rs2348638	*RAPGEF2 | FSTL5*	intergenic	160983514	G	373	-1359.00	328.40	-4.14	4.34×10^{-5}
	10	rs12570727	*SLC39A12 | CACNB2*	intergenic	18465525	A	373	1298.00	318.40	4.08	5.64×10^{-5}
	16	rs116027434	*EDC4*	intron	66475231	T	373	5160.00	1292.00	3.99	7.89×10^{-5}
Lean mass	18	rs2163544	*LOC388474*	intron	35139073	A	373	966.50	220.90	4.38	1.58×10^{-5}
Model 4	18	rs16971807	*LOC388474*	intron	35152687	A	372	996.60	232.00	4.30	2.22×10^{-5}
	8	rs2978056	*LOC100129104 | ZFAT1*	intergenic	134693776	T	373	-909.10	218.10	-4.17	3.83×10^{-5}
	11	rs12288732	*ADAMTS15 | C11orf44*	intergenic	129939477	A	373	-865.10	208.50	-4.15	4.14×10^{-5}
	8	rs602382	*UNC5D | KCNU1*	intergenic	36223678	T	372	-858.00	211.40	-4.06	6.04×10^{-5}
	5	rs1700574	*LOC729506 | LOC100128382*	intergenic	8595619	C	373	1312.00	325.80	4.03	6.90×10^{-5}
	1	rs6683071	*FLJ43505*	coding	220989974	A	373	959.50	238.40	4.02	6.94×10^{-5}
	2	rs2042144	*NAG*	intron	15310767	T	373	-2933.00	735.10	-3.99	7.95×10^{-5}
	11	rs90192	*SIDT2*	intron	116564557	G	373	816.00	205.10	3.98	8.37×10^{-5}
	2	rs7606415	*YSK4*	intron	135479450	T	372	-794.60	201.60	-3.94	9.70×10^{-5}
PFM	3	rs117195838	*BTD*	intron	15625332	T	373	7.80	1.78	4.39	1.46×10^{-5}
Model 1	15	rs78946279	*VPS33B*	intron	89348470	A	373	2.20	0.51	4.29	2.33×10^{-5}
	19	rs4802246	*SFRS16*	intron	50255028	A	373	-2.93	0.69	-4.27	2.54×10^{-5}
	2	rs3843862	*C2orf3 | LRRTM4*	intergenic	76096624	C	373	-3.42	0.84	-4.06	5.91×10^{-5}
	10	rs17106320	*GRID1*	coding	87888709	A	373	-2.54	0.63	-4.03	6.67×10^{-5}
	16	rs6501109	*A2BP1 | LOC283953*	intergenic	8398889	A	371	1.97	0.50	3.95	9.49×10^{-5}
PFM	19	rs4802246	*SFRS16*	intron	50255028	A	373	-3.07	0.68	-4.55	7.47×10^{-6}
Model 2	3	rs117195838	*BTD*	intron	15625332	T	373	7.65	1.75	4.38	1.57×10^{-5}
	15	rs78946279	*VPS33B*	intron	89348470	A	373	2.05	0.51	4.03	6.85×10^{-5}
	16	rs6501109	*A2BP1 | LOC283953*	intergenic	8398889	A	371	1.94	0.49	3.95	9.41×10^{-5}

Table E8. Top associated SNPs with body composition traits in young male adults following linear regression with adjustment for covariates

TRAIT	CHR	SNP	GENE SYMBOL	GENE LOCATION	BP	A1	NMISS	BETA	SE	STAT	P
BMI	19	rs11670765	*LOC339344*	intron	51093838	T	505	0.02	0.00	5.16	$3.50\ 10^{-7}$
Model 1	19	rs8113016	*LOC284417*	intron	60525997	G	505	0.07	0.01	4.90	1.28×10^{-6}
	12	rs7961916	*TB3 | LOC100129020*	intergenic	113839509	A	505	0.02	0.00	4.66	4.06×10^{-6}
	15	rs17598819	*TRPM7*	intron	48701919	T	503	0.12	0.03	4.61	5.17×10^{-6}
	15	rs17598264	*USP50 | TRPM7*	intergenic	48632003	G	505	0.12	0.03	4.60	5.27×10^{-6}
	15	rs62021060	*TRPM7*	utr	48640931	C	505	0.12	0.03	4.60	5.27×10^{-6}
	15	rs62017164	*TRPM7*	intron	48655616	T	505	0.12	0.03	4.60	5.27×10^{-6}
	15	rs62017165	*TRPM7*	intron	48655833	T	505	0.12	0.03	4.60	5.27×10^{-6}
	15	rs1060599	*TRPM7*	intron	48661924	T	505	0.12	0.03	4.60	5.27×10^{-6}
	15	rs62017202	*TRPM7*	intron	48736535	A	505	0.12	0.03	4.60	5.27×10^{-6}
WC	8	rs2013640	*DLC1*	intron	13133461	G	502	3.10	0.57	5.41	9.80×10^{-8}
Model 1	15	rs17598264	*USP50 | TRPM7*	intergenic	48632003	G	502	15.82	3.05	5.19	3.02×10^{-7}
	15	rs62021060	*TRPM7*	utr	48640931	C	502	15.82	3.05	5.19	3.02×10^{-7}
	15	rs62017164	*TRPM7*	intron	48655616	T	502	15.82	3.05	5.19	3.02×10^{-7}
	15	rs62017165	*TRPM7*	intron	48655833	T	502	15.82	3.05	5.19	3.02×10^{-7}
	15	rs1060599	*TRPM7*	intron	48661924	T	502	15.82	3.05	5.19	3.02×10^{-7}
	15	rs62017202	*TRPM7*	intron	48736535	A	502	15.82	3.05	5.19	3.02×10^{-7}
	15	rs62017207	*TRPM7*	intron	48745151	C	502	15.82	3.05	5.19	3.02×10^{-7}
	15	rs62017208	*TRPM7*	intron	48746646	A	502	15.82	3.05	5.19	3.02×10^{-7}
	15	rs62017209	*TRPM7*	intron	48747259	A	502	15.82	3.05	5.19	3.02×10^{-7}
WC	15	rs17598264	*USP50 | TRPM7*	intergenic	48632003	G	502	16.41	2.99	5.50	6.19×10^{-8}
Model 2	15	rs62021060	*TRPM7*	utr	48640931	C	502	16.41	2.99	5.50	6.19×10^{-8}
	15	rs62017164	*TRPM7*	intron	48655616	T	502	16.41	2.99	5.50	6.19×10^{-8}
	15	rs62017165	*TRPM7*	intron	48655833	T	502	16.41	2.99	5.50	6.19×10^{-8}
	15	rs1060599	*TRPM7*	intron	48661924	T	502	16.41	2.99	5.50	6.19×10^{-8}
	15	rs62017202	*TRPM7*	intron	48736535	A	502	16.41	2.99	5.50	6.19×10^{-8}
	15	rs62017207	*TRPM7*	intron	48745151	C	502	16.41	2.99	5.50	6.19×10^{-8}
	15	rs62017208	*TRPM7*	intron	48746646	A	502	16.41	2.99	5.50	6.19×10^{-8}
	15	rs62017209	*TRPM7*	intron	48747259	A	502	16.41	2.99	5.50	6.19×10^{-8}
	15	rs17520350	*TRPM7*	intron	48684965	C	501	16.41	2.99	5.49	6.31×10^{-8}

WHR	6	rs3923725	*LOC100129474 | SLC17A4*	intergenic	25842899	A	501	0.16	0.02	7.39	6.31 1013
Model 1	2	rs115743734	*COBLL1*	intron	165255601	A	502	0.09	0.01	5.91	6.34 1009
	3	rs17651358	*CNTN4*	intron	3039513	A	502	0.08	0.01	5.37	1.19×10^{-7}
	7	rs115252878	*DGKB*	intron	14546425	T	502	0.08	0.01	5.29	1.88×10^{-7}
	7	rs114209770	*DGKB*	intron	14547297	A	502	0.08	0.01	5.29	1.88×10^{-7}
	7	rs76358458	*DGKB*	intron	14548446	G	502	0.08	0.01	5.29	1.88×10^{-7}
	3	rs1543143	*SRGAP3*	intron	9087876	T	502	0.05	0.01	5.24	2.41×10^{-7}
	12	rs76712056	*FLJ40142*	intron	108989096	A	502	0.07	0.01	5.21	2.76×10^{-7}
	7	rs115602681	*DGKB*	intron	14563567	C	501	0.06	0.01	5.19	3.02×10^{-7}
	3	rs2400349	*EPHB1 | LOC645218*	intergenic	136041127	A	502	0.06	0.01	5.16	3.51×10^{-7}
HC	15	rs17598264	*USP50 | TRPM7*	intergenic	48632003	G	502	17.34	3.20	5.42	9.44×10^{-8}
Model 1	15	rs62021060	*TRPM7*	utr	48640931	C	502	17.34	3.20	5.42	9.44×10^{-8}
	15	rs62017164	*TRPM7*	intron	48655616	T	502	17.34	3.20	5.42	9.44×10^{-8}
	15	rs62017165	*TRPM7*	intron	48655833	T	502	17.34	3.20	5.42	9.44×10^{-8}
	15	rs1060599	*TRPM7*	intron	48661924	T	502	17.34	3.20	5.42	9.44×10^{-8}
	15	rs62017202	*TRPM7*	intron	48736535	A	502	17.34	3.20	5.42	9.44×10^{-8}
	15	rs62017207	*TRPM7*	intron	48745151	C	502	17.34	3.20	5.42	9.44×10^{-8}
	15	rs62017208	*TRPM7*	intron	48746646	A	502	17.34	3.20	5.42	9.44×10^{-8}
	15	rs62017209	*TRPM7*	intron	48747259	A	502	17.34	3.20	5.42	9.44×10^{-8}
	15	rs17598819	*TRPM7*	intron	48701919	T	500	17.34	3.20	5.41	9.59×10^{-8}
HC	15	rs17598264	*USP50 | TRPM7*	intergenic	48632003	G	502	18.04	3.12	5.79	1.28×10^{-8}
Model 2	15	rs62021060	*TRPM7*	utr	48640931	C	502	18.04	3.12	5.79	1.28×10^{-8}
	15	rs62017164	*TRPM7*	intron	48655616	T	502	18.04	3.12	5.79	1.28×10^{-8}
	15	rs62017165	*TRPM7*	intron	48655833	T	502	18.04	3.12	5.79	1.28×10^{-8}
	15	rs1060599	*TRPM7*	intron	48661924	T	502	18.04	3.12	5.79	1.28×10^{-8}
	15	rs62017202	*TRPM7*	intron	48736535	A	502	18.04	3.12	5.79	1.28×10^{-8}
	15	rs62017207	*TRPM7*	intron	48745151	C	502	18.04	3.12	5.79	1.28×10^{-8}
	15	rs62017208	*TRPM7*	intron	48746646	A	502	18.04	3.12	5.79	1.28×10^{-8}
	15	rs62017209	*TRPM7*	intron	48747259	A	502	18.04	3.12	5.79	1.28×10^{-8}
	15	rs17598819	*TRPM7*	intron	48701919	T	500	18.05	3.12	5.79	1.29×10^{-8}
Fat mass	1	rs9970334	*ICMT | C1orf211*	intergenic	6218825	T	414	22880.00	4001.00	5.72	2.07×10^{-8}
Model 1	1	rs16852018	*LOC400796 | SEC16B*	intergenic	176098880	A	414	6367.00	1128.00	5.65	3.06×10^{-8}
	11	rs77275360	*SLC22A18*	intron	2894032	T	411	4794.00	923.90	5.19	3.34×10^{-7}
	1	rs6670797	*GP7 | MGC52498*	intergenic	52870169	A	414	9112.00	1817.00	5.02	7.89×10^{-7}

