Family-Based Investigation of the Genetics and Epigenetics of Obesity in Qatar

Mashael Nedham A J Alshafai

A Thesis Submitted for the Degree of Doctor of Philosophy

Department of Genomics of Common Disease

School of Public Health

Imperial College London

July 2015

Abstract

Single nucleotide polymorphisms (SNPs), copy number variations (CNVs) and DNA methylation patterns play a role in the susceptibility to obesity. Despite the alarming figures of obesity in the Arab world, the genetics of obesity remain understudied in Arabs.

Here, I recruited ten multigenerational Qatari families segregating obesity, and generated genome-wide SNP genotyping, whole genome sequencing and genome-wide methylation profiling data using Illumina platforms to investigate the role of rare mutations, CNVs, and DNA methylation patterns in the susceptibility to obesity.

For the identification of obesity mutations, I first identified candidate obesity regions through linkage and run of homozygosity analyses, and then investigated these regions to detect potential deleterious mutations. These analyses highlighted putative rare variants for obesity risk at *PCSK1*, *NMUR2*, *CLOCK* and *RETSAT*. The functional impact of the *PCSK1* mutation on obesity was previously demonstrated while for the other candidates further work is needed to confirm their impact.

For the identification of common obesity CNVs, I performed a genome-wide CNV association analysis with BMI and identified a common duplication of ~5.6kb on 19p13.3 that associates with higher BMI. Moreover, I investigated large (≥500kb) rare CNVs and identified a ~618kb deletion on 16p11.2 in the most extremely obese subject in my samples, which confirms the contribution of the previously reported 16p11.2 deletions to severe obesity beyond European populations.

For the identification of DNA methylation changes in obesity, I attempted to replicate known associations with BMI from large EWASs. I replicated the associations at *ABCG1* (P-value=$6.3X10^{-3}$) and *CPT1A* (P-value=$8.7X10^{-5}$) and compared the effects observed to the TwinsUK cohort through a meta-analysis. I observed low heterogeneity between the two studies, and that increased the associations at *ABCG1* (P-value=$2.5x10^{-13}$) and *CPT1A* (P-value=$1.9x10^{-15}$).

In conclusion, the work represents the first genetic and epigenetic study of obesity in an Arab population. I replicated known genetic and epigenetic obesity susceptibility loci and also discovered novel potential loci.

Table of Contents

Abstract 2
Table of Contents 3
List of Figures 7
List of Tables 9
Acknowledgements 10
Declaration of Originality 11
Copyright Declaration 12
Publications, Presentations and Awards 13
Abbreviations 15
Preface 19
1.1 Epidemic of Obesity 21
1.2 Obesity in Qatar 21
1.3 Physiology of Obesity 23
1.4 Heritability of Obesity 26
1.5 Genetic Basis of Obesity 27
1.5.1 Monogenic Forms of Obesity 27
1.5.1.1 The Role of Leptin and Leptin Receptor in Obesity 27
1.5.1.2 The Role of other Leptin-Melanocortin Pathway Genes in Obesity 29
1.5.2 Syndromic Forms of Obesity 32
1.5.3 Polygenic Form of Obesity 32
1.5.3.1 Linkage Studies on Obesity 33
1.5.3.2 Candidate Gene Association Studies on Obesity 34
1.5.3.3 Genome-Wide Association Studies (GWASs) on Obesity 35
1.7 The Role of Rare Mutations in the Development of Obesity 42
1.7.1 Detection Methods of Rare Mutations 42
1.7.2 Rare Mutations and Obesity 44
1.8 The Role of Copy Number Variations (CNVs) in the Development of Obesity 44
1.8.1 Mechanisms of CNV Formation 45
1.8.2 Methods of CNV Detection 46
1.8.3 Copy Number Variations and Obesity 48

1.9 The Role of DNA Methylation Patterns in the Development of Obesity 50
1.9.1 Formation Mechanisms of DNA Methylation Patterns 51
1.9.2 Detection Methods of DNA Methylation Patterns 52
1.9.3 DNA Methylation Patterns and Obesity 53
1.9.3.1 Candidate Gene Methylation Studies on Obesity 54
1.9.3.2 Genome-Wide Methylation Studies on Obesity 55
1.10 Aims 58
2. Materials and Methods 60
2.1 Ethical Approval of the Study 60
2.2 Selection of Study Subjects 60
2.3 Samples Collection 60
2.4 Phenotypic Data Collection 60
2.5 Extraction of Genomic DNA 64
2.6 Generation of Genomic Data 64
2.6.1 Genome-Wide SNP Genotyping using the HumanOmni2.5-8 Beadchip 65
2.6.2 Whole Genome Sequencing (WGS) from Paired-End Reads 67
2.6.3 Genome-Wide Methylation Profiling using Infinium HumanMethylation450 BeadChip 68
2.7 Additional Genomic Data Sets 71
2.7.1 Data for 108 Qatari Genomes 71
2.7.2 Data from Adult Obesity and Childhood Obesity Case-Control Cohorts 71
2.7.3 Data from the Saudi Extreme Obesity and Mental Retardation Case-Control Sample Set 71
2.7.4 Data from Qatari Families with T2D 71
2.7.5 Data from TwinsUK Cohort 72
2.8 Data Processing and Analysis 72
2.8.1 Identification of Rare Obesity Mutations 72
2.8.1.1 Linkage Analysis 73
2.8.1.2 Run of Homozygosity (ROH) Analysis 80
2.8.1.3 Processing of WGS Data 81
2.8.2 Identification of Common and Rare Obesity CNVs 93
2.8.2.1 CNV Prediction using PennCNV 93
2.8.2.2 Analysis of the 4,481 CNVs Remained after Quality Control 98
2.8.3 Identification of Differentially Methylated Sites in Obesity 104

2.8.3.1 Extraction of DNA Methylation Data 104
2.8.3.2 Quality Control of DNA Methylation Data 105
2.8.3.3 Association of DNA methylation with BMI 111
2.8.3.4 Selection of CpG Associations with BMI for Replication 112
2.8.3.5 Association of DNA Methylation with BMI in the TwinsUK Cohort 112
2.8.3.6 Meta-analysis of the DNA Methylation association studies with BMI using the Qatari and the TwinsUK data 113
3. Identification of Rare Obesity Mutations 115
3.1 Identification of Candidate Obesity Regions 115
3.1.2 Linkage Analysis 116
3.1.3 Run of Homozygosity Analysis 120
3.2 Identification of Candidate Obesity Variants 122
3.2.1 Whole Genome Sequencing and Quality Control 122
3.2.2 Ingenuity Variant Analysis (IVA) 122
3.2.3.1 The Rare Variants in *LEPR* and *PCSK1* Genes 126
3.2.3.2 The Rare Variant in *NMUR2* Gene 126
3.2.3.3 The Rare Variant in *CLOCK* Gene 127
3.2.3.4 The Rare Variant in *RETSAT* Gene 127
3.2.4 Replication of the Candidate Rare Obesity Variants 127
3.3 Discussion 132
3.3.1 Candidate Obesity Regions 132
3.3.2 Whole Genome Sequencing 133
3.3.3 Candidate Rare Obesity Variants 134
3.3.3.1 The Rare Variant in *PCSK1* Gene 134
3.3.3.2 The Rare Variant in *NMUR2* Gene 135
3.3.3.3 The Rare Variant in *CLOCK* Gene 136
3.3.3.4 The Rare Variant in *RETSAT* Gene 136
3.3.4 Limitations 137
3.3.5 Future Directions 137
3.4 Conclusions 138
4. Identification of Common and Rare Obesity CNVs 140
4.1 CNV Prediction 140

4.2 Association of Common CNVs with BMI 141
4.3 Investigation of Rare Obesity CNVs 146
4.4 Confirmation of the 16p11.2 Deletion 152
4.5 Discussion 155
4.5.1 Association of Common CNVs with BMI 155
4.5.2 Candidate Rare Obesity CNVs 155
4.5.3 Limitations 157
4.4.4 Future Directions 158
4.6 Conclusions 158
5. Identification of Differentially Methylated Sites with BMI 161
5.1 Replication of Known Associations between DNA Methylation and BMI in the Qatari Samples 161
5.2 Association of DNA Methylation at *ABCG1* and *CPT1A* with BMI in the TwinsUK Cohort 167
5.3 Meta-analysis of the DNA Methylation Association Studies with BMI using the Qatari and TwinsUK Data 167
5.4 Discussion 168
5.4.1 Replicated Associations between DNA Methylation Levels and BMI 168
5.4.2 Challenges of DNA Methylation Studies 170
5.4.3 Limitations 171
5.4.4 Future Directions 172
5.5 Conclusions 173
6. Conclusions 175
References 178
Appendices 206

List of Figures

Figure 1. The physiological regulation of energy balance

Figure 2. Pedigrees of ten Qatari families participated in the study

Figure 3. Workflow for genome-wide SNP genotyping using the Illumina Infinium Assay

Figure 4. Schematic diagram of Illumina sequence by synthesis next-generation sequencing platform

Figure 5. The two assays employed in the HumanMethylation450 BeadChip

Figure 6. A summary of quality control and marker selection steps using Method A

Figure 7. A summary of quality control and marker selection steps using Method B

Figure 8. Information content for Method A compared to Method B for chromosomes 1, 2 and 3

Figure 9. Information Content (IC) for Method A compared to Method B

Figure 10. Whole genome sequencing analysis pipeline

Figure 11. Sequence quality score across all bases for one sample (QO_F9_01)

Figure 12. GC distribution over all sequences for one sample (QO_F9_01)

Figure 13. Improvement in reported Phred scores after performing BAM file recalibration for one sample (QO_F9_01)

Figure 14. High concordance between the genotypes obtained from SNP-array genotyping and WGS

Figure 15. CNV analysis workflow for the identification of common and rare obesity CNVs

Figure 16. The LRR standard deviation for each of the study samples

Figure 17. Number of CNVs detected in each of the study samples

Figure 18. Schematic representation of the Multiplex Ligation-dependent Probe Amplification (MLPA) technology

Figure 19. Boxplots of CpG sites intensities for the two colour channels per sample

Figure 20. Boxplots of the colour adjusted CpG sites intensities per sample after quantile normalisation

Figure 21. Histograms of DNA methylation profiles (b-values) for one of the study subjects (QO_F7_09) showing the effect of BMIQ normalisation

Figure 22. The NPL signals with LOD scores ≥ 1.0 on chromosomes 3, 13 and 14

Figure 23. The VC linkage signals with LOD ≥ 1.5 on chromosomes 16, 10 and 14

Figure 24. Variant prioritization workflow for the identification of potential rare obesity variants

Figure 25. 1% agarose gel for PCR amplification products and Sanger sequencing traces for candidate variants in: A) *PCSK1*, B) *NMUR2*, C) *CLOCK* and D) *RETSAT*

Figure 26. The genotypes of the candidate variants in the families in which they were detected

Figure 27. The frequency of the different CNV states detected in the obese, overweight and lean subjects

Figure 28. LRR and BAF plots of the seven large non-centromeric CNVs detected in seven study subjects

Figure 29. The normalised ratios of peak heights of the MLPA probes for: A) the 16p11.2 deletion carrier and B) a sibling that does not carry the deletion

Figure 30. Read depth across the 16p11.2 deletion region (chr16:29,581,764-30,199,579 ± 100kb; hg19)

Figure 31. QQ-plot for the association between BMI and methylation sites

Figure 32. Histograms of the methylation b-values distribution for (A) cg06500161 at *ABCG1* and (B) cg00574958 at *CPT1A*

Figure 33. Scatterplots of methylation b-values against BMI for (A) cg06500161 at *ABCG1* and (B) cg00574958 at *CPT1A*

List of Tables

Table 1. Genes involved in monogenic forms of obesity

Table 2. Summary of BMI-associated loci identified through the latest GIANT meta-analysis of GWASs

Table 3. Candidate gene studies and epigenome-wide association studies of DNA methylation on obesity

Table 4. Characteristics of the study participants

Table 5. Summary of the genome-wide datasets generated in this study

Table 6. Primers used for PCR-based validation of candidate obesity variants

Table 7. PCR reagents mix for the validation of candidate obesity variants

Table 8. The 16p11.2 probes and the control probes of the SALSA MLPA P343 Autism-1 probemix kit

Table 9. MLPA reaction mix and thermo-cycler conditions

Table 10. Summary of the signals with LOD score of ≥1 for NPL and VC linkage

Table 11. Summary of overlapping regions between linkage and ROH regions

Table 12. Overview of the variants uploaded into IVA

Table 13. Candidate obesity variants identified in this study

Table 14. The 39 CNV regions overlapping the 2,236 SNP probes that were included for the common CNV association analysis with BMI

Table 15. SNP probes associated with BMI at FDR 5%

Table 16. CNV duplications in the database of genetic variants overlapping with the 19p13.3 duplication detected in my study

Table 17. Large (≥500kb) rare non-centromeric CNVs detected in seven study subjects, and validated by visual examination of LRR and BAF plots

Table 18. CNVs in the DECIPHER that overlap with the large rare CNVs detected in my study

Table 19. Known CpG sites associations with BMI that were attempted using my Qatari sample set

Table 20. The top ten associated CpG sites with BMI in my Qatari data

Acknowledgements

I'm thankful to God for giving me the strength, courage and power to start and complete a PhD journey. During my journey, I was surrounded by supportive people without whom the completion of this thesis would not be possible.

I would like to thank my primary supervisor Dr Mario Falchi for his trust and support, and for being always there when I needed. I would also like to thank my secondary supervisor Professor Philippe Froguel for his support and encouragement. I want to deeply thank my supervisor Professor Khaled Machaca, at Weill Cornell Medical College in Qatar (WCMC-Q), for his continuous help and guidance. I really appreciate it. I would like to express a special thanks to Professor Karsten Suhre, the head of the bioinformatics core at WCMC-Q, for supporting and having me in his team. I learned a lot from you.

Also, I would like to thank the Research Division at Qatar Foundation for giving me the PhD scholarship and for funding my research project. My thanks also extend to the team at Qatar Diabetes Association for their help in sample recruitment and to the participants in this study for their generous contribution.

I want to thank Dr Wadha Al Muftah, the person who shared with me the difficult and nice moments throughout the PhD journey. You are a great friend and I wish you all the success in your PhD. I want to also thank my colleagues at Imperial College London Sadia Saeed, Julia El-Sayed Mustafa and Haya Al-Saud as well as my colleagues at WCMC-Q Noha Al Yousri, Shaza Zaglool, Pankaj Kumar, Anna Halama, Shahina Hayat, Sweety Mathew, Hanif Khalak, Gregory Smith, Khalid Fakhro, Maya Dib and Satanay Hubrack for their support, and for being great friends.