Fat mass	1	rs2753399	*ZFYVE9*	intron	52546945	A	414	8935.00	1817.00	4.92	1.27×10^{-6}
	1	rs72941254	*NEGR1 | LOC100132353*	intergenic	72446505	A	414	3272.00	677.80	4.83	1.95×10^{-6}
	1	rs72941257	*NEGR1 | LOC100132353*	intergenic	72451438	C	414	3272.00	677.80	4.83	1.95×10^{-6}
	1	rs72941270	*NEGR1 | LOC100132353*	intergenic	72465674	A	414	3272.00	677.80	4.83	1.95×10^{-6}
	3	rs115473751	*EIF5A2*	intron	172106758	A	414	4924.00	1025.00	4.80	2.21×10^{-6}
	1	rs72941224	*NEGR1 | LOC100132353*	intergenic	72411598	T	411	3264.00	680.40	4.80	2.26×10^{-6}
Fat mass	1	rs16852018	*LOC400796 | SEC16B*	intergenic	176098880	A	414	6386.00	1122.00	5.69	2.41×10^{-8}
Model 2	1	rs9970334	*ICMT | C1orf211*	intergenic	6218825	T	414	22530.00	3992.00	5.64	3.13×10^{-8}
	11	rs77275360	*SLC22A18*	intron	2894032	T	411	4760.00	920.50	5.17	3.65×10^{-7}
	1	rs6670797	*GP7 | MGC52498*	intergenic	52870169	A	414	9072.00	1809.00	5.01	7.96×10^{-7}
	12	rs2116677	*OCC-1*	intron	104232358	C	414	7179.00	1440.00	4.99	9.15×10^{-7}
	1	rs2753399	*ZFYVE9*	intron	52546945	A	414	8854.00	1811.00	4.89	1.45×10^{-6}
	3	rs115473751	*EIF5A2*	intron	172106758	A	414	4982.00	1021.00	4.88	1.51×10^{-6}
	1	rs72941254	*NEGR1 | LOC100132353*	intergenic	72446505	A	414	3282.00	674.80	4.86	1.64×10^{-6}
	1	rs72941257	*NEGR1 | LOC100132353*	intergenic	72451438	C	414	3282.00	674.80	4.86	1.64×10^{-6}
	1	rs72941270	*NEGR1 | LOC100132353*	intergenic	72465674	A	414	3282.00	674.80	4.86	1.64×10^{-6}
	1	rs72941224	*NEGR1 | LOC100132353*	intergenic	72411598	T	411	3274.00	677.30	4.83	1.90×10^{-6}
Fat mass	1	rs16852018	*LOC400796 | SEC16B*	intergenic	176098880	A	414	6050.00	1054.00	5.74	1.83×10^{-8}
Model 3	1	rs6670797	*GP7 | MGC52498*	intergenic	52870169	A	414	8844.00	1695.00	5.22	2.87×10^{-7}
	11	rs77275360	*SLC22A18*	intron	2894032	T	411	4400.00	864.50	5.09	5.49×10^{-7}
	1	rs2753399	*ZFYVE9*	intron	52546945	A	414	8316.00	1701.00	4.89	1.46×10^{-6}
	2	rs77393265	*SUPT7L*	intron	27733704	A	414	5545.00	1155.00	4.80	2.21×10^{-6}
	1	rs72941254	*NEGR1 | LOC100132353*	intergenic	72446505	A	414	2966.00	636.20	4.66	4.26×10^{-6}
	1	rs72941257	*NEGR1 | LOC100132353*	intergenic	72451438	C	414	2966.00	636.20	4.66	4.26×10^{-6}
	1	rs72941270	*NEGR1 | LOC100132353*	intergenic	72465674	A	414	2966.00	636.20	4.66	4.26×10^{-6}
	4	rs7685796	*LOC152742 | LOC441009*	intergenic	14553996	A	414	6305.00	1357.00	4.65	4.55×10^{-6}
	1	rs72941224	*NEGR1 | LOC100132353*	intergenic	72411598	T	411	2955.00	638.30	4.63	4.96×10^{-6}
Lean mass	19	rs8113016	*LOC284417*	intron	60525997	G	414	8701.00	1655.00	5.26	2.37×10^{-7}
Model 1	1	rs587271	*SSBP3*	intron	54515699	C	414	9424.00	2078.00	4.54	7.58×10^{-6}
	1	rs59391530	*TAF1A | MIA3*	intergenic	220847044	A	414	3428.00	777.40	4.41	1.32×10^{-5}
	1	rs7525548	*TNNI3K*	intron	74774062	T	414	-1860.00	423.70	-4.39	1.44×10^{-5}
	8	rs2409658	*PIN1*	intron	10706375	A	414	-2133.00	498.20	-4.28	2.32×10^{-5}
	6	rs6924854	*PHACTR1*	intron	13229646	A	414	-2166.00	511.40	-4.24	2.81×10^{-5}
	8	rs9969626	*PIN1*	intron	10708211	A	413	-2133.00	504.90	-4.23	2.95×10^{-5}
	3	rs6441886	*CDCP1*	intron	45103101	C	414	-1638.00	387.80	-4.22	2.96×10^{-5}
	17	rs4459609	*CYB561 | LOC342541*	intergenic	58902680	C	414	-1636.00	395.30	-4.14	4.24×10^{-5}

	4	rs1844994	*LOC100132227*	intron	108417809	A	414	-2715.00	656.70	-4.13	4.32×10^{-5}
Lean mass	19	rs8113016	*LOC284417*	intron	60525997	G	414	6562.00	1336.00	4.91	1.31×10^{-6}
Model 2	1	rs9970334	*ICMT | C1orf211*	intergenic	6218825	T	414	19490.00	4371.00	4.46	1.07×10^{-5}
	7	rs62451130	*JAZF1*	intron	27943699	T	410	6552.00	1474.00	4.45	1.14×10^{-5}
	15	rs12917461	*LOC390638 | SV2B*	intergenic	89398076	A	414	-1349.00	306.70	-4.40	1.38×10^{-5}
	4	rs10030759	*FRAS1*	intron	79668559	G	414	-1449.00	331.70	-4.37	1.58×10^{-5}
	1	rs587271	*SSBP3*	intron	54515699	C	414	7271.00	1670.00	4.35	1.70×10^{-5}
	18	rs16951815	*LRRC30 | PTPRM*	intergenic	7456596	G	414	2140.00	511.60	4.18	3.52×10^{-5}
	5	rs9654427	*ANKRD31*	intron	74502589	G	414	9144.00	2198.00	4.16	3.89×10^{-5}
	10	rs7067741	*NT5C2*	intron	104894612	C	414	2015.00	486.90	4.14	4.24×10^{-5}
	10	rs113503123	*NT5C2 | LOC401648*	intergenic	104951623	C	414	1950.00	471.80	4.13	4.36×10^{-5}
	10	rs79562142	*CNNM2*	intron	104725395	C	413	2026.00	493.00	4.11	4.79×10^{-5}
Lean mass	7	rs62451130	*JAZF1*	intron	27943699	T	410	6771.00	1377.00	4.92	1.29×10^{-6}
Model 4	7	rs62451122	*JAZF1*	intron	27921261	T	414	6917.00	1563.00	4.43	1.23×10^{-5}
	15	rs12917461	*LOC390638 | SV2B*	intergenic	89398076	A	414	-1261.00	288.20	-4.38	1.54×10^{-5}
	18	rs9955441	*SLC14A2*	intron	41486630	C	414	-2930.00	672.10	-4.36	1.65×10^{-5}
	20	rs10485674	*DH35 | MAFB*	intergenic	38727246	G	414	5149.00	1204.00	4.28	2.35×10^{-5}
	18	rs16951815	*LRRC30 | PTPRM*	intergenic	7456596	G	414	2051.00	479.90	4.27	2.40×10^{-5}
	4	rs10030759	*FRAS1*	intron	79668559	G	414	-1329.00	312.10	-4.26	2.55×10^{-5}
	14	rs10148443	*LOC644584 | MBIP*	intergenic	35522543	G	414	1505.00	354.10	4.25	2.64×10^{-5}
	10	rs7067741	*NT5C2*	intron	104894612	C	414	1915.00	457.00	4.19	3.42×10^{-5}
PFM	11	rs77275360	*SLC22A18*	intron	2894032	T	411	5.76	1.14	5.06	$6.30\ 10^{-7}$
Model 1	1	rs16852018	*LOC400796 | SEC16B*	intergenic	176098880	A	414	7.01	1.40	5.01	$8.15\ 10^{-7}$
	1	rs6670797	*GP7 | MGC52498*	intergenic	52870169	A	414	10.60	2.25	4.72	3.24×10^{-6}
	1	rs72941254	*NEGR1 | LOC100132353*	intergenic	72446505	A	414	3.90	0.84	4.66	4.20×10^{-6}
	1	rs72941257	*NEGR1 | LOC100132353*	intergenic	72451438	C	414	3.90	0.84	4.66	4.20×10^{-6}
	1	rs72941270	*NEGR1 | LOC100132353*	intergenic	72465674	A	414	3.90	0.84	4.66	4.20×10^{-6}
	1	rs72941224	*NEGR1 | LOC100132353*	intergenic	72411598	T	411	3.89	0.84	4.64	4.82×10^{-6}
	14	rs10134920	*LOC100132612 | C14orf177*	intergenic	97789730	G	414	3.67	0.82	4.49	9.25×10^{-6}
	1	rs2753399	*ZFYVE9*	intron	52546945	A	414	9.99	2.25	4.44	1.15×10^{-5}
	3	rs58574369	*NT5DC2*	intron	52536300	A	414	4.48	1.02	4.40	1.41×10^{-5}
	2	rs7592118	CPS1 | ERBB4	intergenic	211913738	T	412	2.43	0.57	4.28	2.36×10^{-5}
PFM	11	rs77275360	SLC22A18	intron	2894032	T	411	5.78	1.14	5.07	5.99×10^{-7}
Model 2	1	rs16852018	LOC400796 | SEC16B	intergenic	176098880	A	414	7.00	1.40	5.00	8.51×10^{-7}

PFM	1	rs6670797	GP7 | MGC52498	intergenic	52870169	A	414	10.61	2.25	4.72	3.19×10^{-6}
	1	rs72941254	NEGR1 | LOC100132353	intergenic	72446505	A	414	3.90	0.84	4.66	4.35×10^{-6}
	1	rs72941257	NEGR1 | LOC100132353	intergenic	72451438	C	414	3.90	0.84	4.66	4.35×10^{-6}
	1	rs72941270	NEGR1 | LOC100132353	intergenic	72465674	A	414	3.90	0.84	4.66	4.35×10^{-6}
	1	rs72941224	NEGR1 | LOC100132353	intergenic	72411598	T	411	3.89	0.84	4.63	5.00×10^{-6}
	14	rs10134920	LOC100132612 | C14orf177	intergenic	97789730	G	414	3.68	0.82	4.50	8.76×10^{-6}
	1	rs2753399	ZFYVE9	intron	52546945	A	414	10.03	2.25	4.45	1.09×10^{-5}
	3	rs58574369	NT5DC2	intron	52536300	A	414	4.50	1.02	4.41	1.33×10^{-5}
	2	rs7592118	CPS1 | ERBB4	intergenic	211913738	T	412	2.45	0.57	4.31	2.05×10^{-5}

Section C: Results from mixed linear model association (MLMA)

In addition to the linear regression models outlined in Table 3.1. (Chapter 3) all the analyses presented in Tables E9 and E10 are also adjusted for relatedness and ten principle components.

Table E9 Top SNPs associated with body composition traits when datasets are combined.

Chr-chromosome, BP-base pair position, utr-untranslated region, in Plink the A1(affect allele) is the minor allele*, A2-major allele, MAF-minor allele frequency, Beta-refers to per allele change in the phenotype, SE- standard error and *P*-value adjusted for various covariates.

TRAIT	CHR	SNP	*GENE SYMBOL*	GENE LOCATION	BP	A1	A2	MAF	EFFECT (BETA)	SE	P
BMI	17	rs115012414	*APOH | PRKCA*	intergenic	61724373	C	T	0.04	-0.03	0.01	7.53×10^{-6}
Model 1	17	rs77612309	*APOH | PRKCA*	intergenic	61724990	C	T	0.04	-0.03	0.01	1.17×10^{-5}
	21	rs411697	*MULTIPLE_GENES:571.54073*	complex	29636469	A	G	0.48	-0.01	0.00	4.63×10^{-5}
	10	rs11015144	*APBB1IP*	intron	26839957	A	T	0.17	0.01	0.00	5.01×10^{-5}
	19	rs746075	*NUCB1*	intron	54108748	A	G	0.09	-0.02	0.00	5.45×10^{-5}
	2	rs2114591	*SP110*	intron	230758813	T	C	0.40	0.01	0.00	6.62×10^{-5}
	11	rs7106594	*INSC | SO6*	intergenic	15477573	C	T	0.30	0.01	0.00	7.40×10^{-5}

BMI	15	rs4775027	*AQP9 | LIPC*	intergenic	56375396	G	A	0.36	-0.01	0.00	7.48×10^{-5}
	2	rs7581351	*LOC344328 | LOC729141*	intergenic	188520180	A	G	0.42	0.01	0.00	7.77×10^{-5}
	14	rs10146149	*NRN3 | LOC100131580*	intergenic	78578318	T	C	0.11	-0.02	0.00	7.98×10^{-5}
WC	2	rs2114591	*SP110*	intron	230758813	T	C	0.40	1.66	0.37	8.26×10^{-6}
Model 1	10	rs1889516	*C10orf112*	intron	19637710	A	G	0.44	1.60	0.37	1.48×10^{-5}
	9	rs72656779	*LOC729983*	utr	22145640	A	G	0.07	3.21	0.75	1.88×10^{-5}
	9	rs10491772	*NFIB | ZDHHC21*	intergenic	14540138	C	T	0.01	6.84	1.62	2.33×10^{-5}
	14	rs11622292	*PRO1768 | FON3*	intergenic	89136080	T	C	0.14	2.16	0.51	2.67×10^{-5}
	9	rs7021554	*ANRIL | LOC729983*	intergenic	22142884	T	C	0.08	2.83	0.68	2.84×10^{-5}
	14	rs10146149	*NRN3 | LOC100131580*	intergenic	78578318	T	C	0.11	-2.44	0.59	3.85×10^{-5}
	20	rs8122057	*JAG1 | LOC728573*	intergenic	11085719	G	T	0.23	1.76	0.43	4.80×10^{-5}
	5	rs4703034	*SLC2A5 | GPR157*	intergenic	103839971	G	A	0.22	1.80	0.44	4.86×10^{-5}
	1	rs12239636	*STARD4 | C5orf13*	intergenic	9078288	T	C	0.03	-4.12	1.03	6.42×10^{-5}
	5	rs1598818	*NUDT12 | RAB9P1*	intergenic	110992117	T	C	0.37	-1.47	0.37	6.84×10^{-5}
WC	2	rs2114591	*SP110*	intron	230758813	T	C	0.40	1.62	0.37	1.10×10^{-5}
Model 2	10	rs1889516	*C10orf112*	intron	19637710	A	G	0.44	1.56	0.37	1.97×10^{-5}
	9	rs10491772	*NFIB | ZDHHC21*	intergenic	14540138	C	T	0.01	6.76	1.60	2.49×10^{-5}
	9	rs72656779	*LOC729983*	utr	22145640	A	G	0.07	3.12	0.74	2.84×10^{-5}
	14	rs11622292	*PRO1768 | FON3*	intergenic	89136080	T	C	0.14	2.12	0.51	3.27×10^{-5}
	1	rs12239636	*STARD4 | C5orf13*	intergenic	9078288	T	C	0.03	-4.18	1.02	4.44×10^{-5}
	14	rs10146149	*NRN3 | LOC100131580*	intergenic	78578318	T	C	0.11	-2.40	0.59	4.49×10^{-5}
	3	rs12493688	*HES1 | LOC100131551*	intergenic	195401542	G	T	0.01	7.42	1.82	4.74×10^{-5}
	9	rs7021554	*ANRIL | LOC729983*	intergenic	22142884	T	C	0.08	2.73	0.67	4.81×10^{-5}
	3	rs72625020	*IGF2BP2 | SFRS10*	intergenic	187079472	T	C	0.21	1.80	0.44	4.93×10^{-5}
	5	rs4703034	*SLC2A5 | GPR157*	intergenic	103839971	G	A	0.22	1.76	0.44	6.39×10^{-5}
HC	8	rs35584813	*PPP1R3B | LOC100129150*	intergenic	9245040	C	T	0.01	7.19	1.63	1.09×10^{-5}
Model 1	8	rs11778774	*PPP1R3B | LOC100129150*	intergenic	9251517	G	A	0.01	7.01	1.62	1.51×10^{-5}
	21	rs411697	*MULTIPLE_GENES:571.54073*	complex	29636469	A	G	0.48	-1.62	0.38	1.82×10^{-5}
	6	rs2474346	*GUCA1A*	intron	42244235	T	G	0.30	1.69	0.40	2.77×10^{-5}
	10	rs1889516	*C10orf112*	intron	19637710	A	G	0.44	1.59	0.38	3.17×10^{-5}
	8	rs78933755	*PPP1R3B | LOC100129150*	intergenic	9245811	G	A	0.01	7.01	1.69	3.44×10^{-5}
	12	rs3138139	*RDH5*	intron	54402150	G	A	0.31	-1.63	0.41	5.78×10^{-5}
	8	rs73535332	*PPP1R3B | LOC100129150*	intergenic	9255047	C	G	0.01	7.18	1.79	6.17×10^{-5}
	10	rs7072408	*PCBD1 | UNC5B*	intergenic	72419228	G	A	0.24	-1.77	0.44	6.19×10^{-5}
	19	rs12974306	*DNM2*	intron	10691281	T	G	0.13	-2.23	0.56	6.28×10^{-5}
	9	rs7021554	*PRO1768 | FON3*	intergenic	22142884	T	C	0.08	2.80	0.70	6.93×10^{-5}

HC	8	rs11778774	*PPP1R3B | LOC100129150*	intergenic	9251517	G	A	0.01	6.64	1.60	3.25×10^{-5}
Model 2	17	rs115012414	*APOH | PRKCA*	intergenic	61724373	C	T	0.04	-4.18	1.01	3.34×10^{-5}
	12	rs3138139	*RDH5*	intron	54402150	G	A	0.31	-1.64	0.40	4.07×10^{-5}
	10	rs1889516	*C10orf112*	intron	19637710	A	G	0.44	1.55	0.38	4.27×10^{-5}
	17	rs77612309	*APOH | PRKCA*	intergenic	61724990	C	T	0.04	-4.11	1.01	4.34×10^{-5}
	8	rs78933755	*PPP1R3B | LOC100129150*	intergenic	9245811	G	A	0.01	6.75	1.67	5.32×10^{-5}
	2	rs2190373	*RND3 | LOC344332*	intergenic	151482339	A	G	0.13	-2.27	0.57	6.04×10^{-5}
	20	rs115667632	*C20orf174*	intron	572110215	T	C	0.03	-4.37	1.10	6.81×10^{-5}
	8	rs73535332	*PPP1R3B | LOC100129150*	intergenic	9255047	C	G	0.01	6.97	1.77	8.11×10^{-5}
	19	rs12974306	*DNM2*	intron	10691281	T	G	0.13	-2.16	0.55	8.59×10^{-5}
	1	rs56750694	*WARS2*	intron	119382364	T	G	0.05	3.75	0.96	8.79×10^{-5}
WHR	16	rs2037912	*PPL*	coding	4873940	G	C	0.02	0.05	0.01	1.21×10^{-7}
Model 1	4	rs17605986	*RELL1 | LOC727821*	intergenic	37434719	A	G	0.03	0.03	0.01	1.05×10^{-6}
	4	rs11729192	*GRIA2 | C4orf18*	intergenic	158600300	G	A	0.21	0.01	0.00	7.00×10^{-6}
	4	rs17246641	*GRIA2 | C4orf18*	intergenic	158603278	G	A	0.21	0.01	0.00	7.21×10^{-6}
	15	rs11074155	*UNQ9370 | LOC728292*	intergenic	91627775	A	G	0.04	0.02	0.01	1.52×10^{-5}
	7	rs2289055	*DD56*	intron	44578415	G	A	0.04	0.02	0.01	1.74×10^{-5}
	7	rs7723	*DD56*	intron	44584318	A	G	0.04	0.02	0.01	1.74×10^{-5}
	12	rs1625560	*TMTC2 | SLC6A15*	intergenic	83719434	G	T	0.07	-0.02	0.00	2.73×10^{-5}
	3	rs13314361	*SCN5A*	intron	38641586	A	C	0.14	0.01	0.00	2.81×10^{-5}
	16	rs1861554	*FTO*	intron	52607268	G	A	0.07	0.02	0.00	2.91×10^{-5}
	2	rs1949915	*FLJ41046 | FLJ42418*	intergenic	6113406	A	G	0.09	-0.02	0.00	5.01×10^{-5}
	1	rs17023092	*WARS2*	utr	119375976	T	C	0.07	-0.02	0.00	9.11×10^{-5}
	1	rs74112264	*WARS2*	utr	119376488	A	G	0.07	-0.02	0.00	9.11×10^{-5}
	1	rs17023118	*WARS2*	utr	119381509	A	G	0.07	-0.02	0.00	9.11×10^{-5}
Fat mass	1	rs6664268	*LOC400796 | SEC16B*	intergenic	176031329	C	T	0.22	-1813.50	368.47	8.58×10^{-7}
Model 1	1	rs4075235	*LOC400796 | SEC16B*	intergenic	176034945	T	C	0.26	-1663.04	347.89	1.75×10^{-6}
	1	rs4072161	*LOC400796 | SEC16B*	intergenic	176036310	G	T	0.48	1453.94	314.57	3.80×10^{-6}
	9	rs10781091	*GDA | ZFAND5*	intergenic	74064112	A	G	0.08	2503.98	542.09	3.85×10^{-6}
	3	rs157538	*EDEM1 | GRM7*	intergenic	6312837	T	G	0.14	1996.18	446.37	7.75×10^{-6}
	1	rs6425446	*LOC400796 | SEC16B*	intergenic	176031200	A	G	0.42	1371.86	310.27	9.80×10^{-6}
	21	rs411697	*MULTIPLE_GENES:571.54073*	complex	29636469	A	G	0.48	-1300.83	302.41	1.70×10^{-5}
	1	rs11581129	*LOC400796 | SEC16B*	intergenic	176037754	A	G	0.17	-1675.78	404.85	3.48×10^{-5}
	9	rs10118205	*ADAMTSL1*	intron	18815134	C	G	0.03	3487.23	847.91	3.91×10^{-5}
	2	rs2114591	*SP110*	intron	230758813	T	C	0.40	1244.07	305.93	4.77×10^{-5}
	2	rs12470086	*LOC728773 | LOC100133235*	intergenic	147640932	T	A	0.22	1462.10	360.72	5.05×10^{-5}