I am indebted to my great parents and family members who always surrounded me with love, faith, and support. I'm thankful to you Dad, Mom, Samia, Mohammed, Hanadi, Kholoud, Fatima, Mona, Khalid, Mariam and my lovely nieces and nephews. I want to express a special thanks to my wonderful twin Kholoud who shared with me the love of science for her advice and endless support. Thanks a lot for being always there for me. I want to also thank my close friends Essra, Noor and Fatima for their continuous encouragement. It is finally done!

I would like to deeply thank my husband Faisal who loved me and my dreams for making my big dream comes true. I am lucky to have a great, supportive and caring husband like you. One day, we will tell our children stories about the PhD journey☺.

This thesis is dedicated to the two persons I always admired, my mother Fathya Al-Tayeb and father Dr Nedham Al-Shafai.

Declaration of Originality

I hereby declare that the work I am presenting in this thesis is the result of my work undertaken in the Department of Genomics of Common Disease at Imperial College London, and in Weill Cornell Medical College in Qatar (WCMC-Q) during my PhD period. This work was carried-out under the supervision of my supervisors Dr Mario Falchi, Professor Philippe Froguel, Professor Khaled Machaca, and Professor Karsten Suhre. Recognition is also given for other colleagues for their specific contributions to this work, as stated in the relevant sections of this thesis.

Mashael Nedham A J Alshafai

July 2015

Copyright Declaration

The copyright of this thesis rests with the author and is made available under a Creative Commons Attribution Non-Commercial No Derivatives licence. Researchers are free to copy, distribute or transmit the thesis on the condition that they attribute it, that they do not use it for commercial purposes and that they do not alter, transform or build upon it. For any reuse or redistribution, researchers must make clear to others the licence terms of this work.

Publications, Presentations and Awards

Peer-reviewed publications:

Zaghlool SB, **Al-Shafai M**, Al Muftah WA, Kumar P, Falchi M and Suhre K.(2015). Association of DNA methylation with age, gender, and smoking in an Arab population.*Clinical Epigenetics* (7):6.

Kumar P**, Al-Shafai M,** Al Muftah WA**,** Chalhoub N**,** Elsaid MF**,** Abdel Aleem A and Suhre K. (2014). Evaluation of SNP calling using single and multiple-sample calling algorithms by validation against array base genotyping and Mendelian inheritance. *BMC Research Notes* (7):747.

Falchi M, El-Sayed Moustafa JS, Takousis P, Pesce F, Bonnefond A, Andersson-Assarsson JC, Sudmant PH, Dorajoo R, **Al-Shafai MN,** 35 authors and Froguel P. (2014). Low copy number of the salivary amylase gene predisposes to obesity. *Nature Genetics* (46): 492–497.

Al Muftah WA*, **Al-Shafai M*,** Zaghlool SB,Visconti A, Tsai P, Kumar P, Spector T, Bell J, Falchi M, Suhre K. Epigenetic Associations of Type 2 Diabetes and BMI in an Arab Population. *Clinical Epigenetics.* (Accepted Jan 2016, * indicates first co-authorship).

Conference Abstracts:

Thareja G, Kumar P, **Al-Shafai M**, Kronfol Z, Al Muftah WA, Suhre K, Kumar P. Y-chromosome haplogroup diversity in the State of Qatar. Weill Cornell Medical College in Qatar Retreat. 21st February 2015. Doha, Qatar.

Al-Shafai M, Al Muftah W, Kumar P, Machaca K, Suhre K and Falchi M (2014). The 16p deletion in an extremely obese Qatari patient. Qatar Foundation Annual Research Conference 2014 (ARC'14). 18th-19th November 2014. Doha, Qatar.

Zaghlool S, **Al-Shafai M**, Al Muftah WA, Kumar P, and Suhre K. Inheritance of Methylation in the Qatari Population. Qatar Foundation Annual Research Conference 2014 (ARC'14). 18th-19th November 2014. Doha, Qatar.

Al-Shafai M, Al Muftah W, Machaca K, Froguel P, Suhre K and Falchi M. Genetics of obesity in Qatari families. 5th Pan Arab Human Genetics Conference. 17th-19th November 2013. Dubai, UAE.

Zaghlool S, **Al-Shafai M**, Al Muftah W, Kumar P and Suhre K. Study of Methylation Differences in Qatari Population. 5th Pan Arab Human Genetics Conference. 17th-19th November 2013. Dubai, UAE.

Kumar P, **Al-Shafai M** and SuhreK. Whole genome sequencing of Qatar population reveals functional classes for the variants at different allele frequencies and Inbreeding coefficient. International Conference on Systems Biology (ICSB). 29th August- 4thof September 2013. Copenhagen, Denmark.

Presentations:

Al-Shafai M, Al Muftah W, Kumar P, Machaca K, Suhre K and Falchi M (2014). The 16p deletion in an extremely obese Qatari patient. Qatar Foundation Annual Research Conference 2014 (ARC'14). 18th-19th November 2014. Doha, Qatar.

Awards:

Best Student Poster Presentation Award by Qatar Foundation Annual Research Conference 2014 (2014), second position in the Division of Health research. Doha, Qatar. Presentation title: The 16p Deletion in an Extremely Obese Qatari Patient.

Abbreviations

1kGP: 1000 genome project
ABCG1: ATP-binding cassette, sub-family G (WHITE), member 1 gene
aCGH: array comparative genomic hybridisation
AdipoR1: adiponectin receptor 1
ADO: adulthood obesity
ADRB3: peroxisome proliferator-activated receptor γ and β-adrenergic receptor 3 gene
AGPR: agouti-related protein
ALDOA: aldolase A, fructose-bisphosphate gene
*AMD1:*adenosylmethionine decarboxylase 1 gene
AMY1: salivary amylase locus
BAC: bacterial artificial chromosome
BAF: B-allele frequency
BAM: binary alignment map
BBS: Bardet-Biedl syndrmore
BDNF: brain-derived neurotrophic factor gene
BMI: Body Mass Index
BMIQ: beta mixture quantile dilation
BMR: basal metabolic rate
bp: base pairs
BS–seq: Bisulfite Conversion Followed by Sequencing
b-value: beta-value
CART: cocaine-and amphetamine-related transcript
CDCV: common disease–common variant
CDRV: common disease-rare variant
CGH: comparative genomic hybridisation
CHO: childhood obesity
CLOCK: circadian locomotor output cycles kaput gene
CLUH: clustered mitochondria (cluA/CLU1) homolog gene
CM: centimorgan
CN: copy number
CNVRs: copy number variant regions
CNVs: copy number variations
CpG: cytosine–guanine nucleotide pair
CPT1A: carnitine palmitoyltransferase 1A gene
DD: developmental delay
DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources
DGV: database of genomic variants
dNTPs: deoxynucleotide triphosphate
EDTA: ethylenediaminetetraacetic acid
ETV5: Ets variant 5 gene

EWASs: epigenome-wide association studies
FDR: false discovery rate
FFM %: fat-free mass percentage
FFM: fat-free mass
FM %: fat mass percentage
FM: fat mass
FTO: fat mass and obesity associated gene
GABA: gamma-aminobutyric acid
GATK: genome analysis toolkit
GC: guanine-cytosine
GHSRs: secretagogue receptors
GIANT: genetic investigation of anthropometric traits
GNPDA2: glucosamine-6-phosphate deaminase 2 gene
GPR120: G-protein coupled receptor 120 gene
GWASs: genome-wide association studies
HCS: hiSeq control software
Het: heterozygous
HIF3A: hypoxia inducible factor 3, alpha subunit gene
HMM: Hidden Markov Model
hnRNPs: heterogeneous nuclear ribonucleoproteins
HWE: Hardy-Weinberg Equilibrium
IC: information content
IDT: Integrated DNA Technologies
IGF-1: insulin-like growth factor
IRB: Institutional Review Board
IRX3: iroquoishomeobox 3 gene
IVA: Ingenuity Variant Analysis
kb: kilobase
KCTD13: potassium channel tetramerization domain containing 13 gene
KCTD15: potassium channel tetramerization domain containing 15 gene
KIF22: kinesin family member 22 gene
KLF13: kruppel-like factor 13 gene
KSR2: kinase suppressor of ras 2
LD: linkage disequilibrium
LEP: leptin gene
LEPR: leptin receptor gene
LINE-1:Long Interspersed Nucleotide Element 1
LOD scores: logarithm of the odds (to the base 10)
*LPPR3: l*ipid phosphatase-related protein type 3 gene
LRR: log R ratio
LUMA: Luminometric assay
MAF: minor allele frequency
MAF: v-maf avian musculoaponeurotic fibrosarcoma oncogene homolog gene

MC4R: melanocortin 4 receptor
MCH: melanin-concentrating hormone
MeDIP: Methylated DNA Immunoprecipitation
miR-221: microRNA-221
MLPA: Multiplex Ligation-Dependent Probe Amplification
MS PCR: Methylation-specific PCR
MSRA: methionine sulfoxide reductase a gene
MTCH2: mitochondrial carrier homolog 2 gene
NAHR: non-allelic homologous recombination
NEGR1: neuronal growth regulator 1 gene
NGS: next generation sequencing
NHEJ: non-homologous end joining
NMUR2: neuromedin U receptor 2 gene
NPC1: niemann-pick disease type c1 gene
NPL: nonparametric multipoint linkage analysis
NPY: neuropeptide Y
PC1 and PC2: prohormone convertases
PCR: polymerase chain reaction
PCSK1: proprotein convertase subtilisin/kexin type 1 gene
PFB: population frequency of B allele
POLYPHEN-2: Polymorphism Phenotyping
POMC: pro-opiomelanocortin
PopPANTE: Population and Pedigree Association Testing of the Epigenome
PTBP1: polypyrimidine tract binding protein 1 gene
PTER: phosphotriesterase related gene
PTER: phosphotriesterase related gene
pVAAST: Pedigree Variant Annotation, Analysis, and Search Tool
PWS: Prader-Willi syndrome
PYY3–36: peptide YY3–36
QC: quality control
QDA: Qatar Diabetes Association
QN: quantile normalisation
QQ: quantile-quantile
QTLs: quantitative trait loci
r^2 : pairwise correlation coefficient
RETSAT: retinol saturase gene
*RNF145:*ring finger protein 145 gene
ROH: run of homozygosity
ROMA: Representational Oligonucleotide Microarray Analysis
RRBS-seq: Reduced Representation Bisulfite Sequencing
RTA: Real-Time Analysis
SAM: sequence alignment map
SBS: sequencing by synthesis

SD: standard deviation
SDCCAG8: serologically defined colon cancer antigen 8 gene
SEC16B: SEC16 homolog B (S. Cerevisiae) gene
SH2B1: SH2B adaptor protein 1 gene
SIFT: sorting intolerant from tolerant
SIM1: single-minded family bHLH transcription factor 1
SNPs: single nucleotide polymorphisms
SOCS3: suppressor of cytokine signalling 3
T2D: type 2 diabetes mellitus
TAE: tris-acetate EDTA
Tm: annealing temperature
TMEM18: transmembrane protein 18 gene
TNKS: tankyrase trf1-interacting ankyrin-related adp-ribose polymerase gene
TRH: thyrotrophin-releasing hormone
VC: variance component
VCF: variant call format
WCMC-Q: Weill Cornell Medical College in Qatar
WES: whole exome sequencing
WGS: whole genome sequencing
WHO: world health organization
Y1R: Y1 receptor
Y2Rs: Y2 receptors
α-MSH: α-melanocyte-stimulating hormone

Preface

Major advances in the field of genetic research have been achieved with the development of high throughput technologies. These technologies were employed to better understand the phenotypic variation between humans especially in relation to diseases. In this thesis, I investigate the role of several sources of genetic variations including single nucleotide polymorphisms, copy number variations and DNA methylation patterns in the susceptibility to obesity.

Chapter one introduces obesity and the genetic and epigenetic research in the field of obesity. This includes the relevant literature and the progress that have been made in this filed in terms of findings, strategies and technologies, and providing justification for the research carried out in this thesis.

Chapter two gives a description of the materials and methods used to perform the experimental and analytical work of this thesis.

Chapter three shows the results and discussions of linkage, run of homozygosity and rare variants analyses for the purpose of the identification of rare obesity mutations.

Chapter four presents the results and discussions of the CNV analyses that were carried out for the identification of common and rare obesity CNVs.

Chapter five describes the results and discussions of the DNA methylation analysis that was performed to identify differentially methylation CpG sites with BMI.

Chapter six discusses the work presented in this thesis in a wider context and describes potential future research directions to provide better understanding of the contribution of genetic and epigenetic changes to obesity predisposition.

CHAPTER ONE

Introduction

1. Introduction

1.1 Epidemic of Obesity

The prevalence of obesity has reached epidemic levels globally. In 2014, there were more than 1.9 billion overweight adults (age of 20 years or older) and of these more than 600 million were obese (World Health Organization 2015a). In addition, in 2013 there were more than 42 million overweight children, under the age of 5 years, with a higher rate of increase (more than 30%) observed in developing countries compared to developed countries (World Health Organization 2015a). Obesity is characterized by excessive fat accumulation in the body as a result of imbalance between energy intake and expenditure. Both in clinical medicine and in research, obesity is commonly monitored through measuring the Body Mass Index (BMI), which is calculated as weight in kilograms divided by the square of the height in meters squared. According to the World Health Organization (WHO), adults with a BMI of 25kg/m^2 or above are classified as overweight while adults with a BMI of 30kg/m^2 or above are considered obese (World Health Organization 2015a). For subjects below the age of 20 years, the WHO defines overweight and obesity as BMI of more than one standard deviation (SD) and two SDs respectively, above the mean BMI for the specific age and gender (World Health Organization 2015b).

The increasing incidence of obesity has become a major health concern since obesity is a prominent risk factor for other public-health burdening conditions such as type 2 diabetes mellitus (T2D), cardiovascular diseases, hypertension and certain forms of cancer (Haslam and James 2005). In addition to the medical consequences, obesity is associated with significant social and economic challenges particularly in terms of the cost required for treatment of these obesity-associated diseases (Wang et al. 2011).

Obesity is a complex disease, which results from compound interactions between genetic and environmental factors such as diet and physical activity. Genetic risk factors are thought to predispose a proportion of the population to complex diseases like obesity, while environmental risk factors are believed to modulate the disease progression and severity (Panoutsopoulou and Zeggini 2009).

1.2 Obesity in Qatar

The State of Qatar is a peninsula located amid the western coast of the Arabian Gulf with an area of 11,521 square kilometers. The population size is about two million residents (Tadmouri et al. 2009), of which ~300,000 Qataris owe their ancestry to a combination of three major subgroups: The Bedouin,

who originally migrated from the Nagid Plateau, the Persian who originally migrated along the coast of the Arabian Gulf, and the Africans who originally migrated to Qatar from Africa through trading routes or brought as slaves (Hunter-Zinck et al. 2010).