Fat mass	1	rs6664268	*LOC400796 | SEC16B*	intergenic	176031329	C	T	0.22	-1795.48	366.27	9.48×10^{-7}
Model 2	1	rs4075235	*LOC400796 | SEC16B*	intergenic	176034945	T	C	0.26	-1649.56	345.76	1.84×10^{-6}
	1	rs4072161	*LOC400796 | SEC16B*	intergenic	176036310	G	T	0.48	1454.81	312.25	3.18×10^{-6}
	9	rs10781091	*GDA | ZFAND5*	intergenic	74064112	A	G	0.08	2408.25	539.52	8.06×10^{-6}
	3	rs157538	*EDEM1 | GRM7*	intergenic	6312837	T	G	0.14	1965.69	444.33	9.69×10^{-6}
	1	rs6425446	*LOC400796 | SEC16B*	intergenic	176031200	A	G	0.42	1360.32	308.08	1.01×10^{-5}
	21	rs411697	*MULTIPLE_GENES:571.54073*	complex	29636469	A	G	0.48	-1274.29	301.09	2.31×10^{-5}
	9	rs10118205	*ADAMTSL1*	intron	18815134	C	G	0.03	3562.01	843.64	2.42×10^{-5}
	1	rs11581129	*LOC400796 | SEC16B*	intergenic	176037754	A	G	0.17	-1690.15	402.53	2.68×10^{-5}
	2	rs12470086	*LOC728773 | LOC100133235*	intergenic	147640932	T	A	0.22	1471.27	359.01	4.16×10^{-5}
	1	rs4311843	*LOC400796 | SEC16B*	intergenic	176035387	C	A	0.33	1315.94	326.26	5.50×10^{-5}
Fat mass	1	rs6425453	*LOC400796 | SEC16B*	intergenic	176085789	G	A	0.48	1004.70	230.50	1.31×10^{-5}
Model 3	1	rs4072161	*LOC400796 | SEC16B*	intergenic	176036310	G	T	0.48	1006.85	231.45	1.36×10^{-5}
	20	rs6047259	*C20orf74 | C20orf19*	intergenic	21046319	T	C	0.46	924.98	216.77	1.98×10^{-5}
	2	rs2339853	*KLHL29 | ATAD2B*	intergenic	23811668	T	C	0.50	892.60	211.87	2.52×10^{-5}
	1	rs4075235	*LOC400796 | SEC16B*	intergenic	176034945	T	C	0.26	-1070.22	254.86	2.68×10^{-5}
	3	rs6802030	*LOC730168 | TBL1R1*	intergenic	177944881	C	T	0.49	937.03	223.69	2.80×10^{-5}
	1	rs1854288	*LOC400796 | SEC16B*	intergenic	176069710	A	G	0.47	965.52	231.64	3.07×10^{-5}
	1	rs12092449	*LOC400796 | SEC16B*	intergenic	176081551	C	T	0.47	964.83	231.61	3.10×10^{-5}
	1	rs2068973	*LOC400796 | SEC16B*	intergenic	176059338	A	G	0.43	954.84	230.67	3.48×10^{-5}
	1	rs6664268	*LOC400796 | SEC16B*	intergenic	176031329	C	T	0.22	-1110.31	269.81	3.87×10^{-5}
	11	rs7127524	*LOC727869 | MMP7*	intergenic	101858402	A	G	0.29	-1011.08	245.98	3.95×10^{-5}
Lean mass	9	rs10781091	*GDA | ZFAND5*	intergenic	74064112	A	G	0.08	1633.50	350.03	3.06×10^{-6}
Model 1	5	rs247544	*STARD4 | C5orf13*	intergenic	110923874	A	G	0.10	-1480.44	338.76	1.24×10^{-5}
	2	rs2114591	*SP110*	intron	230758813	T	C	0.40	836.43	198.42	2.49×10^{-5}
	14	rs10144702	*NRN3 | LOC100131580*	intergenic	78582869	G	A	0.16	1102 .45	266.55	3.53×10^{-5}
	17	rs7207980	*ABR*	intron	971144	C	A	0.22	-966.38	237.36	4.67×10^{-5}
	19	rs746075	*NUCB1*	intron	54108748	A	G	0.09	-1392.20	344.68	5.37×10^{-5}
	3	rs6441886	*CDCP1*	intron	45103101	T	C	0.50	805.94	202.46	6.87×10^{-5}
	7	rs6967593	*IFRD1*	intron	111884028	C	T	0.38	787.11	202.17	9.89×10^{-5}
Lean mass	9	rs10781091	*GDA | ZFAND5*	intergenic	74064112	A	G	0.08	1374.36	309.73	9.11×10^{-6}
Model 2	2	rs2114591	*SP110*	intron	230758813	T	C	0.40	774.48	175.24	9.89×10^{-6}
	17	rs7207980	*ABR*	intron	971144	C	A	0.22	-876.17	210.01	3.02×10^{-5}
	9	rs56372464	*GBGT1 | OBP2B*	intergenic	135051234	G	A	0.14	1051.44	256.11	4.04×10^{-5}
	6	rs4248166	*BTNL2*	intron	32474399	C	T	0.14	-1089.19	267.18	4.57×10^{-5}

	6	rs12529049	*C6orf10 | BTNL2*	intergenic	32465693	T	C	0.14	-1083.50	268.06	5.30×10^{-5}
	6	rs4895391	*TCF21 | TBPL1*	intergenic	134280295	A	C	0.45	694.03	174.46	6.95×10^{-5}
Lean mass	22	rs1534891	*CSNK1E*	intron	37025045	T	C	0.04	1486.67	333.52	8.29×10^{-6}
Model 4	20	rs10485674	*DH35 | MAFB*	intergenic	38727246	G	A	0.01	2348.38	548.24	1.84×10^{-5}
	9	rs1926367	*JMJD2C | LOC158345*	intergenic	7353762	T	G	0.14	731.83	177.44	3.72×10^{-5}
	2	rs75276762	*KIAA1486 | IRS1*	intergenic	226755929	T	C	0.26	-665.56	162.88	4.38×10^{-5}
	5	rs10040989	*CDC23 | GFRA3*	intergenic	137601624	A	G	0.10	-827.54	204.78	5.32×10^{-5}
	3	rs11928440	*MUC4*	intron	196986392	G	A	0.34	526.68	130.73	5.61×10^{-5}
	2	rs1550109	*SRBD1*	intron	45633112	G	T	0.07	-955.50	237.78	5.86×10^{-5}
	5	rs17409588	*RNASEN*	intron	31564384	C	T	0.12	-757.11	191.79	7.90×10^{-5}
	17	rs7207980	*ABR*	intron	971144	C	A	0.22	-596.27	151.43	8.23×10^{-5}
PFM	1	rs6664268	*LOC400796 | SEC16B*	intergenic	176031329	C	T	0.22	-1.33	0.27	1.18×10^{-6}
Model 1	1	rs4075235	*LOC400796 | SEC16B*	intergenic	176034945	T	C	0.26	-1.26	0.26	1.24×10^{-6}
	1	rs4072161	*LOC400796 | SEC16B*	intergenic	176036310	G	T	0.48	1.10	0.23	2.55×10^{-6}
	1	rs2068973	*LOC400796 | SEC16B*	intergenic	176059338	A	G	0.43	1.01	0.23	1.35×10^{-5}
	1	rs11581129	*LOC400796 | SEC16B*	intergenic	176037754	A	G	0.17	-1.29	0.30	1.94×10^{-5}
	2	rs2339853	*KLHL29 | ATAD2B*	intergenic	23811668	T	C	0.50	0.92	0.22	2.39×10^{-5}
	3	rs157538	*EDEM1 | GRM7*	intergenic	6312837	T	G	0.14	1.39	0.33	2.66×10^{-5}
	20	rs8122057	*JAG1 | LOC728573*	intergenic	11085719	G	T	0.23	1.12	0.27	2.80×10^{-5}
	17	rs115012414	*APOH | PRKCA*	intergenic	61724373	C	T	0.04	-2.45	0.59	3.20×10^{-5}
	1	rs6425453	*LOC400796 | SEC16B*	intergenic	176085789	G	A	0.48	0.95	0.23	4.25×10^{-5}
	6	rs2820232	*ANKS1A*	intron	35111581	A	C	0.01	-5.20	1.27	4.27×10^{-5}
PFM	1	rs6664268	*LOC400796 | SEC16B*	intergenic	176031329	C	T	0.22	-1.35	0.27	7.90×10^{-7}
Model 2	1	rs4075235	*LOC400796 | SEC16B*	intergenic	176034945	T	C	0.26	-1.27	0.26	9.05×10^{-7}
	1	rs4072161	*LOC400796 | SEC16B*	intergenic	176036310	G	T	0.48	1.10	0.23	2.50×10^{-6}
	1	rs12068973	*LOC400796 | SEC16B*	intergenic	176059338	A	G	0.43	1.01	0.23	1.35×10^{-5}
	20	rs8122057	*JAG1 | LOC728573*	intergenic	11085719	G	T	0.23	1.14	0.27	1.93×10^{-5}
	1	rs11581129	*LOC400796 | SEC16B*	intergenic	176037754	A	G	0.17	-1.28	0.30	2.00×10^{-5}
	3	rs157538	*EDEM1 | GRM7*	intergenic	6312837	T	G	0.14	1.40	0.33	2.04×10^{-5}
	2	rs2339853	*KLHL29 | ATAD2B*	intergenic	23811668	T	C	0.50	0.91	0.22	2.60×10^{-5}
	1	rs6425453	*LOC400796 | SEC16B*	intergenic	176085789	G	A	0.48	0.96	0.23	3.31×10^{-5}
	1	rs6425446	*LOC400796 | SEC16B*	intergenic	176031200	A	G	0.42	0.95	0.23	3.67×10^{-5}