Similar to many modern societies, the Qatari population is facing increasing rates of complex disorders such as obesity, T2D and cardiovascular diseases. Today, Qatar is one of the countries with highest obesity prevalence in the world with 42% of the Qataris being obese and 70% being overweight (World Health Organization 2015c). The country is also ranked high in the incidence of obesity-associated diseases such as T2D, in which the country ranks first globally with an estimated incidence of 23% among Qataris (Scully 2012).

A major contributor to the problem is the rapid urbanization in the country after the discovery of oil and gas, resulting in lifestyle changes towards reduced physical activity and increased food consumption. However, as seen in other populations, genetic factors are likely to play a role in determining individual susceptibility to obesity, in such an "obesegenic" environment.

One of the hypotheses proposed by geneticists to explain the burden of metabolic diseases such as obesity and T2D and the differential susceptibility between different ethnic populations is the thrifty gene hypothesis (Diamond 2003; Neel 1999). The hypothesis suggests that variants which predispose individuals to such diseases were historically advantageous in societies with food scarcity; therefore evolutionary pressures have favoured their persistence to prevent starvation (Neel 1999; Walley, Asher, Froguel 2009). This hypothesis might be relevant for the Qatari society, which has experienced long periods of food scarcity in the past, before the discovery of oil and gas, followed continuous abundance of food in the last 50 years following the rapid development in the country due to oil and gas revenues. However, the thrifty gene hypothesis has not been supported by genetic studies which showed absence of variation in the prevalence of some obesity risk variants, like those close to *FTO,* in isolated populations with high obesity prevalence (Walley, Asher, Froguel 2009).

Demographic studies suggest that the rapid increase in obesity incidence in the globe within the last 2-3 generations cannot be explained by genetic processes only, and the genetic studies which showed that obesity risk variants cannot account for much of the disease heritability have supported this idea (Drong, Lindgren, McCarthy 2012; Jaenisch and Bird 2003; Manolio et al. 2009). Epigenetic processes such as changes in DNA methylation patterns are believed to contribute to the susceptibility of complex disease (Drong, Lindgren, McCarthy 2012; Jaenisch and Bird 2003). DNA methylations patterns can be affected

by genetic factors as well as environmental factors including intra-uterine conditions and early life exposures. In addition, DNA methylation patterns might also show transgenerational inheritance (Drong, Lindgren, McCarthy 2012; McRae et al. 2014).

Despite the alarming figures of obesity in the Qatari population, no genetic or epigenetic studies on obesity have been performed on the Qatari population so far making this study the first in that area. Conducting such studies on the Qatari population has many advantages. First, it can identify novel obesity variants, genes, and/or pathways, since the population has been understudied in the obesity research field, and can also highlight conserved findings from studies on other ethnicities. Second, it confers a good opportunity to map recessive obesity genetic variants due to the high rates of consanguinity in the population, which is estimated at about 54% (Bener and Alali 2006). Third, the Qatari population has more uniform genetic and environmental backgrounds, compared to outbred populations, which may simplify the study of complex traits like obesity. Fourth, Qatari families are usually large intact families, due to cultural traditions and reduced rates of emigrations among Qataris, which is advantageous for conducting family-based studies in particular.

1.3 Physiology of Obesity

The physiological mechanisms underlying obesity are complex and involve multiple organ systems and signalling regulators, and pathways. One of the key players in the process are the adipocytes which store surplus fat in the body as triacylglycerol and release non-esterified fatty acids into the bloodstream when fat stores are mobilized (Walley, Asher, Froguel 2009). Both adipocytes number and size have been found to be important in relation to obesity. A striking study conducted by Spalding *et al.* (Spalding et al. 2008) has shown that the number of adipocytes in the human body increases steadily during childhood and adolescence irrespective of weight while it remains constant during adulthood, which suggests that changes in fat mass in adults are determined by adipocytes size rather than number. Also, it means that people with early-onset obesity have a larger number of adipocytes when they reach adulthood compared to lean children. This may explain why about 75% of obese children remain obese during adulthood while only 10% of lean children become obese as adults (Freedman et al. 2001).

Recently, obesity has also been viewed as a neuroendocrine disorder of energy balance controlled mainly by the brain. The process involves the interpretation and integration of several neural and hormonal signals, from different tissues, in the arcuate nucleus of the hypothalamus, which contains two sets of neurons: neuropeptide Y (NPY) and agouti-related protein (AGPR) neurons, and the pro-

opiomelanocortin (POMC) and cocaine-and amphetamine-related transcript (CART) neurons (Barsh and Schwartz 2002). Activation of NPY/AGPR neurons has orexigenic effect, promoting food intake and reducing energy expenditure, while the activation of the POMC/CART neurons has the opposite anorexigenic effect (Barsh and Schwartz 2002).

For example, leptin, which is secreted from adipose tissue, and insulin, which is secreted from the pancreas exert their anorexigenic effect through the activation of the POMC/CART neurons and the inhibition of the NPY/AGPR neurons (Friedman and Halaas 1998). Another anorexigenic signalling molecule in the process is the peptide YY3–36 (PYY3–36), which is secreted from the distal gastrointestinal tract, and exerts an inhibitory influence on the NPY/AGRP neurons though Y2 receptors (Y2Rs) present on those neurons. On the other hand, the orexigenic peptide ghrelin, which is secreted primarily by the stomach and duodenum, activates the NPY/AGPR neurons through their hormone secretagogue receptors (GHSRs) and consequently promotes food intake (Bell, Walley, Froguel 2005; Kohno et al. 2003). Moreover, the NPY/AGRP neurons exhibit an inhibitory influence on the POMC/CART neurons mediated through the release of gamma-aminobutyric acid (GABA).

The signals produced by POMC/CART neurons and NPY/AGPR neurons are then sent to downstream effector neurons including the melanin-concentrating hormone (MCH) neurons, orexin or hypocretin neurons and thyrotrophin-releasing hormone (TRH) neurons, that express receptors including Y1 receptor (Y1R) and the melanocortin 4 receptor (MC4R) (Bell, Walley, Froguel 2005). These neurons also receive further input signals from dopamine, serotonin and endocannabinoid to ultimately determine the balance between energy expenditure and food intake. The physiological regulation of energy balance via the arcuate nucleus neurons is summarized in Figure 1.

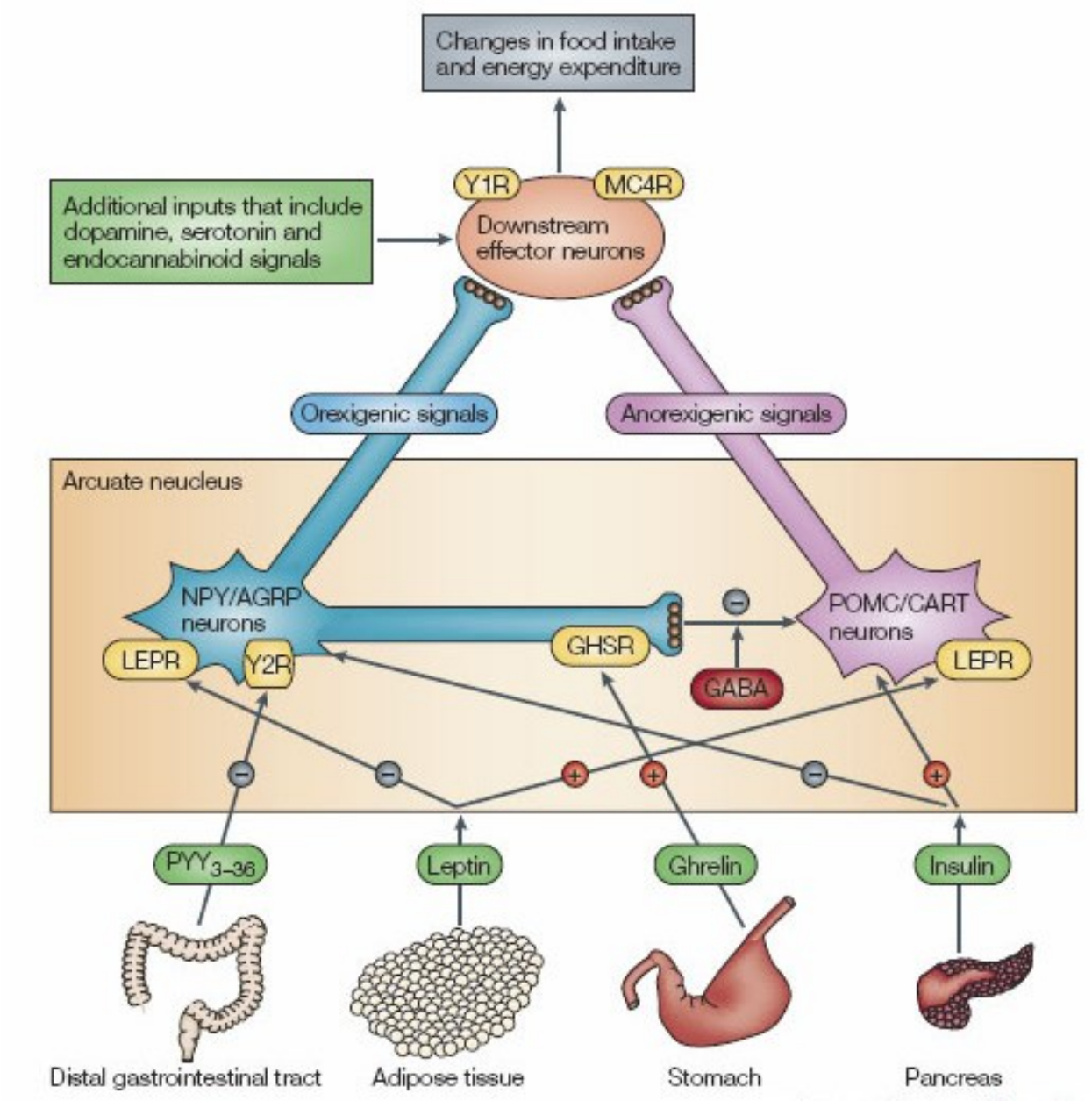


Figure 1. The physiological regulation of energy balance. Detailed description of the figure is given in Section 1.3. Figure is reproduced from Bell *et al* (Bell, Walley, Froguel 2005) with permission (See Appendix 2.1).

1.4 Heritability of Obesity

Although the increasing prevalence of obesity worldwide has been driven by environmental changes, particularly changes in diet and lifestyle leading to over-nutrition and reduced physical activity, inter-individual differences in susceptibility or resistance to obesity is strongly influenced by genetic factors. Early indications of the genetic component of obesity came from the observation of variation in obesity prevalence between different racial groups, from about 5% among Caucasian and Asian populations to about 50% in Pima Indians and South Sea Island populations (Knowler et al. 1990), as well as the long-noted familial aggregation of obesity occurrence.

Twin and adoption studies have estimated the heritability of obesity, the proportion of phenotypic variance attributable to genetic variance, at 40-70% (Elks et al. 2012; Maes, Neale, Eaves 1997). Even though these estimates might be inflated, they show that obesity is one of the strongest genetically influenced traits (Willyard 2014). In addition, the existence of a number of rare monogenic and syndromic forms of obesity has clearly highlighted the genetic aspect of obesity, as will be discussed in Sections 1.5.1 and 1.5.2 respectively.

Environmental factors also play a role in modulating the effect of some of the genetic risk factors in obesity. For example, it has been shown that changes in lifestyle and physical activity can dilute or amplify the effect of the obesity risk variants near the fat mass and obesity-associated (*FTO*) gene (Silventoinen et al. 2009). Moreover, other factors like age and gender, which can be assimilated as specific environmental conditions, have also been reported to modulate the effect of obesity risk variants. For example, carriers of pathogenic obesity mutations in melanocortin 4 receptor (*MC4R*), which cause a monogenic form of obesity, have displayed age-dependent penetrance where the likelihood of obesity increases with age, as well as, gender-dependent effect on BMI where females tend to have twice the variant effect size seen in males (Stutzmann et al. 2008). Such gender-specific differences in the effect sizes of risk obesity variants may reflect in part the fact that females are at higher risk of developing morbid obesity than males (Ogden et al. 2006). These studies highlight the existence of interaction between genetic and environmental factors in determining obesity susceptibility, and hence underscore the complexity of the disease.

1.5 Genetic Basis of Obesity

On the genetic level, obesity can be classified into three groups: 1) Monogenic obesity, which results from single gene mutations, 2) Syndromic obesity, where obesity presents in some genetic syndromes as a central clinical feature (Perrone et al. 2010), and 3) Polygenic obesity, which is associated with genetic variants at multiple loci conferring susceptibility to obesity. These classes of obesity will be introduced in sections 1.5.1, 1.5.2 and 1.5.3 respectively.

1.5.1 Monogenic Forms of Obesity

Monogenic forms of obesity provided direct evidence for the genetic role in the development of obesity, and contributed to our understanding of some of the cellular pathways involved in obesity development and in body weight regulation. Monogenic obesity represent about 2% of obesity cases and 5-6% of severe obesity cases (Froguel and Blakemore 2008), and it is usually caused by rare mutations (minor allele frequency (MAF) < 1%) with large effect size. Cases of monogenic obesity exhibit an extreme obesity phenotype with early onset and they often have other associated features like hyperphagia (Froguel and Blakemore 2008; Walley, Asher, Froguel 2009).

1.5.1.1 The Role of Leptin and Leptin Receptor in Obesity

Animal studies of monogenic obesity gave the first confirmation that single gene mutations can result in obesity. Studies of a severely obese strain of mice, ob/ob, provided early insights about the first obesity gene (Ingalls, Dickie, Snell 1950). In 1994, this strain has been characterized by a recessive mutation in the *ob* gene, which is expressed predominantly in white adipose tissue (Ingalls, Dickie, Snell 1950). The cloning and characterization of the *ob* gene facilitated the identification of its homologue in humans, the leptin gene (*LEP*) (Zhang et al. 1994).

Leptin is a hormone-like protein that is secreted primarily by adipocytes in concentrations proportional to body adipose stores (Considine et al. 1996), and it acts as a key component of the leptin-melanocortin system, which regulates food intake and energy homeostasis (Farooqi et al. 2003). This system involves the secretion of leptin by adipocytes in the bloodstream. Circulating leptin in the bloodstream binds to leptin receptors (*LEPR*) in the hypothalamic arcuate nucleus leading to the activation of the POMC/CART neurons and the inhibition of the NPY/AGPR neurons, and the regulation of the expression of many key genes encoding downstream effectors in the pathway. One of the genes that are up-regulated is pro-opiomelanocortin (*POMC*). The POMC protein is usually cleaved by prohormone convertases (PC1 and

PC2) to generate the α-melanocyte-stimulating hormone (α-MSH), the ligand for the melanocortin 4 receptor (MC4R). The binding of α-MSH to MC4R results in the activation of downstream signals to inhibit anabolic pathways and to stimulate catabolic pathways, leading to reduced food intake and increased energy expenditure, respectively (Farooqi et al. 2003).

Further studies on the ob/ob strain have shown that the exogenous administration of recombinant leptin results in complete reversal of the obesity phenotype in this strain (Campfield et al. 1995; Halaas et al. 1995). The recombinant leptin is also currently applied as a treatment in humans (Farooqi et al. 1999).

In relation to obesity in humans, one of the first breaks came in 1997, with the identification of a rare homozygous mutation in *LEP* in patients with severe early-onset obesity (Montague et al. 1997). This discovery provided the first evidence that monogenic obesity also exist in man. The mutation was identified in two obese cousins from a highly consanguineous family of Pakistani origin. The mutation resulted in very low serum leptin (< 1ng/ml) levels in these patients despite their elevated fat mass, a condition termed as congenital leptin deficiency (Montague et al. 1997). This condition is usually associated with hyperphagia, hyperinsulinaemia, hypothyroidism and hypogonadism (Gibson et al. 2004; Ozata, Ozdemir, Licinio 1999; Saeed et al. 2012; Strobel et al. 1998). Screening subjects with severe early-onset obesity for mutations in *LEP*, as a candidate gene, allowed the identification of additional mutations causing leptin deficiency (Gibson et al. 2004; Ozata, Ozdemir, Licinio 1999; Saeed et al. 2012; Strobel et al. 1998). These mutations include missense, nonsense as well as farmeshift mutations and they are usually inherited in an autosomal recessive manner.

Currently, recombinant leptin is available as a treatment to reverse the phenotypic aspects of leptin deficiency and about 30 patients worldwide were cured of their obesity phenotype so far (Farooqi et al. 1999). Despite the rarity of this condition, the ability to provide effective personalized medicine for this group of patients represents a successful and motivating example in the field of translational medicine in obesity.

Animal studies on another strain of mice (db/db) with severe diabetes, obesity and dislipidaemia, have revealed a causative recessive mutation in the ob (leptin) receptor gene (Chua et al. 1996). The leptin receptor (LEPR) is abundantly expressed in the hypothalamus. The same study showed that the administration of leptin could not reverse the obesity phenotype in this strain consistent with the molecular phenotype (Chua et al. 1996). Homozygous mutations in the *LEPR* have also been found in

patients with severe obesity usually associated with some other features such as hyperphagia, reduced insulin-like growth factor (IGF-1), growth abnormalities and delayed or absent puberty (Clement et al. 1998; Farooqi et al. 2007; Mazen et al. 2011). Similar to mutations in *LEP*, *LEPR* mutations also showed an autosomal recessive pattern of inheritance.

1.5.1.2 The Role of other Leptin-Melanocortin Pathway Genes in Obesity

Further investigations of patients with severe early-onset obesity have shed light on four monogenic obesity genes other than *LEP* and *LEPR*. Interestingly, these genes also encode for effectors involved in the leptin-melanocortin pathway. One of these monogenic obesity genes is *POMC* in which homozygous loss of function mutations resulting in POMC deficiency have been identified in subjects with severe early onset obesity, and with other associated features such as red hair pigmentation, pale skin and hyperphagia (Challis et al. 2002; El-Sayed Moustafa and Froguel 2013; Krude et al. 1998; Krude and Gruters 2000). These mutations displayed a recessive mode of inheritance. Similar to the phenotype of *POMC*-deficient patients, *pomc* knock-out mice are characterized with obesity, defective adrenal development and altered pigmentation, which supports the conserved role of POMC across different species (Yaswen et al. 1999).

Furthermore, homozygous, compound heterozygous and heterozygous loss-of-function mutations in proprotein convertase subtilisin/kexin type 1 (*PCSK1)* gene, which encodes the enzyme responsible for the cleavage of POMC, were also reported to cause severe obesity (Farooqi et al. 2007; Jackson et al. 1997; Philippe et al. 2015). Phenotypes associated with obesity due to *PCSK1* mutations include abnormal glucose homeostasis, reduced plasma insulin, hypogonadism and hypocortisolaemia (Farooqi et al. 2007; Jackson et al. 1997). The causative mutations in *PCSK1* showed autosomal recessive or autosomal dominant modes of inheritance (Farooqi et al. 2007; Jackson et al. 1997; Philippe et al. 2015). In addition, heterozygous mutations causing partial PCSK1 deficiency confer increased risk for obesity (Creemers et al. 2012). Surprisingly, *pcsk1* knock-out mice do not develop obesity and are characterized with dwarfism and defects in growth hormone processing (Zhu et al. 2002), suggesting a potential variation in the function of PCSK1 across different species. However, Lloyd *et al.* (Lloyd, Bohan, Gekakis 2006) have reported a novel mutation in *pcsk1* in mice (N222D) that leads to impaired catalytic activity of PC1, reduced levels of α-MSH, and the development of obesity, similar to PCSK1 deficiency in humans.

In addition, mutations in the melanocortin 4 receptor (*MC4R*) are the most common cause of monogenic obesity affecting 2-4% of severely obese individuals (Farooqi et al. 2003; Hinney et al. 2006; Vaisse et al. 1998; Yeo et al. 1998). Loss of function mutations in *MC4R* have been identified in subjects with severe obesity (Dubern et al. 2001; Farooqi et al. 2000; Lubrano-Berthelier et al. 2006; Stutzmann et al. 2008). Unlike most of the monogenic obesity forms discussed above, which are recessively inherited, mutations in *MC4R* are inherited in autosomal dominant, co dominant or even autosomal recessive manner (Dubern et al. 2001; Farooqi et al. 2000; Hinney et al. 2006; Saeed et al. 2012; Stutzmann et al. 2008; Vaisse et al. 1998).

Knock out studies of *mc4r* in mice have demonstrated that the disruption of a single copy of the gene can result in increased food intake, severe obesity, and hyperinsulinemia, which confirmed the preserved role of MC4R in the regulation of body weight across species (Huszar et al. 1997). Rhythm, a pharmaceutical company based on Boston, has developed a compound called ”RM-493” for the treatment of obesity in patients with MC4R deficiency due to heterozygous mutations in *MC4R* gene (Rhythm 2013). The compound is a novel MC4R agonist that aims to restore MC4R function through increasing the signalling activity of the normal copy of the receptor. In 2013, the company launched a clinical trial to investigate the effect of RM-493 on weight loss in obese subjects with normal copies of *MC4R*, and they demonstrated its efficacy on increasing energy expenditure and weight loss (Rhythm 2013). The follow-up clinical trials will address the same question but in patients with MC4R deficiency.

Furthermore, monogenic obesity can result from loss of function mutations in single-minded family bHLH transcription factor 1 (*SIM1*) gene, which encodes a transcription factor that has essential role in neurogenesis (Bonnefond et al. 2013; Holder, Butte, Zinn 2000; Montagne et al. 2014; Ramachandrappa et al. 2013; Zegers et al. 2014). This transcription factors is believed to act upstream and downstream of MC4R, and contributes to the regulation of energy homeostasis. Severely obese subjects due to mutations in *SIM1* are characterized by other associated features such as hypotonia and developmental delay (Bonnefond et al. 2013; Holder, Butte, Zinn 2000; Montagne et al. 2014). Functional studies have shown that *Sim1* deficiency is sufficient to cause hyperphagic obesity in mice and have highlighted the role of sim1-expressing neurons in regulating food intake through the leptin-melanocortin pathway (Kublaoui et al. 2006; Tolson et al. 2010). The genes involved in the monogenic forms of obesity to date are summarized in Table 1.

Collectively, monogenic obesity studies have drawn a picture of obesity as a neuroendocrine disorder controlled mainly by the functioning of the hypothalamus and caused by appetite dysfunction due to the leptin-melanocortin pathway impairment. In addition, the large effect size of the various monogenic obesity mutations gives them useful predictive value in clinical setting at the level of individual cases. Thus, genetic screening for known monogenic obesity mutations can be useful particularly in cases with extreme early onset obesity, and especially in cases displaying other features such as hyperphagia or developmental delay (Blakemore and Froguel 2010; El-Sayed Moustafa and Froguel 2013).

Gene (gene symbol)	Cytoband	Protein role	Mode of inheritance	Knock-out animal models	Reference
Leptin (*LEP)*	7q31.3	A hormone-like protein involved in regulation of food intake and energy expenditure, through leptin-melanocortin pathway.	Autosomal recessive	Obese	(Gibson et al. 2004; Montague et al. 1997; Ozata, Ozdemir, Licinio 1999; Saeed et al. 2012; Strobel et al. 1998)
Leptin Receptor (*LEPR)*	1p31	A receptor for leptin involved in hypothalamic signaling and appetite control.	Autosomal recessive	Obese	(Clement et al. 1998; Farooqi et al. 2007; Mazen et al. 2011)
Melanocortin-4 receptor (*MC4R*)	18q21.32	A receptor for melanocortin family, integral in hypothalmatic signaling for appetite control.	Autosomal dominant/ co-dominant/ recessive	Obese	(Dubern et al. 2001; Farooqi et al. 2000; Huszar et al. 1997; Lubrano-Berthelier et al. 2006; Saeed et al. 2012; Stutzmann et al. 2008)
Pro-opiomelanocortin (*POMC*)	2p23.3	A polypeptide hormone precursor for MC4R ligand (α-MSH), a key player in leptin-melanocortin pathway.	Autosomal recessive	Obese	(Challis et al. 2002; El-Sayed Moustafa and Froguel 2013; Krude et al. 1998; Krude and Gruters 2000; Yaswen et al. 1999)
Pro-hormone convertase subtilisin/kexin type 1 (*PCSK1*)	5q15	An enzyme involved in post-translational processing of POMC.	Autosomal dominant/ recessive	Not obese	(Farooqi et al. 2007; Jackson et al. 1997; Philippe et al. 2015; Zhu et al. 2002)
Single-minded homolog 1 (*SIM1*)	6q16.3	A DNA-binding protein that acts upstream and downstream of MC4R with essential role in neurogenesis and regulation of energy homeostasis.	Autosomal dominant	Obese	(Bonnefond et al. 2013; Holder, Butte, Zinn 2000; Kublaoui et al. 2006; Montagne et al. 2014; Ramachandrappa et al. 2013; Tolson et al. 2010; Zegers et al. 2014)

Table 1.Genes involved in monogenic forms of obesity.

1.5.2 Syndromic Forms of Obesity

Unlike monogenic forms of obesity in which obesity is present as the principle clinical phenotype, syndromic forms of obesity are characterized with obesity as an associated clinical feature with a group of complex clinical phenotypes such as mental retardation, dysmorphic features and developmental delay. To date, there are more than 30 genetic syndromes associated with obesity due to genetic defects or chromosomal abnormalities (O'Rahilly et al. 2003).

Prader-Willi syndrome (PWS) is the most common example of obesity syndromes that occurs in approximately 1 out of 50,000 people (Farooqi 2005). PWS is neurodevelopmental syndrome characterized by developmental delay, inflate hypotonia, hypogonadism and early onset obesity associated with hyperphagia. PWS is caused by deletions or defects at an imprinted region on chromosome 15 (q11-q13), which is usually paternally expressed (Farooqi 2005). This region contains several protein coding genes as well as multiple non-coding snoRNAs. The gene responsible for the PWS phenotype in the region is believed to contain the HBII-85 snoRNA cluster, suggesting a role of this particular family of non-coding RNAs in growth and energy homeostasis (de Smith et al. 2009).

Bardet-Biedl syndrome (BBS) is a rare syndrome characterized with obesity, retinopathy, renal and cardiac defects, polydactyly, learning disabilities and hypogonadism. The prevalence of BBS is 1 in 125,000–160,000 in people of European descent and is higher in consanguineous populations (Guo and Rahmouni 2011). About 80% of BBS cases result from mutations in BBS genes (*BBS1-BBS15*) encoding BBS proteins that are involved in leptin signalling in the hypothalamus (Guo and Rahmouni 2011). In addition, a case-control association study between variants in the *BBS* genes and obesity revealed association between single nucleotide polymorphisms (SNPs) in *BBS2*, *BBS4* and *BBS6* with common obesity (Benzinou et al. 2006), suggesting a role of *BBS* genes even in non-syndromic obesity**.**

1.5.3 Polygenic Form of Obesity

Monogenic and syndromic forms of obesity account for only a small proportion of obesity cases. The majority of obesity cases, on the other hand, are polygenic, resulting from complex interactions between susceptibility variants in multiple genes with environmental risk factors where each individual susceptibility variant has subtle effect on weight (Froguel and Blakemore 2008).

Until 2006, studies aiming to define susceptibility loci underlying complex diseases like common obesity were mainly based on linkage and candidate gene approaches. However, these approaches have been of

limited utility in revealing genes underlying common obesity. The main advances in the field of common obesity came with the development of high throughput genotyping array platforms, which allow the implementation of hypothesis-free association approaches at the genome-wide level. These approaches, known as genome-wide association studies (GWASs), have dramatically increased the speed of gene discovery in the field (Locke et al. 2015; Shungin et al. 2015). These different types of genetic approaches are described below.

1.5.3.1 Linkage Studies on Obesity

Linkage analysis is a family-based approach that aims to locate a genomic region that is co-inherited with a trait of interest in families. This approach can be performed at the genome wide level or at particular candidate regions, and it usually involves the genotyping of family members for polymorphic markers such as SNPs or microsatellites (El-Sayed Moustafa and Froguel 2013). This type of analysis takes advantage of the inverse relationship between the frequency of recombination events between any two genomic loci and their physical distance. Thus, the co-inheritance of markers with the trait of interest suggests that these markers are located close to the trait/disease locus.