Table E10. Top SNPs associated with body composition traits when female datasets are combined (caregivers and young adult females)

TRAIT	CHR	SNP	GENE SYMBOL	GENE LOCATION	BP	A1	A2	MAF	EFFECT (BETA)	SE	P
BMI	17	rs115012414	*APOH | PRKCA*	intergenic	61724373	C	T	0.03	-0.04	0.01	1.68×10^{-5}
Model 1	17	rs77612309	*APOH | PRKCA*	intergenic	61724990	C	T	0.03	-0.04	0.01	2.85×10^{-5}
	19	rs661821	*ZNF358*	intron	7488649	T	C	0.39	-0.01	0.00	3.22×10^{-5}
	8	rs4246126	*PTK2*	intron	141860074	C	A	0.21	0.02	0.00	4.26×10^{-5}
	15	rs4775027	*AQP9 | LIPC*	intergenic	56375396	G	A	0.37	-0.01	0.00	7.63×10^{-5}
	11	rs7106594	*INSC | SO6*	intergenic	15477573	C	T	0.30	0.01	0.00	8.10×10^{-5}
	9	rs10781091	*GDA | ZFAND5*	intergenic	74064112	A	G	0.08	0.02	0.01	8.86×10^{-5}
	2	rs2114591	*SP110*	intron	230758813	T	C	0.40	0.01	0.00	9.06×10^{-5}
WC	2	rs2114591	*SP110*	intron	230758813	T	C	0.40	2.16	0.48	5.33×10^{-6}
Model 1	9	rs7021554	*ANRIL | LOC729983*	intergenic	22142884	T	C	0.09	3.69	0.86	1.85×10^{-5}
	9	rs72656779	*LOC729983*	utr	22145640	A	G	0.07	4.03	0.95	2.05×10^{-5}
	1	rs4625336	*PADI3 | PADI4*	intergenic	17503551	A	C	0.13	2.93	0.70	2.57×10^{-5}
	14	rs10146149	*NRN3 | LOC100131580*	intergenic	78578318	T	C	0.11	-3.15	0.76	3.08×10^{-5}
	14	rs11622292	*PRO1768 | FON3*	intergenic	89136080	T	C	0.14	2.73	0.66	3.73×10^{-5}
	3	rs12493688	*HES1 | LOC100131551*	intergenic	195401542	G	T	0.01	10.18	2.47	3.85×10^{-5}
	3	rs117195838	*BTD*	intron	15625332	T	C	0.03	6.33	1.54	4.05×10^{-5}
	18	rs4550540	*ZNF521*	intron	21145629	A	G	0.49	-1.94	0.48	4.68×10^{-5}
	7	rs33951980	*MLIPL*	intron	72667373	T	C	0.03	5.48	1.38	6.80×10^{-5}
WC	2	rs2114591	*SP110*	intron	230758813	T	C	0.40	2.13	0.47	6.40×10^{-6}
Model 2	3	rs12493688	*HES1 | LOC100131551*	intergenic	195401542	G	T	0.01	10.74	2.46	1.26×10^{-5}
	5	rs6865951	*SH3PD2B | LOC100130394*	intergenic	171815909	T	C	0.07	3.99	0.92	1.55×10^{-5}
	9	rs72656779	*LOC729983*	utr	22145640	A	G	0.07	3.96	0.94	2.53×10^{-5}
	9	rs7021554	*ANRIL | LOC729983*	intergenic	22142884	T	C	0.09	3.58	0.86	2.91×10^{-5}
	1	rs4625336	*PADI3 | PADI4*	intergenic	17503551	A	C	0.13	2.85	0.69	3.84×10^{-5}
	3	rs117195838	*BTD*	intron	15625332	T	C	0.03	6.29	1.53	4.06×10^{-5}
	14	rs10146149	*NRN3 | LOC100131580*	intergenic	78578318	T	C	0.11	-3.08	0.75	4.13×10^{-5}
	3	rs72625020	*IGF2BP2 | SFRS10*	intergenic	187079472	T	C	0.21	2.34	0.57	4.54×10^{-5}
	14	rs11622292	*PRO1768 | FON3*	intergenic	89136080	T	C	0.14	2.63	0.66	6.41×10^{-5}
HC	9	rs7021554	*ANRIL | LOC729983*	intergenic	22142884	T	C	0.09	4.05	0.89	5.97×10^{-6}

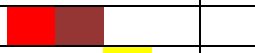
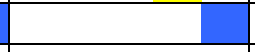
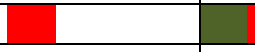
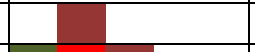
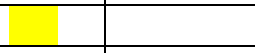
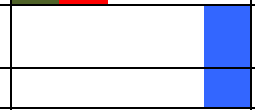
HC	8	rs78933755	*PPP1R3B | LOC100129150*	intergenic	9245811	G	A	0.01	9.07	2.13	2.06×10^{-5}
Model 1	9	rs72656779	*LOC729983*	utr	22145640	A	G	0.07	4.17	0.98	2.16×10^{-5}
	18	rs12326471	*LMAN1 | CCBE1*	intergenic	55191615	C	T	0.26	-2.26	0.55	3.53×10^{-5}
	18	rs385769	*PTPRM*	intron	7790480	G	T	0.11	3.12	0.77	5.34×10^{-5}
	8	rs35584813	*PPP1R3B | LOC100129150*	intergenic	9245040	C	T	0.01	8.37	2.08	5.70×10^{-5}
	11	rs58174260	*BDNFOS*	intron	27547828	T	G	0.04	4.93	1.24	6.61×10^{-5}
	8	rs73535332	*PPP1R3B | LOC100129150*	intergenic	9255047	C	G	0.01	8.88	2.23	6.97×10^{-5}
	21	rs411697	*MULTIPLE_GENES:571.5407 3*	complex	29636469	A	G	0.48	-1.92	0.48	7.58×10^{-5}
	14	rs10146149	*NRN3 | LOC100131580*	intergenic	78578318	T	C	0.11	-3.09	0.78	8.15×10^{-5}
	8	rs11778774	*PPP1R3B | LOC100129150*	intergenic	9251517	G	A	0.01	8.06	2.05	8.55×10^{-5}
HC	9	rs7021554	*ANRIL | LOC729983*	intergenic	22142884	T	C	0.09	3.89	0.88	1.07×10^{-5}
Model 2	18	rs12326471	*LMAN1 | CCBE1*	intergenic	55191615	C	T	0.26	-2.31	0.54	1.94×10^{-5}
	8	rs78933755	*PPP1R3B | LOC100129150*	intergenic	9245811	G	A	0.01	8.84	2.10	2.65×10^{-5}
	9	rs72656779	*LOC729983*	utr	22145640	A	G	0.07	4.07	0.97	2.70×10^{-5}
	2	rs2190373	*RND3 | LOC344332*	intergenic	151482339	A	G	0.13	-3.00	0.73	3.68×10^{-5}
	11	rs58174260	*BDNFOS*	intron	27547828	T	G	0.04	4.92	1.22	5.61×10^{-5}
	20	rs115667632	*C20orf174*	intron	57210215	T	C	0.03	-5.68	1.43	7.01×10^{-5}
	11	rs12574325	*BDNFOS*	intron	27569624	A	G	0.06	4.28	1.08	7.10×10^{-5}
	11	rs16917135	*BDNFOS*	intron	27571281	T	C	0.06	4.28	1.08	7.10×10^{-5}
	6	rs2474346	*GUCA1A*	intron	42244235	T	G	0.30	1.98	0.50	7.87×10^{-5}
	21	rs411697	*MULTIPLE_GENES:571.5407 3*	complex	29636469	A	G	0.48	-1.89	0.48	8.01×10^{-5}
WHR	12	rs1625560	*TMTC2 | SLC6A15*	intergenic	83719434	G	T	0.07	-0.03	0.01	5.16×10^{-6}
Model 1	7	rs2289055	*DD56*	intron	44578415	G	A	0.04	0.03	0.01	7.11×10^{-6}
	7	rs7723	*DD56 | TMED4*	intergenic	44584318	A	G	0.04	0.03	0.01	7.11×10^{-6}
	15	rs11074155	*UNQ9370 | LOC728292*	intergenic	91627775	A	G	0.04	0.03	0.01	1.40×10^{-5}
	16	rs2037912	*PPL*	coding nonsyn c|s	4873940	G	C	0.02	0.05	0.01	2.19×10^{-5}
	16	rs1861554	*FTO*	intron	52607268	G	A	0.07	0.02	0.01	2.75×10^{-5}
	5	rs17307165	*NDUFS4 | ARL15*	intergenic	53113382	G	A	0.09	0.02	0.01	3.56×10^{-5}
	15	rs12438098	*LOC728292 | MCTP2*	intergenic	92188493	C	G	0.13	0.02	0.00	4.79×10^{-5}
	16	rs1861358	*FTO*	intron	52602704	A	C	0.06	0.03	0.01	4.83×10^{-5}
	16	rs2111116	*FTO*	intron	52606753	A	G	0.06	0.03	0.01	4.83×10^{-5}
	6	rs9271100	*HLA-DRB1 | HLA-DQA1*	intergenic	32684456	T	C	0.16	0.02	0.00	5.53×10^{-5}
	6	rs9271209	*HLA-DRB1 | HLA-DQA1*	intergenic	32687013	G	A	0.16	0.02	0.00	5.53×10^{-5}
	9	rs10781091	*GDA | ZFAND5*	intergenic	74064112	A	G	0.08	3172.33	702.51	6.31×10^{-6}