Over 60 genome-wide linkage scans have been performed for different obesity-related traits, and they revealed 253 putative obesity quantitative trait loci (QTLs) (Rankinen et al. 2006). These QTLs are mapped to all chromosomes except the Y chromosome and are summarized in the "Human Genome Map" (Rankinen et al. 2006). Although some functionally promising candidates were identified through linkage signals, replication of these results has proved difficult as only 20% of signals were supported by more than one study (Rankinen et al. 2006). A large meta-analysis that combined data from 37 independent obesity linkage studies containing more than 31,000 individuals, have been conducted to increase the statistical power for linkage detection. However, no regions with consistently significant evidence for linkage have been identified (indicated by logarithm of odds score (LOD score) of 3.3 or more) and few suggestive linkage signals (indicated by LOD score of 1.9 or more) were detected at 13q13.2-13q33.1, 12q23-12q24.3 and 16q12.2 (Saunders et al. 2007). The lack of linkage evidence, despite sufficient detection power, suggests high heterogeneity underlying obesity and small effect size of common risk variants on the phenotype. Some of the linkage signals for obesity-related traits have been observed in multiple studies (El-Sayed Moustafa and Froguel 2013) including the signals on 1p36 (Deng et al. 2002; Gorlova et al. 2003; Liu et al. 2004a; Norris et al. 2005; Wilson et al. 1991), 2p21-2p23 (Cai et al. 2004; Moslehi et al. 2003; Palmer et al. 2003; Rotimi et al. 1999), 2q14 (Deng et al. 2002; Fox et al. 2004; Liu et al. 2004b), 3q27 (Kissebah et al. 2000; Luke et al. 2003; Vionnet et al. 2000; Walder et

al. 2000; Wu et al. 2002), 6q21-6q23 (Atwood et al. 2002; Fox et al. 2004; Meyre et al. 2004), 7q31-7q32 (Cai et al. 2004; Liu et al. 2004b; Moslehi et al. 2003; Palmer et al. 2003; Rotimi et al. 1999), 10p11-10p12 (Hager et al. 1998; Hinney et al. 2007; Price et al. 2001), 11q14-11q24 (Atwood et al. 2002; Hanson et al. 1998; Norman et al. 1997) and 20q11-20q13 (Dong et al. 2003; Hunt et al. 2001; Lee et al. 1999; Norman et al. 1998).

1.5.3.2 Candidate Gene Association Studies on Obesity

Candidate gene approach is a hypothesis-driven approach that tests the association between a trait of interest (either discrete or quantitative) and markers in a predefined candidate gene. Candidate genes might be selected based on gene function, pathway information, evidence from linkage or association studies, and/or animal model studies (El-Sayed Moustafa and Froguel 2013). Unlike linkage analysis, candidate gene approach can be carried out in unrelated individuals.

Hundreds of candidate gene association studies have been performed for obesity or obesity-related traits in the past two decades, and they revealed associations with markers in more than 127 candidate genes. Most of these associations were not consistent across different studies. However, the most widely replicated associations were with variants in *MC4R* (Chambers et al. 2008; Loos et al. 2007; Ma et al. 2004), *PCSK1* (Benzinou et al. 2008; Creemers et al. 2012), brain-derived neurotrophic factor (*BDNF*) (El-Gharbawy et al. 2006; Gray et al. 2006; Ribases et al. 2005; Shugart et al. 2009), peroxisome proliferator-activated receptor γ gene (*PPARγ*) (Deeb et al. 1998; Ristow et al. 1998; Tonjes et al. 2006) and β-adrenergic receptor 3 (*ADRB3*) (Fujisawa et al. 1998; Kurokawa et al. 2008). Most of the associations identified through candidate gene studies did not reach genome-wide significance, which is the association P-value threshold after adjustment for multiple testing (the number of pairwise association tests performed at genome-wide level), using well-established statistical approaches such as Bonferroni correction (Bland and Altman 1995), even in meta-analyses. This suggested that genetic variations at many of the candidate obesity genes have modest effect sizes on obesity.

Recently, Ichimura *et al.* (Ichimura et al. 2012) performed a candidate gene study on the G-protein coupled receptor 120 (*GPR120*) gene, which encodes for an omega-3 fatty acids receptor involved in several biological processes including inflammation, adipogenesis and appetite control. The study identified a rare non-synonymous mutation (p.R270H) in *GPR120* that

confers an increased risk of obesity in European populations. In vitro and in vivo functional studies on this variant demonstrated that it results in the loss of function of the receptor, as a sensor of dietary fat, in a dominant negative fashion which highlighted the importance of environmental sensors as obesity candidates.

1.5.3.3 Genome-Wide Association Studies (GWASs) on Obesity

The international HapMAP project, which aimed to catalogue the common genetic variation (MAF ≥ 5%) in the human genome, served as a foundation for the implementation of genome-wide association studies (GWASs) (International HapMap Consortium 2003). Information on the common genetic variants and the patterns of linkage disequilibrium (LD), which is the non-random association of alleles located in adjacent chromosomal locations, were utilized by companies such as Illumina Inc. (San Diego, USA) and Affymetrix Inc. (Santa Clara, USA) for the development of array-based genome-wide genotyping platforms (International HapMap Consortium 2005; International HapMap Consortium et al. 2007). Such platforms facilitated the genotyping of hundreds of thousands SNPs for large number of samples in a fast, efficient and cost effective manner, making the implementation of GWASs possible (McCarthy et al. 2008). Both the Illumina and Affymetrix platforms have the same genotyping principle that relies on using high-density oligonucleotide arrays and comparing the signal intensity from each of the two SNP alleles for each SNP oligonucleotide to derive a genotyping call, but they apply different chemistries in their workflow.

GWASs are hypothesis-free approach that test for the association between a trait of interest (either discrete or quantitative) and common allelic variants (MAF ≥ 5%) at the genome level. In GWASs, the genotype frequencies at each of these SNPs are tested for association with the trait of interest. Associations are considered significant only when they reach genome-wide levels, therefore a large number of samples is typically needed to enhance the power of association detection (Ku et al. 2010). Association hits that are obtained from GWASs are not necessarily causative but are instead in strong LD with the causal variants (Ku et al. 2010; Walley, Asher, Froguel 2009). A group of variants or genes are usually selected from the initial GWAS for replication and follow-up purposes.

One of the limitations of the GWASs is that they rely on the common disease–common variant (CDCV) hypothesis. This hypothesis argues that the major contributors to the genetic susceptibility to common diseases are common genetic variations, with relatively low effect sizes, and this hypothesis has recently

been challenged (McCarthy et al. 2008). Despite that, GWASs had major contribution towards the identification of genetic variants associated with several common complex diseases.

The first important association to common obesity was mapped to chromosome 16q12 close to the fat mass and obesity-associated (*FTO*) gene (Dina et al. 2007). The association with the *FTO* locus was initially discovered in a GWAS of T2D conducted in 2007 (Frayling et al. 2007). However, the association with T2D was not detected with matched BMI, which suggested that *FTO* effect on T2D is through its effect on BMI (Frayling et al. 2007). Since then, associations with the *FTO* locus were the strongest and most widely replicated with obesity and BMI (Bradfield et al. 2012; Cotsapas et al. 2009; Dorajoo et al. 2012; Heard-Costa et al. 2009; Meyre et al. 2009; Paternoster et al. 2011; Scherag et al. 2010; Scuteri et al. 2007; Tan et al. 2008; Wang et al. 2012).

The risk allele of the *FTO* SNP rs9939609 has the strongest known effect on increased BMI, particularly within European populations (Frayling et al. 2007). Homozygous individuals for the A risk allele show approximately 3kg increase in weight compared to individual homozygous for the normal allele, and have a 1.67-fold increased risk of developing obesity (Frayling et al. 2007). However, this variant accounts for only about 1% of the total heritability estimated for BMI (Frayling et al. 2007).

Currently, the function of *FTO* is largely unknown despite the fact that its effect on obesity has been extensively demonstrated even in animal models (Church et al. 2010; Fischer et al. 2009). Studies on mice have shown that fto is widely expressed in the brain, particularly in the hypothalamus (Gerken et al. 2007), and that its overexpression induces obesity (Church et al. 2010), while the lack of its expression protects from obesity (Fischer et al. 2009; Speakman 2010), suggesting the involvement of *FTO* in the regulation of energy homeostasis and metabolism. However, no direct connection has been made between the obesity-associated variants close to *FTO* and the expression or function of the gene. Smemo *et al.* (Smemo et al. 2014) has recently shown that the real effect of the risk variants close to *FTO* is actually driven by altering the expression of another gene located at a megabase distance from *FTO* and involved in neural development, called Iroquois homeobox 3 (*IRX3*). Such long-range interactions between *FTO* intron and *IRX3* were also conserved in other vertebrate genomes. Moreover, the study showed that *Irx3*-deficient mice have a reduction in body weight of 25 to 30% compared to wild types, demonstrating the role of IRX3 as a regulator of body mass and composition.

Meta-analyses of GWASs, which combine data from multiple independent GWASs, have been conducted in order to achieve large sample size and increase the statistical power for association

detection. The first meta-analysis conducted by the Genetic Investigation of ANthropometric Traits (GIANT) consortium has identified a strong association between SNPs located near *MC4R* and BMI (Loos et al. 2008). In parallel, the SNPs near *MC4R* were reported in association with waist circumference in Indian Asians (Chambers et al. 2008). Associations of *MC4R* variants in relation to obesity have also been reported by other studies (Bradfield et al. 2012; Dorajoo et al. 2012; Heard-Costa et al. 2009; Locke et al. 2015; Meyre et al. 2009; Paternoster et al. 2011).

A larger meta-analysis carried out by the GIANT consortium (Willer et al. 2009) identified six novel obesity loci including transmembrane protein 18 (*TMEM18*), potassium channel tetramerization domain containing 15 (*KCTD15*), glucosamine-6-phosphate deaminase 2 (*GNPDA2*), SH2B adaptor protein 1 (*SH2B1*), mitochondrial carrier homolog 2 (*MTCH2*) and neuronal growth regulator 1 (*NEGR1*).

Further effort by researchers from deCODE genetics (Thorleifsson et al. 2009) confirmed the associations at *TMEM18*, *NEGR1*, *KCTD15* and *SH2B1* with BMI and reported additional associations at SEC16 homolog B (S. Cerevisiae) *(SEC16B)*, brain-derived neurotrophic factor (*BDNF*) and near Ets variant 5 (*ETV5*). Smaller GWASs (Cotsapas et al. 2009; Meyre et al. 2009; Scherag et al. 2010) have confirmed the associations observed at *FTO*, *TMEM18* and *MC4R*, and identified novel associations at v-maf avian musculoaponeurotic fibrosarcoma oncogene homolog (*MAF*), phosphotriesterase related (*PTER*), niemann-pick disease type c1 (*NPC1*), serologically defined colon cancer antigen 8 (*SDCCAG8*), tankyrase trf1-interacting ankyrin-related adp-ribose polymerase (*TNKS*), and methionine sulfoxide reductase a (*MSRA*).

In addition, a larger meta-analysis study on BMI was conducted by the GIANT consortium and it consisted of 46 GWASs and 249,769 individuals (Speliotes et al. 2010). The study revealed 32 SNPs robustly associated with BMI, 14 of these associations were previously reported while 18 were novel. The largest and most recent meta-analysis with BMI carried out by the GIANT consortium included 339,224 individuals from 125 GWASs and has expanded the catalogue of BMI-associated loci to 97 loci (Locke et al. 2015). Of these loci, 41 were previously reported (Berndt et al. 2013; Bradfield et al. 2012; Guo et al. 2013; Kilpelainen et al. 2011; Locke et al. 2015; Monda et al. 2013) while 56 were novel (Locke et al. 2015), as summarized in Table 2. The associations detected provided strong support for the role of the central nervous system in obesity susceptibility as well as other new pathways involved in insulin function, glutamate signalling, adipogenesis and others (Locke et al. 2015).

Interestingly, GWASs have provided links between monogenic and polygenic forms of obesity as they have demonstrated that some common variants located in or near previously recognized monogenic obesity genes modulate BMI and confer risk for common obesity. For example, common variants in *MC4R* have been reported in association with BMI, waist circumference, insulin resistance and other obesity related traits (Bradfield et al. 2012; Chambers et al. 2008; Dorajoo et al. 2012; Heard-Costa et al. 2009; Locke et al. 2015; Loos et al. 2008; Meyre et al. 2009; Paternoster et al. 2011). In addition, Wheeler *et al.* (Wheeler et al. 2013) have identified a common variant in *LEPR* in association with severe early-onset obesity. This suggests that monogenic obesity genes are good candidates for common obesity, and this also highlights the importance of the leptin-melanocortin 4 receptor pathways and the central nervous system in modulating obesity.

The rapid shift towards the application of GWASs had a major contribution in identifying several putative genes and variants that contributes to the pathogenesis of obesity. However, the 97 SNPs robustly associated with BMI all together, for example, accounts only for 2.7% of phenotypic variance of BMI (Locke et al. 2015). The modest effect size of these variants limits their utility in clinical prediction.

In addition, the functional relevance of many of these variants to obesity is unclear since most of them are mapped to intergenic or intronic regions (Drong, Lindgren, McCarthy 2012). Therefore, additional follow-up studies are necessary to fully understand the genes and mechanisms deriving the effect on obesity, as in the case of the *FTO* locus.

SNP	Chr.	Position (bp)	Nearest gene	Risk allele	Risk allele frequency	P-value	Variance explained
rs1558902	16	52,361,075	*FTO*	A	0.413	$7.51X10^{-153}$	0.325%
rs6567160	18	55,980,115	*MC4R*	C	0.238	$3.93X10^{-53}$	0.111%
rs13021737	2	622,348	*TMEM18*	G	0.827	$1.11X10^{-50}$	0.103%
rs10938397	4	44,877,284	*GNPDA2*	G	0.429	$3.21X10^{-38}$	0.079%
rs543874	1	176,156,103	*SEC16B*	G	0.194	$2.62X10^{-35}$	0.072%
rs2207139	6	50,953,449	*TFAP2B*	G	0.174	$4.13X10^{-29}$	0.058%
rs11030104	11	27,641,093	*BDNF*	A	0.789	$5.56X10^{-28}$	0.056%
rs3101336	1	72,523,773	*NEGR1*	C	0.609	$2.66X10^{-26}$	0.053%
rs7138803	12	48,533,735	*BCDIN3D*	A	0.384	$8.15X10^{-24}$	0.047%
rs10182181	2	25,003,800	*ADCY3*	G	0.460	$8.78X10-^{24}$	0.047%
rs3888190	16	28,796,987	*ATP2A1*	A	0.402	$3.14X10^{-23}$	0.046%
rs1516725	3	187,306,698	*ETV5*	C	0.871	$1.89X10^{-22}$	0.045%
rs12446632	16	19,842,890	*GPRC5B*	G	0.866	$1.48X10^{-18}$	0.038%
rs2287019	19	50,894,012	*QPCTL*	C	0.806	$4.59X10^{-18}$	0.041%
rs16951275	15	65,864,222	*MAP2K5*	T	0.781	$1.91X10^{-17}$	0.033%
rs3817334	11	47,607,569	*MTCH2*	T	0.408	$5.15X10^{-17}$	0.033%
rs2112347	5	75,050,998	*POC5*	T	0.633	$6.19X10^{-17}$	0.032%
rs12566985	1	74,774,781	*FPGT-TNNI3K*	G	0.442	$3.28X10^{-15}$	0.029%
rs3810291	19	52,260,843	*ZC3H4*	A	0.665	$4.81X10^{-15}$	0.036%
rs7141420	14	78,969,207	*NRXN3*	T	0.522	$1.23X10^{-14}$	0.028%
rs13078960	3	85,890,280	*CADM2*	G	0.196	$1.74X10^{-14}$	0.028%
rs10968576	9	28,404,339	*LINGO2*	G	0.317	$6.61X10^{-14}$	0.027%
rs17024393	1	109,956,211	*GNAT2*	C	0.037	$7.03X10^{-14}$	0.033%
rs657452	1	49,362,434	*AGBL4*	A	0.392	$5.48X10^{-13}$	0.025%
rs12429545	13	53,000,207	*OLFM4*	A	0.133	$1.09X10^{-12}$	0.026%
rs12286929	11	114,527,614	*CADM1*	G	0.522	$1.31X10^{-12}$	0.023%
rs13107325	4	103,407,732	*SLC39A8*	T	0.071	$1.83X10^{-12}$	0.030%
rs11165643	1	96,696,685	*PTBP2*	T	0.582	$2.07X10^{-12}$	0.023%
rs7903146	10	114,748,339	*TCF7L2*	C	0.713	$1.11X10^{-11}$	0.022%
rs10132280	14	24,998,019	*STXBP6*	C	0.682	$1.14X10^{-11}$	0.023%
rs17405819	8	76,969,139	*HNF4G*	T	0.700	$2.07X10^{-11}$	0.021%
rs6091540	20	50,521,269	*ZFP64*	C	0.719	$2.15X10^{-11}$	0.014%
rs1016287	2	59,159,129	*LINC01122*	T	0.286	$2.25X10^{-11}$	0.021%
rs4256980	11	8,630,515	*TRIM66*	G	0.644	$2.90X10^{-11}$	0.020%
rs17094222	10	102,385,430	*HIF1AN*	C	0.209	$5.94X10^{-11}$	0.021%
rs12401738	1	78,219,349	*FUBP1*	A	0.353	$1.15X10^{-10}$	0.020%
rs7599312	2	213,121,476	*ERBB4*	G	0.724	$1.17X10^{-10}$	0.019%
rs2365389	3	61,211,502	*FHIT*	C	0.581	$1.63X10^{-10}$	0.019%
rs205262	6	34,671,142	*C6orf106*	G	0.270	$1.75X10^{-10}$	0.019%

SNP	Chr.	Position (bp)	*Nearest gene*	Risk allele	Risk allele frequency	P-value	Variance explained
rs2820292	1	200,050,910	*NAV1*	C	0.557	$1.83X10^{-10}$	0.019%
rs12885454	14	28,806,589	*PRKD1*	C	0.639	$1.94X10^{-10}$	0.020%
rs9641123	7	93,035,668	*CALCR*	C	0.430	$2.08X10^{-10}$	0.018%
rs12016871	13	26,915,782	*MTIF3*	T	0.203	$2.29X10^{-10}$	0.029%
rs16851483	3	142,758,126	*RASA2*	T	0.066	$3.55X10^{-10}$	0.029%
rs1167827	7	75,001,105	*HIP1*	G	0.559	$6.33X10^{-10}$	0.020%
rs758747	16	3,567,359	*NLRC3*	T	0.266	$7.47X10^{-10}$	0.020%
rs1928295	9	119,418,304	*TLR4*	T	0.549	$7.91X10^{-10}$	0.018%
rs9925964	16	31,037,396	*KAT8*	A	0.620	$8.11X10^{-10}$	0.017%
rs11126666	2	26,782,315	*KCNK3*	A	0.278	$1.33X10^{-09}$	0.017%
rs2650492	16	28,240,912	*SBK1*	A	0.304	$1.92X10^{-09}$	0.018%
rs6804842	3	25,081,441	*RARB*	G	0.579	2.48X10-^{09}	0.017%
rs12940622	17	76,230,166	*RPTOR*	G	0.570	$2.49X10^{-09}$	0.016%
rs7164727	15	70,881,044	*LOC100287559*	T	0.688	$3.92X10^{-09}$	0.014%
rs11847697	14	29,584,863	*PRKD1*	T	0.043	$3.99X10^{-09}$	0.019%
rs4740619	9	15,624,326	*C9orf93*	T	0.544	$4.56X10^{-09}$	0.016%
rs492400	2	219,057,996	*USP37*	C	0.425	$6.78X10^{-09}$	0.012%
rs13191362	6	162,953,340	*PARK2*	A	0.876	$7.34X10^{-09}$	0.016%
rs3736485	15	49,535,902	*DMXL2*	A	0.455	$7.41X10^{-09}$	0.015%
rs17001654	4	77,348,592	*SCARB2*	G	0.154	$7.76X10^{-09}$	0.024%
rs11191560	10	104,859,028	*NT5C2*	C	0.090	$8.45X10^{-09}$	0.015%
rs2080454	16	47,620,091	*CBLN1*	C	0.400	$8.60X10^{-09}$	0.014%
rs7715256	5	153,518,086	*GALNT10*	G	0.420	$8.85X10^{-09}$	0.013%
rs2176040	2	226,801,046	*LOC646736*	A	0.370	$9.99X10^{-09}$	0.009%
rs1528435	2	181,259,207	*UBE2E3*	T	0.632	$1.20X10^{-08}$	0.015%
rs2075650	19	50,087,459	*TOMM40*	A	0.850	$1.25X10^{-08}$	0.017%
rs1000940	17	5,223,976	*RABEP1*	G	0.316	$1.28X10^{-08}$	0.016%
rs2033529	6	40,456,631	*TDRG1*	G	0.288	$1.39X10^{-08}$	0.015%
rs11583200	1	50,332,407	*ELAVL4*	C	0.392	$1.48X10^{-08}$	0.015%
rs7239883	18	38,401,669	*LOC284260*	G	0.388	$1.51X10^{-08}$	0.013%
rs2836754	21	39,213,610	*ETS2*	C	0.620	$1.61X10^{-08}$	0.013%
rs9400239	6	109,084,356	*FOXO3*	C	0.693	$1.61X10^{-08}$	0.015%
rs10733682	9	128,500,735	*LMX1B*	A	0.474	$1.83X10^{-08}$	0.015%
rs11688816	2	62,906,552	*EHBP1*	G	0.525	$1.89X10^{-08}$	0.015%
rs11057405	12	121,347,850	*CLIP1*	G	0.902	$2.02X10^{-08}$	0.017%
rs9914578	17	1,951,886	*SMG6*	G	0.209	$2.07X10^{-08}$	0.013%
rs977747	1	47,457,264	*TAL1*	T	0.392	$2.18X10^{-08}$	0.013%
rs2121279	2	142,759,755	*LRP1B*	T	0.147	$2.31X10^{-08}$	0.015%
rs29941	19	39,001,372	*KCTD15*	G	0.671	$2.41X10^{-08}$	0.015%
rs11727676	4	145,878,514	*HHIP*	T	0.912	$2.55X10^{-08}$	0.021%

SNP	Chr.	Position (bp)	Nearest gene	Risk allele	Risk allele frequency	P-value	Variance explained
rs3849570	3	81,874,802	*GBE1*	A	0.357	$2.60X10^{-08}$	0.016%
rs9374842	6	120,227,364	*LOC285762*	T	0.744	$2.67X10^{-08}$	0.013%
rs6477694	9	110,972,163	*EPB41L4B*	C	0.364	$2.67X10^{-08}$	0.014%
rs4787491	16	29,922,838	*INO80E*	G	0.510	$2.70X10^{-08}$	0.013%
rs1441264	13	78,478,920	*MIR548A2*	A	0.604	$2.96X10^{-08}$	0.015%
rs7899106	10	87,400,884	*GRID1*	G	0.053	$2.96X10^{-08}$	0.015%
rs2176598	11	43,820,854	*HSD17B12*	T	0.249	$2.97X10^{-08}$	0.015%
rs2245368	7	76,446,079	*PMS2L11*	C	0.178	$3.19X10^{-08}$	0.030%
rs17203016	2	207,963,763	*CREB1*	G	0.196	$3.41X10^{-08}$	0.014%
rs17724992	19	18,315,825	*PGPEP1*	A	0.744	$3.42X10^{-08}$	0.014%
rs7243357	18	55,034,299	*GRP*	T	0.817	$3.86X10^{-08}$	0.014%
rs16907751	8	81,538,012	*ZBTB10*	C	0.918	$3.89X10^{-08}$	0.019%
rs1808579	18	19,358,886	*C18orf8*	C	0.535	$4.17X10^{-08}$	0.014%
rs13201877	6	137,717,234	*IFNGR1*	G	0.141	$4.29X10^{-08}$	0.013%
rs2033732	8	85,242,264	*RALYL*	C	0.747	$4.89X10^{-08}$	0.014%
rs9540493	13	65,103,705	*MIR548X2*	A	0.452	$4.97X10^{-08}$	0.015%
rs1460676	2	164,275,935	*FIGN*	C	0.172	$4.98X10^{-08}$	0.011%
rs6465468	7	95,007,450	*ASB4*	T	0.307	$4.98X10^{-08}$	0.012%

Table 2. Summary of BMI-associated loci identified through the latest GIANT meta-analysis of GWASs. Coordinates are given according to the GRCh36/hg18 genome assembly, allele coding based on the positive strand and P-value represent the most significant P-value obtained.

Many scientists expected that the heritability of obesity and other common diseases would be explained by common genetic variants, as described by the common disease-common variant (CDCV) hypothesis. However, GWASs proved that the situation is more complex. Most common variants individually or in combination confer relatively modest effect sizes and explain only a small proportion of the heritability. The 97 BMI variants detected through the largest meta-analysis of GWASs, for example, can explain only 2.7% of the phenotypic variance of BMI (Locke et al. 2015). Such mismatch between the high heritability estimates for BMI and the small proportion of variation explained through the variants identified in GWASs is descried by the phenomenon of "missing heritability" which is also seen in other common diseases (Manolio et al. 2009).

Various sources have been proposed for the missing heritability. A recent study suggested that the accumulated effect of many hundreds of associated loci, including weaker associations that do not reach genome-wide significance, may explain about 20% of the total variance in BMI (Yang et al. 2012). Other

explanations include the interaction of risk variants with environmental conditions, which are not taken into account in GWASs. For instance, the risk effect of the *FTO* variant is reduced in physically active people, and this has only become clear with the follow-up studies on the *FTO* variant (Willyard 2014). Other sources of genetic variation including rare SNPs (MAF <1%) and copy number variations (CNVs), and epigenetic variations such as DNA methylation are also believed to play a role in obesity susceptibility, as will be discussed in Sections 1.7, 1.8 and 1.9 respectively.

1.7 The Role of Rare Mutations in the Development of Obesity

The missing heritability of complex diseases is largely believed to be attributable to moderately highly penetrant rare mutations (MAF< 1%) of relatively recent origin, that have not been identified through GWASs (Bodmer and Bonilla 2008; Pritchard 2001; Schork et al. 2009). This hypothesis is known as the common disease-rare variant (CDRV) hypothesis. Under this hypothesis, the severity of the disease in the subjects carrying the risk disease variants may be modified by other genetic loci or by environmental factors (Bodmer and Bonilla 2008; Steinberg and Adewoye 2006). In relation to obesity, it has been shown that rare variants can drive monogenic obesity cases, as discussed in Section 1.5.1, and they are also implicated in more common forms of obesity, as will be discussed in Section 1.7.2.

1.7.1 Detection Methods of Rare Mutations

Rare mutations are not detected in GWASs since SNP genotyping arrays are usually designed to capture common SNPs with MAF≥ 5% in the general population (Drong, Lindgren, McCarthy 2012). The detection and investigation of the role of rare variants is possible using sequencing technologies. Recent advances in next generation sequencing (NGS) technologies are allowing researchers to extend their genome-wide analyses beyond common SNPs to assess the contribution of rare variants in disease predisposition in a fast way. However, population-based approaches to mapping rare variants for complex disease from NGS data require a large number of samples (thousands or many thousands) to provide sufficient statistical power for rare variants detection (Cirulli and Goldstein 2010). The reason for that is the dilution of the frequency of the rare variants and their effect sizes at the population level as a result of the high locus and allelic heterogeneity (Cirulli and Goldstein 2010; Heinzen et al. 2012; Need et al. 2012)

Until the cost of whole genome sequencing (WGS) drops enough for conducting population-based whole genome association studies with large number of samples, study designs based on extreme phenotypes and/or families can be exploited (Cirulli and Goldstein 2010). This is because these study designs offer an

opportunity for increasing the frequency of rare variants (and their effects) in the study and such enrichment results in improved power for variants detection.

Extreme phenotype study designs involves the sequencing of a small number of individuals at the ends of the extremes of the phenotype distribution from the general population, and the identification of rare variants that are enriched in frequency in cases versus controls (Cirulli and Goldstein 2010). Candidate rare disease variants are then selected based on their potential functional impact and typed in larger case-control population cohorts for confirmation.

Family-based study designs involve the sequencing of multiple affected individuals in families, and the screening of the list of variants that are shared between affected individuals to identify candidate rare disease variants. This enrichment for rare variants between affected individuals in families results in increased locus specific heritability and improved power for detection of causal variants. Candidate rare variants are usually selected based on features like the expected variant effect, gene function and co-segregation evidence with diseases in families (Cirulli and Goldstein 2010). Co-segregation evidence of a candidate variant with disease in other families or association evidence in large case-control cohorts is usually required for successful variant discovery (Cirulli and Goldstein 2010).

In addition, family-based designs offer the possibility of performing linkage analysis and determining the genetic basis of the linkage signals using sequencing. This approach could facilitate the identification of genes involved in obesity that are otherwise undetectable using case-control designs (Cirulli and Goldstein 2010). For example, Sobreira *et al.* (Sobreira et al. 2010) applied this strategy to metachondromatosis, a dominantly inherited Mendelian disease with incomplete penetrance, and successfully identified the causing variant, which was located within the second-highest LOD score of linkage regions.

In this study, I have applied a similar strategy to obesity in order to identify rare obesity variants in families with multiple obese individuals from Qatar. My strategy involved the identification of candidate obesity genomic regions through performing linkage analysis and run of homozygosity (ROH) analyses, using high-density SNP genotyping data. ROH regions are long stretches of consecutive homozygous genotypes that probably reflect identity by descent segments and are likely to be enriched for deleterious variants (Szpiech et al. 2013). Then, I have screened the candidate obesity regions and genes for harbouring potential rare obesity variants using WGS data, and attempted to replicate the identified variants in a Qatari population cohort, as will be discussed in details in Chapter 3.

1.7.2 Rare Mutations and Obesity

There is now sufficient evidence that both common and rare variants contribute to common complex diseases. Rare variants of strong effect have been reported to influence the susceptibility to several complex diseases including infectious diseases (Khor et al. 2007), type I diabetes (Nejentsev et al. 2009) as well as obesity (Blakemore et al. 2009; Creemers et al. 2012; Ichimura et al. 2012; Pearce et al. 2013).