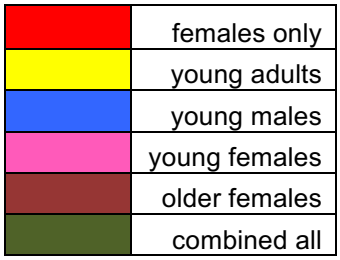
Fat mass	1	rs4075235	*LOC400796 | SEC16B*	intergenic	176034945	T	C	0.26	-1997.59	445.14	7.20×10^{-6}
Model 1	1	rs6664268	*LOC400796 | SEC16B*	intergenic	176031329	C	T	0.22	-2085.36	466.88	7.95×10^{-6}
	3	rs157538	*EDEM1 | GRM7*	intergenic	6312837	T	G	0.14	2475.25	568.68	1.35×10^{-5}
	1	rs4072161	*LOC400796 | SEC16B*	intergenic	176036310	G	T	0.47	1739.79	406.15	1.84×10^{-5}
	1	rs4311843	*LOC400796 | SEC16B*	intergenic	176035387	C	A	0.32	1743.95	425.85	4.22×10^{-5}
	2	rs2114591	*SP110*	intron	230758813	T	C	0.40	1555.03	391.41	7.10×10^{-5}
	2	rs6541885	*LOC728241 | CNTNAP5*	intergenic	123668890	A	G	0.03	-4274.74	1081.24	7.70×10^{-5}
	14	rs10141212	*MDGA2 | RPS29*	intergenic	48007525	C	T	0.13	-2273.95	578.81	8.54×10^{-5}
Fat mass	1	rs6425446	*LOC400796 | SEC16B*	intergenic	176031200	A	G	0.41	1809.22	398.26	5.55×10^{-6}
Model 2	1	rs4075235	*LOC400796 | SEC16B*	intergenic	176034945	T	C	0.26	-1982.54	442.82	7.57×10^{-6}
	1	rs6664268	*LOC400796 | SEC16B*	intergenic	176031329	C	T	0.22	-2055.32	464.45	9.63×10^{-6}
	9	rs10781091	*GDA | ZFAND5*	intergenic	74064112	A	G	0.08	3025.39	699.38	1.52×10^{-5}
	3	rs157538	*EDEM1 | GRM7*	intergenic	6312837	T	G	0.14	2409.64	566.10	2.08×10^{-5}
	1	rs4072161	*LOC400796 | SEC16B*	intergenic	176036310	G	T	0.47	1710.93	403.88	2.27×10^{-5}
	1	rs4311843	*LOC400796 | SEC16B*	intergenic	176035387	C	A	0.32	1728.45	423.45	4.47×10^{-5}
	14	rs10141212	*MDGA2 | RPS29*	intergenic	48007525	C	T	0.13	-2316.47	576.24	5.82×10^{-5}
	2	rs6541885	*LOC728241 | CNTNAP5*	intergenic	123668890	A	G	0.03	-4266.37	1076.42	7.39×10^{-5}
	1	rs2404717	*PPAP2B*	intron	56751927	G	T	0.15	-2156.07	546.69	8.02×10^{-5}
	2	rs2114591	*SP110*	intron	230758813	T	C	0.40	1533.83	389.66	8.27×10^{-5}
Fat mass	7	rs4729039	*ANKIB1 | GATAD1*	intergenic	91880904	C	A	0.10	2027.36	423.69	1.71×10^{-6}
Model 3	3	rs6802030	*LOC730168 | TBL1R1*	intergenic	177944881	C	T	0.49	1139.91	261.40	1.30×10^{-5}
	6	rs7745594	*FARS2 | NRN1*	intergenic	5843149	T	C	0.19	1365.38	329.71	3.46×10^{-5}
	3	rs7610881	*ROBO2*	intron	77619036	G	A	0.03	-3206.52	789.34	4.86×10^{-5}
	11	rs115124293	*FCHSD2*	intron	72354213	C	A	0.07	1984.73	493.91	5.86×10^{-5}
	11	rs114705375	*FCHSD2*	intron	72383069	T	A	0.07	1984.73	493.91	5.86×10^{-5}
	11	rs80059745	*CHST1 | DKFZp779M0652*	intergenic	45679635	C	G	0.08	-1835.02	456.75	5.88×10^{-5}
	11	rs75409252	*FCHSD2*	intron	72246468	T	C	0.07	1983.39	496.43	6.46×10^{-5}
	1	rs6425446	*LOC400796 | SEC16B*	intergenic	176031200	A	G	0.41	1057.83	264.95	6.54×10^{-5}
	6	rs11243038	*FARS2 | NRN1*	intergenic	5842767	C	T	0.19	1318.07	331.36	6.96×10^{-5}
	13	rs6563563	*POSTN | TRPC4*	intergenic	37093430	T	G	0.07	1948.73	493.53	7.86×10^{-5}
Lean mass	9	rs10781091	*GDA | ZFAND5*	intergenic	74064112	A	G	0.08	2240.17	407.69	3.91×10^{-8}
Model 1	7	rs17443228	*LRRN3 | IMMP2L*	intergenic	110774133	A	G	0.08	-1767.14	420.78	2.67×10^{-5}
	2	rs2114591	*SP110*	intron	230758813	T	C	0.40	899.49	227.32	7.59×10^{-5}
	20	rs6035771	*C20orf74 | C20orf19*	intergenic	20981250	A	G	0.29	-998.63	254.51	8.72×10^{-5}
Lean mass	9	rs10781091	*GDA | ZFAND5*	intergenic	74064112	A	G	0.08	1868.52	369.67	4.31×10^{-7}

Lean mass	5	rs185105	*ANKRD34B | DHFR*	intergenic	79906483	T	C	0.04	2184.80	510.10	1.84×10^{-5}
Model 2	8	rs73535324	*PPP1R3B | LOC100129150*	intergenic	9252643	A	C	0.01	3790.77	919.74	3.76×10^{-5}
	8	rs73535332	*PPP1R3B | LOC100129150*	intergenic	9255047	C	G	0.01	3808.12	934.66	4.61×10^{-5}
	2	rs2114591	*SP110*	intron	230758813	T	C	0.40	832.99	206.02	5.27×10^{-5}
	8	rs78933755	*PPP1R3B | LOC100129150*	intergenic	9245811	G	A	0.01	3641.13	900.75	5.29×10^{-5}
	8	rs11778774	*PPP1R3B | LOC100129150*	intergenic	9251517	G	A	0.01	3451.69	864.75	6.56×10^{-5}
	16	rs388997	*CDH8 | LOC390735*	intergenic	61204594	C	T	0.42	-814.29	206.06	7.76×10^{-5}
	8	rs35584813	*PPP1R3B | LOC100129150*	intergenic	9245040	C	T	0.01	3459.27	877.38	8.06×10^{-5}
	8	rs36090863	*PPP1R3B | LOC100129150*	intergenic	9246378	T	C	0.01	3459.27	877.38	8.06×10^{-5}
	10	rs1472135	*LOC727878 | PFKP*	intergenic	2552642	C	T	0.18	-1011.66	257.58	8.58×10^{-5}
Lean mass	15	rs17270501	*RORA*	intron	58907943	T	C	0.03	2003.19	413.04	1.24×10^{-6}
Model 4	11	chr11:45676254	*CHST1 | DKFZp779M0652*	intergenic	45676254	A	T	0.40	-639.25	148.56	1.68×10^{-5}
	11	chr11:45675448	*CHST1 | DKFZp779M0652*	intergenic	45675448	C	T	0.38	-599.31	147.80	5.01×10^{-5}
	3	rs2606738	*ATG7*	intron	11371562	G	A	0.37	557.31	137.86	5.28×10^{-5}
	22	rs1534891	*CSNK1E*	intron	37025045	T	C	0.04	1449.79	360.78	5.86×10^{-5}
	3	rs7610881	*ROBO2*	intron	77619036	G	A	0.03	1650.33	416.09	7.30×10^{-5}
	8	rs2978056	*LOC100129104 | ZFAT1*	intergenic	134693776	T	C	0.31	-575.91	146.00	8.00×10^{-5}
	7	rs4729039	*ANKIB1 | GATAD1*	intergenic	91880904	C	A	0.10	-875.99	223.74	9.03×10^{-5}
PFM	1	rs4075235	*LOC400796 | SEC16B*	intergenic	176034945	T	C	0.26	-1.37	0.31	1.18×10^{-5}
Model 1	1	rs6425446	*LOC400796 | SEC16B*	intergenic	176031200	A	G	0.41	1.22	0.28	1.39×10^{-5}
	1	rs2068973	*LOC400796 | SEC16B*	intergenic	176059338	A	G	0.42	1.21	0.28	1.54×10^{-5}
	14	rs11846244	*TRDV2 | TRDD1*	intergenic	21965715	A	G	0.24	1.35	0.31	1.62×10^{-5}
	20	rs8122057	*JAG1 | LOC728573*	intergenic	11085719	G	T	0.23	1.37	0.32	2.14×10^{-5}
	15	rs74024597	*ADAMTSL3*	intron	82353771	G	A	0.11	1.87	0.44	2.49×10^{-5}
	1	rs10913437	*LOC400796 | SEC16B*	intergenic	176062763	T	C	0.43	1.19	0.28	2.64×10^{-5}
	14	rs10141212	*MDGA2 | RPS29*	intergenic	48007525	C	T	0.13	-1.68	0.41	3.41×10^{-5}
	1	rs6664268	*LOC400796 | SEC16B*	intergenic	176031329	C	T	0.22	-1.36	0.33	3.51×10^{-5}
	3	rs157538	*EDEM1 | GRM7*	intergenic	6312837	T	G	0.14	1.64	0.40	4.12×10^{-5}
	6	rs2820232	*ANKS1A*	intron	35111581	A	C	0.01	-5.94	1.45	4.40×10^{-5}
PFM	1	rs4075235	*LOC400796 | SEC16B*	intergenic	176034945	T	C	0.26	-1.38	0.31	8.89×10^{-6}
Model 2	1	rs6425446	*LOC400796 | SEC16B*	intergenic	176031200	A	G	0.41	1.24	0.28	9.53×10^{-6}
	20	rs8122057	*JAG1 | LOC728573*	intergenic	11085719	G	T	0.23	1.41	0.32	1.18×10^{-5}
	1	rs2068973	*LOC400796 | SEC16B*	intergenic	176059338	A	G	0.42	1.21	0.28	1.49×10^{-5}
	3	rs157538	*EDEM1 | GRM7*	intergenic	6312837	T	G	0.14	1.69	0.40	2.05×10^{-5}
	1	rs6664268	*LOC400796 | SEC16B*	intergenic	176031329	C	T	0.22	-1.38	0.33	2.31×10^{-5}
	1	rs10913437	*LOC400796 | SEC16B*	intergenic	176062763	T	C	0.43	1.18	0.28	2.77×10^{-5}

	15	rs74024597	*ADAMTSL3*	intron	82353771	G	A	0.11	1.84	0.44	3.13×10^{-5}
	1	rs4072161	*LOC400796 | SEC16B*	intergenic	176036310	G	T	0.47	1.18	0.28	3.23×10^{-5}
	14	rs11846244	*TRDV2 | TRDD1*	intergenic	21965715	A	G	0.24	1.27	0.31	4.38×10^{-5}
	1	rs6425453	*LOC400796 | SEC16B*	intergenic	176085789	G	A	0.47	1.14	0.28	4.44×10^{-5}

Table E11. Sub-groups (represented in various colours in the key) used in the analysis together with the different phenotypes such that age- and sex- specific associations are illustrated. *-refer to loci that were associated with a single phenotype

Loci	Phenotypes						
	BMI	***WC***	***HC***	***WHR***	***FM***	***LM***	***PFM***
APOH | PRKCA							
LOC728241 | CNTNAP5							
ZFYVE9							
TRPM7							
SP110							
NRXN3 | LOC100131580							
LPAL2							
PPP1R3B | LOC100129150							
WARS2							
*BDNFOS**							
LOC400796 | SEC16B							
*FTO**							
*LOC100129474 | SLC17A4**							
*COBLL1**							
LOC727944 | TMEM18							
NEGR1 | LOC100132353							
NBEAL1							



APPENDIX F: Scripts for calculating narrow-sense heritability estimates using Metabochip data.