For example, Pearce *et al.* (Pearce et al. 2013) have recently explored the kinase suppressor of ras 2 (*KSR2*) gene, which encodes an intracellular scaffolding protein involved in multiple regulatory pathways, as a candidate obesity gene in humans. They performed whole exome sequencing on 2,101 individuals with severe early-onset obesity and 1,536 controls, and identified multiple rare variants in *KSR2* that disrupt the signalling of the protein through the Raf-MEK-ERK pathway. Carriers of the mutation exhibited hyperphagia during childhood, low heart rate, low basal metabolic rate as well as severe insulin resistance. The study underscored the role of KSR2 as a regulator of energy intake and expenditure in humans and, in general, the role of rare variants in common obesity.

Another interesting example that illustrates the wide spectrum of variants contributing to obesity is the *PCSK1* gene. As discussed before, rare loss-of-function mutations in *PCSK1* cause recessive (Farooqi et al. 2007; Jackson et al. 1997) and dominant (Philippe et al. 2015) monogenic early-onset obesity. Moreover, common coding variants in *PCSK1* were found to mildly increase the risk for common obesity (Benzinou et al. 2008; Villalobos-Comparan et al. 2012). In addition to that, rare heterozygous *PCSK1* mutations causing partial prohormone convertase 1 deficiency were reported to confer an increased risk for obesity (Creemers et al. 2012; Pickett et al. 2013).

Given the role of rare variants in obesity predisposition, I have searched for potential obesity variants in my Qatari families' dataset. To achieve this goal, I have combined linkage and ROH evidence in these families to focus my search for rare obesity variants to a reduced number of genomic regions and thus variants to those that are most promising, as will be discussed in details in Chapter 3.

1.8 The Role of Copy Number Variations (CNVs) in the Development of Obesity

Copy number variations are DNA segments of 1kb or more that exist in variable numbers within the genome and they usually refer to deletions and duplications. CNVs have been identified in humans as well as in other organisms (Shastry 2009), and they have emerged as a notable source of genetic variation that is present in high abundance in the human genome, including coding and non-coding

regions (Beckmann, Sharp, Antonarakis 2008). Yeo *et al.* have shown that CNV with a size of about 500kb are present in 5-10% of the population (Yoo et al. 2005). As many CNVs encompass genes and/or regulatory regions, they may result in alteration of gene dosage and/or disruption of gene structure and/or regulation (Stranger et al. 2007). Therefore, CNVs may account for significant normal phenotypic variation within a species, susceptibility or resistance to disease, variation in drug response as well as the evolution of genomes (Shastry 2009; Stranger et al. 2007).

1.8.1 Mechanisms of CNV Formation

CNV structure can be simple with stable regions flanked by consistent boundaries, or more complex, containing multiple CNVs with overlapping boundaries, called copy number variant regions (CNVRs). Although the mechanisms of CNV formation are not fully elucidated, these different types of CNVs may arise through different processes (de Smith et al. 2008).

For example, about 50% of reported CNVs are associated with regions of segmental duplication (Cooper, Nickerson, Eichler 2007; Redon et al. 2006; Sharp et al. 2005). These are genomic regions of > 1kb that have sequence similarity of >90% with another genomic region (Conrad and Hurles 2007). The formation of this type of CNVs is likely to be mediated by a process known as non-allelic homologous recombination (NAHR), whereby crossing over occur between non-allelic regions of high sequence similarity, such as segmental duplications (Conrad and Hurles 2007) resulting in copy number changes. Complex CNVs have a greater enrichment of segmental duplications compared to more simple CNVs, therefore they are likely to occur via NAHR (Redon et al. 2006).

Other non-homology-based mechanisms have been proposed to explain the formation of other CNVs. One of these mechanisms is the non-homologous end joining (NHEJ), which is driven by errors in DNA break repair mechanisms (Korbel et al. 2007; Perry et al. 2008). A recent study of the breakpoint junctions of CNVs suggests that the majority of CNVs were formed through this mechanism (Perry et al. 2008). A recent analysis on the origin of common deletion variants in the genome have suggested a role of Alu repeats in CNV formation (de Smith et al. 2008). The study reported an enrichment of Alu repeats at the break points of this group of CNVs. These repeats are prone to single or double strand breakage that could result in NHEJ and hence CNV formation. The role of Alu repeats in CNV formation is likely to be greater in gene CNVs than non-gene CNVs since Alu repeats correlate with CG rich genomic regions (de Smith et al. 2008; Lander et al. 2001).

It is apparent that the distribution of CNVs in the genome is not random, with the existence of hot and cold CNV spots (Cooper, Nickerson, Eichler 2007). In particular, CNVs are more frequent towards centromeres and telomeres (Nguyen, Webber, Ponting 2006). This variation might be due to differences in the type and number of nearby repeat elements, such as Alu repeats and segmental duplications, in the genome. The nature of the sequence at a particular region could also influence the type of CNVs they may overlap. For instance, complex CNVRs are more likely to exist in highly repetitive regions, while ancient stable CNVs that are inherited with conserved boundaries are likely to present in less repetitive regions (de Smith et al. 2008).

Family studies have shown that CNVs may vary in their inheritance. The majority of CNVs are inherited from parents but CNVs may also occur *de novo* during meiosis (Locke et al. 2006). In addition, recent studies have reported CNVs between monozygotic twins (Bruder et al. 2008) and also between tissues within the same individual (Piotrowski et al. 2008) suggesting that not all CNVs arise during meiosis. The mechanisms underlying the formation of somatic CNVs are unknown, which adds a further complexity to the CNV field.

1.8.2 Methods of CNV Detection

Several technologies can be used to study copy number variations. One of the earliest methods was the comparative genomic hybridisation (CGH) technology, developed in 1992 (Kallioniemi et al. 1992). This method compares the fluorescence ratio between a test and a reference DNA sample, each labelled with a particular fluorescent dye, hybridised to a metaphase chromosome spread from a healthy individual. The low resolution of the conventional CGH technology (5-10Mb) was a major disadvantage, therefore the technology was further developed in 2004 to replace the metaphase spread with microarrays (Carter 2007). The new wave of technology was called array comparative genomic hybridisation (aCGH), and they provided greater resolution (down to ~30kb) and less labour-intensive procedure for the detection of CNVs (Carter 2007). One of the popular arrays was the bacterial artificial chromosome (BAC) array CGH, in which DNA obtained from BAC clones was spotted on the array (Ren et al. 2005b). In addition, other types of arrays that implement oligonucleotide probes were developed such as Representational Oligonucleotide Microarray Analysis (ROMA) (Lucito et al. 2003). In 2006, high-resolution CGH (HR-CGH) arrays were developed that can detect CNVs at a resolution of 200bp (Urban et al. 2006).

In 2004, two striking studies on CNVs were published. The first study used BAC array CGH and identified 255 CNVs in 39 subjects (Iafrate et al. 2004), while the other study used ROMA and found 76 CNVs in 20 individuals (Sebat et al. 2004). These two studies have shown that CNVs are more common in the genome than previously speculated, giving further appreciation of CNVs as a frequent and widespread source of genetic variation that may have phenotypic effect. Iafrate *et al.* also developed the Database of Genomic Variants (DGV) to catalogue the CNVs in the human genome (Iafrate et al. 2004).

A number of CNV studies have been published after that, using a variety of CNV detection techniques (Conrad et al. 2006; Hinds et al. 2006; Korbel et al. 2007; McCarroll and Altshuler 2007; Mills et al. 2006; Redon et al. 2006; Tuzun et al. 2005; Wong et al. 2007). These studies have revealed a wealth of CNVs in the genome, many of which overlap with multiple genes. Some of these studies (Conrad et al. 2006; Hinds et al. 2006; Korbel et al. 2007; McCarroll and Altshuler 2007; Mills et al. 2006; Redon et al. 2006; Tuzun et al. 2005; Wong et al. 2007) were carried out using high-resolution genome-wide SNP genotyping arrays.

The principle behind CNV prediction from high-density SNP genotyping platforms is based on the signal intensity measures. For each SNP probe, the total signal intensity from each of the two alleles is normalised and processed to calculate Log R Ratio (LRR) and B Allele Frequency (BAF) (Peiffer et al. 2006). LRR is a measure proportional to the total signal intensity and it reflects the copy number at a particular position while the BAF is a measure of the relative contribution of the B allele to the total signal intensity and it reflects the allelic composition at each SNP. Analysis of these measures for the probes across the genome is used to identify CNVs and to determine their copy number states (Peiffer et al. 2006).

One of the advantages of using the SNP arrays for CNV detection is the ability to utilize the genotyping data generated for other purposes such as linkage analyses and GWASs. In addition, there are many well-established CNV detection algorithms from SNP array data including PennCNV (Wang et al. 2007), FamCNV (Eleftherohorinou et al. 2011), CNVHap (Coin et al. 2010) and CNV Partition (Peiffer et al. 2006).

Following the development of next generation sequencing technologies, CNV analyses can be performed today at the nucleotide level using whole exome sequencing (WES) and whole genome sequencing (WGS) technologies. Several computational approached have been developed for CNV prediction from these sequencing based-technologies such as CNV-seq (Xie and Tammi 2009), SegSeq (Chiang et al. 2009) and readDepth (Miller et al. 2011), and they are usually based on short read information. Due to

the relatively young age of these approaches, efforts have also been made to compare their performance for different study designs and purposes (Duan et al. 2013).

1.8.3 Copy Number Variations and Obesity

One of the key effects of CNVs is that they can result in alteration of gene expression patterns. It has been estimated that at least 17.7% of the heritable variation in gene expression is due to CNVs (Stranger et al. 2007). This suggests that CNVs may account for a substantial proportion of phenotypic variations including susceptibility to diseases. CNVs have been found to be involved in a broad range of human diseases. This includes genetic disorders such as Charcot-Marie-Tooth disease and DiGeorge syndrome (Baldini 2005; Lupski et al. 1991), in which large deletions and duplications result in distinct syndromic phenotypes, as well as other non-syndromic disorders such as schizophrenia (Stefansson et al. 2008) and autism (Kumar et al. 2008).

In relation to obesity, several common (observed in more than 5% of the population) bialleleic CNVs have been reported in association with BMI and obesity (Jarick et al. 2011; Sha et al. 2009; Speliotes et al. 2010; Willer et al. 2009). Similar to common SNPs, the identified common CNVs had small effect sizes on obesity risks. Recently, Falchi *et al.* identified a common multi-allelic CNV at the salivary amylase (*AMY1*) locus associated with BMI in European and Asian adults, through a CNV association study of gene expression levels in adipose tissue (Falchi et al. 2014). This study showed that low copy number of *AMY1* results in decreased salivary amylase levels and increased risk of obesity, with relatively high odd ratio of about 1.19 per copy of *AMY1.* The association was even stronger than the associations observed for *FTO* SNPs with BMI, and was also recently replicated in Mexican children (Mejia-Benitez et al. 2015). Salivary amylase functions as a key catalytic enzyme for the initiation of carbohydrate digestion in the oral cavity and interestingly, earlier studies reported low blood amylase levels in obese humans (Kondo et al. 1988) and rats (Schneeman, Inman, Stern 1983). The study underscores the importance to *AMY1* CNV locus in particular and CNVs in general in relation to obesity.

On the other hand, rare CNVs (present in less than 1% of the population) have been found to cause highly penetrant forms of obesity (Bochukova et al. 2010; Glessner et al. 2010; Walters et al. 2013; Wang et al. 2010a; Yu et al. 2011). The most well-known CNVs in relation to obesity are at the 16p11.2 locus. The short arm of chromosome 16 is enriched for large segmental duplication that can facilitate the occurrence of the nonallelic homologous recombination (NAHR), and hence the formation of recurrent CNVs at this particular region. Bochukova *et al.* (Bochukova et al. 2010) performed CNV

analysis on 300 Caucasian children with early-onset obesity (half of them had also developmental delay) and reported a rare heterozygous deletion of 220kb at 16p11.2 (28.7-28.9 Mb, hg18; NCBI build 36) associated with obesity.

The minimal deleted region encompasses nine annotated genes; some of them involved in neurological diseases and immunity. Interestingly, the region contains SH2B adapter protein 1 (*SH2B1)*, which encodes an adaptor protein involved in leptin and insulin signalling (Jamshidi et al. 2007; Ren et al. 2007). Knockout of sh2b1 in mice results in extreme obesity, hyperphagia and insulin resistance (Ren et al. 2005a), and SNPs within this gene have also been associated with BMI (Speliotes et al. 2010). These findings suggest a potential involvement of this particular gene in causing the obesity phenotype in subjects carrying the deletion.

Our group (Walters et al. 2013) identified a separate deletion on 16p11.2 (29.5-30.1Mb.hg18; NCBI build 36) associated with a highly penetrant form of obesity. The deletion was initially identified in 2.9% (9 out of 312) of the study subjects who had malformations and/or developmental delay (DD) in addition to obesity, and was also investigated in other independent cohorts for obesity and/or DD. The data from the obesity cohorts was pooled with data obtained from general population cohort (of 11,856 subjects) and a case-control association analysis was carried out, giving association at genome-wide significance.

This deletion confers about 43-fold increased risk of severe obesity in adults and interestingly, Jacquemont *et al.* has recently found that reciprocal duplication at the 16p11.2 confers an increased risk (about 8.3 fold) for being underweight, showing that gene dosage at this particular locus have opposing effect on BMI (Jacquemont et al. 2011). The same study carried out expression analyses in carriers of the 16p11.2 deletions and duplications and found that the expression levels of 27 genes in the region (out of 29 genes) are positively correlated with gene dosage.

The 16p11.2 deletions demonstrate the potential importance of rare variants in the susceptibility to common diseases like obesity. In addition, the large effect size of some of the rare CNVs, including the 16p11.2 deletions, makes them a useful platform for clinical disease prediction similar to monogenic obesity variants (El-Sayed Moustafa and Froguel 2013). Conducting functional studies is crucial to fully characterize the functionally relevant loci within CNVs in relevance to the disease phenotype, particularly, large CNVs that may encompass multiple potential disease loci.

Given the significant role of both common and rare CNVs in relevance to obesity, I have decided to investigate obesity CNVs in my Qatari families, using the intensity data from the high-density SNP

genotyping arrays. I have carried out a genome-wide CNV association analysis with BMI to identify common obesity CNVs, and I have also searched for potential large rare obesity CNVs in my dataset, as will be discussed in details in Chapter 4.