Script 1: *find_duplicates.R*

```
source("../scripts/functions.R")

g <- read_grm("../data/clean")
dup_values <- g$grm[g$grm$id1 != g$grm$id2 & g$grm$grm > 0.8, ]
dup_ids <- g$id[unlist(dup_values[,1:2]),]
write.table(dup_ids, "../data/duplicate_genos.txt", row=F, col=F,
qu=F)
```

Script 2*: remove_unrelateds.R*

```
' Read binary GRM files into R
   #'
   #' @param rootname
   #' @export
   #' @return List of GRM and id data frames
   readGRM <- function(rootname)
   {
   bin.file.name <- paste(rootname, ".grm.bin", sep="")
   n.file.name <- paste(rootname, ".grm.N.bin", sep="")
   id.file.name <- paste(rootname, ".grm.id", sep="")

   cat("Reading IDs\n")
   id <- read.table(id.file.name)
   n <- dim(id)[1]
   cat("Reading GRM\n")
   bin.file <- file(bin.file.name, "rb")
   grm <- readBin(bin.file, n=n*(n+1)/2, what=numeric(0), size=4)
```

```
close(bin.file)
cat("Reading N\n")
n.file <- file(n.file.name, "rb")
N <- readBin(n.file, n=n*(n+1)/2, what=numeric(0), size=4)
close(n.file)

cat("Creating data frame\n")
l <- list()
for(i in 1:n)
{
   l[[i]] <- 1:i
}
col1 <- rep(1:n, 1:n)
col2 <- unlist(l)
grm <- data.frame(id1=col1, id2=col2, N=N, grm=grm)

ret <- list()
ret$grm <- grm
ret$id <- id
return(ret)
}

#' Write readGRM style output back to binary GRM for use with
GCTA
#'
#' @param grm Output from \link{readGRM}
#' @param rootname
#' @export
writeGRM <- function(grm, rootname)
{
bin.file.name <- paste(rootname, ".grm.bin", sep="")
n.file.name <- paste(rootname, ".grm.N.bin", sep="")
id.file.name <- paste(rootname, ".grm.id", sep="")
write.table(grm$id, id.file.name, row=F, col=F, qu=F)
n <- dim(grm$id)[1]
bin.file <- file(bin.file.name, "wb")
writeBin(grm$grm$grm, bin.file, size=4)
close(bin.file)
n.file <- file(n.file.name, "wb")
writeBin(grm$grm$N, n.file, size=4)
```

```
close(n.file)
}


setUnrelsZero <- function(grm, threshold)
{
   index <- grm$grm$grm < threshold
   grm$grm$grm[index] <- 0
   return(grm)
}


arguments <- commandArgs(T)

infile <- arguments[1]
outfile <- arguments[2]
threshold <- as.numeric(arguments[3])

grm <- readGRM(infile)
grm <- setUnrelsZero(grm, threshold)
writeGRM(grm, outfile)
```

Script 3: *geno_data.sh*

```
#!/bin/bash

#PBS -N Heritability
#PBS -q WitsLong
#PBS -l walltime=20:00:00,mem=3GB
#PBS -l nodes=1:ppn=1
set -e

cd /home/venesa/archive/data

# Get summary statistics
plink1.90      --bfile      ../data/prunedbmerge      --freq      --out
../data/prunedbmerge
plink1.90      --bfile      ../data/prunedbmerge      --hardy      --out
../data/prunedbmerge
```

```
plink1.90 --bfile ../data/prunedbmerge --missing --out
../data/prunedbmerge
gzip ../data/prunedbmerge.hwe ../data/prunedbmerge.frq
../data/prunedbmerge.imiss ../data/prunedbmerge.lmiss
# Clean data
plink1.90 --bfile ../data/prunedbmerge --maf 0.01 --hwe 1e-6 --out
../data/clean --make-bed

# Generate GRM
plink1.90 --bfile ../data/clean --make-grm-bin --out ../data/clean

# Find duplicates
R --no-save < find_duplicates.R

# Remove possible duplicates
plink1.90 --bfile ../data/clean --remove ../data/duplicate_genos.txt --
make-bed --out ../data/clean2

# Generate GRM again without duplicates
plink1.90 --bfile ../data/clean2 --make-grm-bin --out ../data/clean2

# Principal components
gcta64 --grm ../data/clean2 --pca 10 --out ../data/clean2

# Make GRM of relateds
Rscript remove_unrelateds.R ../data/clean2 ../data/clean_relateds2 0.05

# Make MGRM file
echo -e "../data/clean2" > ../data/mgrm.txt
echo -e "../data/clean_relateds2" >> ../data/mgrm.txt

# Make GRM of unrelateds by removing a random person from each pair of
individuals with relationship > 0.05
gcta64 --grm ../data/clean2 --grm-cutoff 0.05 --out
../data/clean_unrelateds2 --make-grm
```

Script 4: *pheno_data_all.R*

```
## ---- setup ----
```

```
library(GenABEL)
library(plyr)
source("../scripts/functions.R")


## ---- read_data ----

dups <- read.table("../data/duplicate_genos.txt",
stringsAsFactors=FALSE)
phen <- read.table("../data/pheno.txt", he=T,
stringsAsFactors=FALSE)
phen <- subset(phen, ! IID %in% dups$V2)

## ---- write_height ----

write.table(subset(phen, select=c(FID, IID, HEIGHT_1)),
file="../data/hieght.phen", row=F, col=F, qu=F)


## ---- write_weight ----

write.table(subset(phen, select=c(FID, IID, WEIGHT)),
file="../data/weight.phen", row=F, col=F, qu=F)


## ---- rank_transform_bmi ----

phen$rnbmi[phen$BMI != -9] <- rntransform(phen$BMI[phen$BMI !=
-9])


## ---- write_bmi ----

write.table(subset(phen, select=c(FID, IID, BMI, LOGBMI,
rnbmi)), file="../data/bmi.phen", row=F, col=F, qu=F)


## ---- write_wc ----
```

```
write.table(subset(phen, select=c(FID, IID, WC)),
file="../data/wc.phen", row=F, col=F, qu=F)


## ---- write_HC ----

write.table(subset(phen, select=c(FID, IID, HC)),
file="../data/hc.phen", row=F, col=F, qu=F)


## ---- write_whr ----

write.table(subset(phen, select=c(FID, IID, WHR)),
file="../data/whr.phen", row=F, col=F, qu=F)


## ---- write_fatmass ----

write.table(subset(phen, select=c(FID, IID, FATMASS)),
file="../data/fatmass.phen", row=F, col=F, qu=F)

## ---- write_leanmass ----

write.table(subset(phen, select=c(FID, IID, LEANMASS)),
file="../data/leanmss.phen", row=F, col=F, qu=F)


## ---- write_pbf ----

write.table(subset(phen, select=c(FID, IID, PBF)),
file="../data/pbf.phen", row=F, col=F, qu=F)



## ---- covariates ----

pcs <- read.table("../data/clean2.eigenvec",
stringsAsFactors=FALSE)
stopifnot(all(pcs$V1 == phen$FID))
pcs$sex <- phen$SEX
```

```
pcs$age <- phen$AGE


## ---- write_covs ----

write.table(pcs, file="../data/covariates.txt", row=F, col=F,
qu=F)
```

Script 5: *run_allpheno_analysis.sh*

```
#!/bin/bash

#PBS -N Heritability
#PBS -q WitsLong
#PBS -l walltime=20:00:00,mem=3G
#PBS -l nodes=1:ppn=1


set -e

cd /home/venesa/archive/data


# Run GCTA BMI zaitlen
gcta64 \
        --mgrm ../data/mgrm.txt \
        --reml --reml-no-lrt \
        --pheno ../data/bmi.phen \
        --mpheno 1 \
        --qcovar ../data/covariates.txt \
        --out ../results2/bmi_zaitlen

gcta64 \
        --mgrm ../data/mgrm.txt \
        --reml --reml-no-lrt \
        --pheno ../data/bmi.phen \
        --mpheno 2 \
        --qcovar ../data/covariates.txt \
        --out ../results2/logbmi_zaitlen
```

```
gcta64 \
        --mgrm ../data/mgrm.txt \
        --reml --reml-no-lrt \
        --pheno ../data/bmi.phen \
        --mpheno 3 \
        --qcovar ../data/covariates.txt \
        --out ../results2/rnbmi_zaitlen

# Run GCTA BMI all
gcta64 \
        --grm ../data/clean2 \
        --reml --reml-no-lrt \
        --pheno ../data/bmi.phen \
        --mpheno 1 \
        --qcovar ../data/covariates.txt \
        --out ../results2/bmi_all

gcta64 \
        --grm ../data/clean2 \
        --reml --reml-no-lrt \
        --pheno ../data/bmi.phen \
        --mpheno 2 \
        --qcovar ../data/covariates.txt \
        --out ../results2/logbmi_all

gcta64 \
        --grm ../data/clean2 \
        --reml --reml-no-lrt \
        --pheno ../data/bmi.phen \
        --mpheno 3 \
        --qcovar ../data/covariates.txt \
        --out ../results2/rnbmi_all

# Run GCTA BMI relateds
gcta64 \
        --grm ../data/clean_relateds2 \
        --reml --reml-no-lrt \
        --pheno ../data/bmi.phen \
        --mpheno 1 \
```

```
        --qcovar ../data/covariates.txt \
        --out ../results2/bmi_relateds

gcta64 \
        --grm ../data/clean_relateds2 \
        --reml --reml-no-lrt \
        --pheno ../data/bmi.phen \
        --mpheno 2 \
        --qcovar ../data/covariates.txt \
        --out ../results2/logbmi_relateds

gcta64 \
        --grm ../data/clean_relateds2 \
        --reml --reml-no-lrt \
        --pheno ../data/bmi.phen \
        --mpheno 3 \
        --qcovar ../data/covariates.txt \
        --out ../results2/rnbmi_relateds

# Run GCTA BMI unrelateds
gcta64 \
        --grm ../data/clean_unrelateds2 \
        --reml --reml-no-lrt \
        --pheno ../data/bmi.phen \
        --mpheno 1 \
        --qcovar ../data/covariates.txt \
        --out ../results2/bmi_unrelateds

gcta64 \
        --grm ../data/clean_unrelateds2 \
        --reml --reml-no-lrt \
        --pheno ../data/bmi.phen \
        --mpheno 2 \
        --qcovar ../data/covariates.txt \
        --out ../results2/logbmi_unrelateds

gcta64 \
        --grm ../data/clean_unrelateds2 \
        --reml --reml-no-lrt \
        --pheno ../data/bmi.phen \
```

```
        --mpheno 3 \
        --qcovar ../data/covariates.txt \
        --out ../results2/rnbmi_unrelateds


# Run GCTA WC zaitlen
gcta64 \
        --mgrm ../data/mgrm.txt \
        --reml --reml-no-lrt \
        --pheno ../data/wc.phen \
        --mpheno 1 \
        --qcovar ../data/covariates.txt \
        --out ../results2/wc_zaitlen



# Run GCTA WC all
gcta64 \
        --grm ../data/clean2 \
        --reml --reml-no-lrt \
        --pheno ../data/wc.phen \
        --mpheno 1 \
        --qcovar ../data/covariates.txt \
        --out ../results2/wc_all


# Run GCTA WC relateds
gcta64 \
        --grm ../data/clean_relateds2 \
        --reml --reml-no-lrt \
        --pheno ../data/wc.phen \
        --mpheno 1 \
        --qcovar ../data/covariates.txt \
        --out ../results2/wc_relateds


# Run GCTA WC unrelateds
gcta64 \
        --grm ../data/clean_unrelateds2 \
        --reml --reml-no-lrt \
```

```
        --pheno ../data/wc.phen \
        --mpheno 1 \
        --qcovar ../data/covariates.txt \
        --out ../results2/wc_unrelateds


# Run GCTA hc zaitlen
gcta64 \
   --mgrm ../data/mgrm.txt \
   --reml --reml-no-lrt \
   --pheno ../data/hc.phen \
   --mpheno 1 \
   --qcovar ../data/covariates.txt \
   --out ../results2/hc_zaitlen


# Run GCTA hc all
gcta64 \
   --grm ../data/clean2 \
   --reml --reml-no-lrt \
   --pheno ../data/hc.phen \
   --mpheno 1 \
   --qcovar ../data/covariates.txt \
   --out ../results2/hc_all


# Run GCTA hc relateds
gcta64 \
   --grm ../data/clean_relateds2 \
   --reml --reml-no-lrt \
   --pheno ../data/hc.phen \
   --mpheno 1 \
   --qcovar ../data/covariates.txt \
   --out ../results2/hc_relateds


# Run GCTA hc unrelateds
gcta64 \
   --grm ../data/clean_unrelateds2 \
   --reml --reml-no-lrt \
```

```
   --pheno ../data/hc.phen \
   --mpheno 1 \
   --qcovar ../data/covariates.txt \
   --out ../results2/hc_unrelateds


# Run GCTA whr zaitlen
gcta64 \
        --mgrm ../data/mgrm.txt \
        --reml --reml-no-lrt \
        --pheno ../data/whr.phen \
        --mpheno 1 \
        --qcovar ../data/covariates.txt \
        --pheno ../data/whr.phen \
        --mpheno 1 \
        --qcovar ../data/covariates.txt \
        --out ../results2/whr_zaitlen


# Run GCTA whr all
gcta64 \
        --grm ../data/clean2 \
        --reml --reml-no-lrt \
        --pheno ../data/whr.phen \
        --mpheno 1 \
        --qcovar ../data/covariates.txt \
        --out ../results2/whr_all

# Run GCTA whr relateds
gcta64 \
        --grm ../data/clean_relateds2 \
        --reml --reml-no-lrt \
        --pheno ../data/whr.phen \
        --mpheno 1 \
        --qcovar ../data/covariates.txt \
        --out ../results2/whr_relateds


# Run GCTA whr unrelateds
gcta64 \
```

```
        --grm ../data/clean_unrelateds2 \
        --reml --reml-no-lrt \
        --pheno ../data/whr.phen \
        --mpheno 1 \
        --qcovar ../data/covariates.txt \
        --out ../results2/whr_unrelateds


# Run GCTA fatmass zaitlen
gcta64 \
        --mgrm ../data/mgrm.txt \
        --reml --reml-no-lrt \
        --pheno ../data/fatmass.phen \
        --mpheno 1 \
        --qcovar ../data/covariates.txt \
        --pheno ../data/fatmass.phen \
        --mpheno 1 \
        --qcovar ../data/covariates.txt \
        --out ../results2/fatmass_zaitlen



# Run GCTA fatmass all
gcta64 \
        --grm ../data/clean2 \
        --reml --reml-no-lrt \
        --pheno ../data/fatmass.phen \
        --mpheno 1 \
        --qcovar ../data/covariates.txt \
        --out ../results2/fatmass_all



# Run GCTA fatmass relateds
gcta64 \
        --grm ../data/clean_relateds2 \
        --reml --reml-no-lrt \
        --pheno ../data/fatmass.phen \
        --mpheno 1 \
        --qcovar ../data/covariates.txt \
```

```
        --out ../results2/fatmass_relateds


# Run GCTA fatmass unrelateds
gcta64 \
        --grm ../data/clean_unrelateds2 \
        --reml --reml-no-lrt \
        --pheno ../data/fatmass.phen \
        --mpheno 1 \
        --qcovar ../data/covariates.txt \
        --out ../results2/fatmass_unrelateds


# Run GCTA leanmass zaitlen
gcta64 \
        --mgrm ../data/mgrm.txt \
        --reml --reml-no-lrt \
        --pheno ../data/leanmass.phen \
        --mpheno 1 \
        --qcovar ../data/covariates.txt \
        --pheno ../data/leanmass.phen \
        --mpheno 1 \
        --qcovar ../data/covariates.txt \
        --out ../results2/leanmass_zaitlen



# Run GCTA leanmass all
gcta64 \
        --grm ../data/clean2 \
        --reml --reml-no-lrt \
        --pheno ../data/leanmass.phen \
        --mpheno 1 \
        --qcovar ../data/covariates.txt \
        --out ../results2/leanmass_all

# Run GCTA leanmass relateds
gcta64 \
        --grm ../data/clean_relateds2 \
        --reml --reml-no-lrt \
```

```
        --pheno ../data/leanmass.phen \
        --mpheno 1 \
        --qcovar ../data/covariates.txt \
        --out ../results2/leanmass_relateds


# Run GCTA leanmass unrelateds
gcta64 \
        --grm ../data/clean_unrelateds2 \
        --reml --reml-no-lrt \
        --pheno ../data/leanmass.phen \
        --mpheno 1 \
        --qcovar ../data/covariates.txt \
        --out ../results2/leanmass_unrelateds



# Run GCTA PBF zaitlen
gcta64 \
        --mgrm ../data/mgrm.txt \
        --reml --reml-no-lrt \
        --pheno ../data/pbf.phen \
        --mpheno 1 \
        --qcovar ../data/covariates.txt \
        --pheno ../data/pbf.phen \
        --mpheno 1 \
        --qcovar ../data/covariates.txt \
        --out ../results2/pbf_zaitlen



# Run GCTA pbf all
gcta64 \
        --grm ../data/clean2 \
        --reml --reml-no-lrt \
        --pheno ../data/pbf.phen \
        --mpheno 1 \
        --qcovar ../data/covariates.txt \
        --out ../results2/pbf_all
```

```
# Run GCTA pbf relateds
gcta64 \
        --grm ../data/clean_relateds2 \
        --reml --reml-no-lrt \
        --pheno ../data/pbf.phen \
        --mpheno 1 \
        --qcovar ../data/covariates.txt \
        --out ../results2/pbf_relateds


# Run GCTA pbf unrelateds
gcta64 \
        --grm ../data/clean_unrelateds2 \
        --reml --reml-no-lrt \
        --pheno ../data/pbf.phen \
        --mpheno 1 \
        --qcovar ../data/covariates.txt \
        --out ../results2/pbf_unrelateds


# Run GCTA weight zaitlen
gcta64 \
        --mgrm ../data/mgrm.txt \
        --reml --reml-no-lrt \
        --pheno ../data/weight.phen \
        --mpheno 1 \
        --qcovar ../data/covariates.txt \
        --pheno ../data/weight.phen \
        --mpheno 1 \
        --qcovar ../data/covariates.txt \
        --out ../results2/weight_zaitlen



# Run GCTA weight all
gcta64 \
        --grm ../data/clean2 \
        --reml --reml-no-lrt \
        --pheno ../data/weight.phen \
```

```
        --mpheno 1 \
        --qcovar ../data/covariates.txt \
        --out ../results2/weight_all


# Run GCTA weight relateds
gcta64 \
        --grm ../data/clean_relateds2 \
        --reml --reml-no-lrt \
        --pheno ../data/weight.phen \
        --mpheno 1 \
        --qcovar ../data/covariates.txt \
        --out ../results2/weight_relateds


# Run GCTA weight unrelateds
gcta64 \
        --grm ../data/clean_unrelateds2 \
        --reml --reml-no-lrt \
        --pheno ../data/weight.phen \
        --mpheno 1 \
        --qcovar ../data/covariates.txt \
        --out ../results2/weight_unrelateds

# Run GCTA height zaitlen
gcta64 \
        --mgrm ../data/mgrm.txt \
        --reml --reml-no-lrt \
        --pheno ../data/height.phen \
        --mpheno 1 \
        --qcovar ../data/covariates.txt \
        --pheno ../data/height.phen \
        --mpheno 1 \
        --qcovar ../data/covariates.txt \
        --out ../results2/height_zaitlen



# Run GCTA height all
gcta64 \
```

```
        --grm ../data/clean2 \
        --reml --reml-no-lrt \
        --pheno ../data/height.phen \
        --mpheno 1 \
        --qcovar ../data/covariates.txt \
        --out ../results2/height_all

# Run GCTA height relateds
gcta64 \
        --grm ../data/clean_relateds2 \
        --reml --reml-no-lrt \
        --pheno ../data/height.phen \
        --mpheno 1 \
        --qcovar ../data/covariates.txt \
        --out ../results2/height_relateds


# Run GCTA height unrelateds
gcta64 \
        --grm ../data/clean_unrelateds2 \
        --reml --reml-no-lrt \
        --pheno ../data/height.phen \
        --mpheno 1 \
        --qcovar ../data/covariates.txt \
        --out ../results2/height_unrelateds
```

Appendix G: Published Paper attached.

Appendix H: Ethics approval, consent forms and relevant permissions obtained from journals for re-use of figures in thesis, attached