1.9 The Role of DNA Methylation Patterns in the Development of Obesity

Epigenetics is the study of changes in the genome that do not alter the primary DNA sequence itself. The major epigenetic modifications are DNA methylation, histone methylation and chromatin modifications and they are believed to provide a biological link between an individual's environmental exposures and phenotypes (Drong, Lindgren, McCarthy 2012). DNA methylation was the first epigenetic modification to be discovered (HOTCHKISS 1948) and it is the most studied so far. Historically, DNA methylation was recognized as DNA restriction-modification mechanism in bacteria that protects host DNA from being cleaved by their cognate restriction enzymes that digests unmodified foreign DNAs like viral DNAs (Meselson, Yuan, Heywood 1972).

In humans, the predominant form of DNA methylation occurs on the 5^{th} carbon of the cytosine in the cytosine–guanine nucleotide pair (CpG) sites along the DNA chain (Liu et al. 2010). However, CpH methylation (where H=C/A/T) also exists mainly in embryonic stem cells (Ramsahoye et al. 2000). The total number of CpG sites in the human genome is estimated to be around 28 million CpGs present in a nonrandom distribution throughout the genome, and 60–80% of them are generally methylated (Smith and Meissner 2013). Less than 10% of CpGs occur in CG-rich regions, known as CpG islands. These CpG islands accounts for most of the nonmethylated CpG sites and they are prevalent at transcription start sites of housekeeping genes as well as developmental regulator genes (Deaton and Bird 2011). On the other hand, CpG sites that exist outside CpG islands are largely methylated (Deaton and Bird 2011).

In 1975, it has been proposed that DNA methylation might be responsible for maintaining expression patterns through mitosis (Holliday and Pugh 1975; Riggs 1975). Since then, DNA methylation has gained a lot of attention and it became apparent that hypomethylated DNA regions are involved in active transcription, while the hypermethylated DNA regions tend to have an opposing effect on gene expression (Jaenisch and Bird 2003; Razin and Cedar 1977; Stein et al. 1983; Urnov and Wolffe 2001; Wolffe and Matzke 1999). For example, CpG regions at the promoter regions of tissue-specific genes are usually highly unmethylated in their tissue of expression while they are highly methylated in other tissues where they are not expressed (Jaenisch and Bird 2003). Similarly, housekeeping genes were

found to have constitutively unmethylated CpG islands at their promoters regardless of tissue types (Smith and Meissner 2013; Stein et al. 1983).

In some instances, the effect of DNA methylation on gene expression is mediated through blocking the binding sites of transcription factors. However, since most transcription factors do not have CpGs in their binding sites, the effect of DNA methylation is believed to be largely mediated through affecting gene accessibility at the level of chromatin structure, where methylated CpGs result in closely packed chromatin and unmethylated CpGs result in more open chromatin structure (Razin and Cedar 1977). For example, the presences of a methyl groups can enhance the binding of methyl-binding proteins such as MeCP2 and MBD2 (Klose and Bird 2006) and/or prevent the binding of other types of proteins such as the chromatin proteins CTCF and Cfp1 (Thomson et al. 2010). Given its critical effect on gene expression, DNA methylation is considered a key mechanism that contributes to normal development, phenotypic variations as well as susceptibility to diseases.

1.9.1 Formation Mechanisms of DNA Methylation Patterns

DNA methylation is mainly mediated by two types of methyltransferases: (i) DNMT1, which is responsible for methylation after every cellular DNA replication cycle to maintain DNA methylation patterns, and (ii) DNMT3a and DNMT3b, which are responsible of *de novo* methylation early in development (Thompson, Fazzari, Greally 2010).

Although the precise mechanism of DNA methylation is not fully elucidated, studies have shown that DNA methylation modifications are typically established during early developmental stages. In the gametes and preimplantation embryo, almost all methylation modifications are erased, but imprinting regions preserve their methylation on one allele (and also throughout development). At the time of implantation, the entire genome gets methylated but CpG islands stay unmethylated. Postimplantation, pluripotency genes undergo *de novo* methylation. The postimplantation development is also characterized by changes in DNA methylation patterns that are tissue specific, where tissue-specific genes get unmethylated in their tissues of expression. This phenomenon of acquiring two or more different epigenetic states at a particular genomic locus in different cells or tissues within the same organism is known as epigenetic mosaicism and it adds a further complexity to the field of DNA methylation as it can contribute to phenotypic variation among people (Stoger 2008). Changes in methylation patterns may also arise in reaction to environmental exposures or as a function of aging, resulting in global and/or local changes of methylation patterns (Ng et al. 2010; Radford et al. 2014).

The process of erasure and reestablishment of DNA methylation modification in early developmental stages is known as “DNA methylation programing”. A recent study has shown that methylation patterns at some genomic regions can escape the reprogramming process and thus can be inherited through generations (Hackett et al. 2013).

There are two types of inheritance of epigenetic state: i) the genetic inheritance where the individual’s underlying genetic sequence affects the epigenetic state, for example, as in the case of the occurrence of genetic variations within a CpG locus which disrupts the methylation site, and ii) the epigenetic inheritance which refers to the sequence-independent transmission of epigenetic mark that may happen through escaping the methylation reprogramming events (McRae et al. 2014).

Bell *et al.* (Bell et al. 2012) investigated the heritability of methylation in twins using methylation profiles derived from whole blood-DNA. Monozygotic twins have a significantly higher genome-wide average correlation in methylation levels compared to dizygotic twins. This supported the idea of the heritability of methylation patterns and the heritability estimate obtained was about 18%. However, a recent study conducted by McRae *et al.* (McRae et al. 2014) has investigated the heritability of DNA methylation in whole-blood DNA using multigenerational families, which allowed the conductance of both within and across generations’ comparisons of heritability patterns. The study showed that the heritability of DNA methylation is actually largely driven by the underlying genetic sequence (i.e. type i of methylation inheritance).

1.9.2 Detection Methods of DNA Methylation Patterns

Similar to other research fields, the development in the field of epigenetics is technology-driven. Technologies that are based on repeat interspersed regions, namely Alu and Long Interspersed Nucleotide Element 1 (LINE-1), like MethyLight (Weisenberger et al. 2005), or based on restriction enzymes like the Luminometric assay (LUMA) (Karimi et al. 2006) were able to estimate DNA methylation at the global level. For site-specific information on DNA methylation, several methods including Methylation-specific PCR (MS PCR) (Herman et al. 1996) and Pyrosequencing (Colella et al. 2003) were developed. These methods rely on bisulfite treatment of DNA, which converts the vast majority of unmethylated cytosines to uraciles, whereas the methylated cytosines remain protected from the conversion, and are usually used to investigate the methylation status of candidate genes or regions.

The genome-wide assessment of DNA methylation states became feasible with the development of two groups of technologies. The first group identifies enriched DNA methylation sites through utilizing methylation-sensitive restriction enzymes such as Methylated DNA Binding Domain sequencing (MBD-seq) technology (Lan et al. 2011), or through immunoprecipitation with antibodies against the methylated cytosines as in Methylated DNA Immunoprecipitation (MeDIP) (Mohn et al. 2009). The second group relies on using bisulfite-converted DNA such as Bisulfite Conversion Followed by Sequencing (BS–seq) (Lister et al. 2009) and Reduced Representation Bisulfite Sequencing (RRBS-seq) (Gu et al. 2010). In both groups of technologies, microarray and sequencing approaches have been implemented (Ladd-Acosta et al. 2010).

Genome-wide sequencing-based approaches of DNA methylation can provide comprehensive information of the methylation state of every cytosine in the entire genome, including repetitive sequences, in a sequence specific way. However, the high cost of these approaches is still limiting their utility in epigenetic research. On the other hand, microarray-based profiling approaches allow the analysis of large number of samples at an affordable cost. Illumina Infinium HumanMethylation450 BeadChip, which is designed to target more than 450,000 CpG sits across the genome, is currently a common option for genome-wide DNA methylation studies (Bibikova et al. 2009; Bibikova et al. 2011).

Despite the advancement in the methylation profiling technologies, designing appropriate DNA methylation studies remains challenging due to the instability of the epigenetic marks over time and also between different cell and tissue types, and in response to environmental changes (Sandholt, Hansen, Pedersen 2012). DNA methylation profiles from whole blood for example does not necessarily reflects the profiles in more disease relevant tissues such as adipose tissues in the case obesity, and also different cellular subtypes within whole blood for example can have distinct methylation profiles (Houseman et al. 2012). As studies can be limited by the available type of sample or tissue, these issues should be considered and addressed in the study design (Houseman et al. 2012; Michels et al. 2013).

1.9.3 DNA Methylation Patterns and Obesity

Epigenetic abnormalities have been found to be causative factors in different diseases including cancer, autoimmune diseases and aging (Drong, Lindgren, McCarthy 2012). In addition, the epigenetic effect has long been known in syndromic obesity as in the case of Prader-Willi Syndrome that is caused by deletions of an imprinted region on chromosome 15 (Farooqi 2005), as was discussed in Section 1.5.2.

I addition, the rapid increase of the prevalence of common diseases including obesity and T2D in the past 2-3 generations, as well as the missing heritability seen for such diseases suggested a role of epigenetic changes triggered by environmental changes as a potential mechanism driving this phenomenon. Therefore, several studies have been conducted to investigate the effect of epigenetic changes in relation to common diseases.

A study conducted in Pima Native American community in Arizona has shown that offsprings of mothers with T2D during their pregnancy are at higher risk for developing T2D and obesity compared to offspring's of mothers without diabetes during pregnancy (Pettitt et al. 1983). Moreover, offspring's born to mothers with T2D are at higher risk of obesity compared to their siblings who were born when the mother had no T2D. The variation observed was preserved in children born to the same mother after the mother was diagnosed with T2D compared to their siblings born when the mother was nondiabetic, which suggests that that the variation observed is unlikely to completely reflect genetic transmission (Dabelea et al. 2000). Some studies on animal models suggest that such effect of intra-uterine environment on the susceptibility of T2D and obesity can be mediated through epigenetic processes (Morgan et al. 1999; Seki et al. 2012).

Morgan *et al.* has described the epigenetic inheritance at the agouti locus in mice, which is responsible for fur coat colour (Morgan et al. 1999). Mutations at this locus have pleiotropic effects that include obesity and yellow coat in mice. The study by Morgan *et al.* showed that despite identical genetic makeup of mice carrying the viable yellow A^{vy} allele at the agouti locus, their fur coat colours ranges from yellow to brown due to an epigenetic modification at the A^{vy} allele. The variation in fur colour was also correlated with adult body weight. The epigenetic modification at the A^{vy} allele was inherited trans-generationally through the female germline. This gives an example of an inherited epigenetic modification that skips the reprogramming process during early developmental stages.

1.9.3.1 Candidate Gene Methylation Studies on Obesity

To date, a small number of studies have investigated the role of epigenetic variations, particularly DNA methylation, in relation to obesity susceptibility and many of them were conducted on candidate genes for obesity (Godfrey et al. 2011; Groom et al. 2012; Kuehnen et al. 2012; Relton et al. 2012; Souren et al. 2011; Stepanow et al. 2011; Zhao, Goldberg, Vaccarino 2013), as summarized in Table 3. For example, Kuehnen *et al*. (Kuehnen et al. 2012) studied the methylation status of *POMC*, a monogenic obesity gene that encodes for proopiomelanocortin involved in the melanocortin 4 receptor pathway, in blood

leukocytes of obese children. They reported a hypermethylation variation associated with obesity. The methylation status at this position influences the binding activity of the histone acetyltransferase P300 complex, required for gene activation, and thus the expression of *POMC*. This illustrates how changes in DNA methylation can affect gene expression patterns and subsequently susceptibility to disease. In addition, it shows that for a key obesity gene like *POMC* both genetic as well as epigenetic mechanisms can modulate obesity predisposition.

1.9.3.2 Genome-Wide Methylation Studies on Obesity

The development of DNA methylation profiling microarrays such as the Illumina HumanMethylation450 BeadChip (Bibikova et al. 2009), and the next generation sequencing technologies coupled with bisulfide treatment, paved the way for conducting DNA methylation analysis at the genome wide level, widely recognized as epigenome-wide association studies (EWASs). The era of EWASs using large number of samples has begun and is expected to provide a more comprehensive view of the contribution of DNA methylation to disease pathogenesis (Drong, Lindgren, McCarthy 2012).

In relevance to obesity, few EWASs have been performed to date (Almen et al. 2012; Almen et al. 2014; Feinberg et al. 2010; Sun.D. et al. 2014; Wang et al. 2010b), as summarized in Table 3. Recently, a seminal paper by Dick *et al.* (Dick et al. 2014) has provided a systematic analysis of the association between variation in DNA methylation and BMI. Their initial discovery cohort consisted of 479 European individual who were typed, using whole-blood DNA, with the Infinium HumanMethylation450 array. They reported a replicated significant association between three probes, targeting CpG sites within the first intron of hypoxia inducible factor 3, alpha subunit (*HIF3A),* and BMI. HIF3A is a component of the hypoxia inducible transcription factor that regulates the cellular and physiological response to hypoxia through regulating the expression downstream target genes (Greer et al. 2012). For the most significantly associated probe, cg22891070, every 10% increase in methylation gives about 3.6% increase in BMI, which is about 0.98kg/m^2. This paper highlighted the importance of *HIF3A* pathway in relation to obesity as well as the role of epigenetic variations in determining susceptibility to obesity.

Given the important role of DNA methylation patterns in the susceptibility to obesity as well as the lack of methylation studies on obesity in Arab populations, I have sought to attempt replicating known associations between DNA methylation and BMI from large EWASs (Dick et al. 2014; Sun.D. et al. 2014) in my Qatari dataset. This will help identifying conserved associations between Caucasians and Arabs and could highlight key obesity genes. For this purpose, I have performed differential methylation

analysis with BMI in my sample set using the genome-wide DNA methylation profiling data generated with the Infinium HumanMethylation450 BeadChip, as will be discussed in details in Chapter 5.

Study target	Trait	Samples case/control (c/c) or quantitative(qt)	DNA source	Technology	Replication case/control (c/c) or quantitative (qt)	Reference
ALOX12, ALPL, BCL2A1, CASP10, CAV1, CCL3, CD9, CDKN1C, DSC2, EPHA1, EVI2A, HLA, IRF5,KRT1, LCN2, MLLT4, MMP9, MPL, NID1, NKX31, PMP22, S100A12, TAL1, VIM	BMI, fat mass, and lean mass	qt: 178	Umbilical cord blood	GoldenGate® Cancer Panel I Array	-	(Relton et al. 2012)
KCNQ1OT1, H19, IGF2, GRB10, MEST, SNRPN, GNAS	BMI (discordance in twins)	c/c: 16/16	Saliva	Bisulfite sequencing	-	(Souren et al. 2011)
MCHR1	BMI	qt: 49	Whole blood	Bisulfite sequencing	-	(Stepanow et al. 2011)
POMC	Obesity	c/c: 71/36	Whole blood	Bisulfite sequencing	c/c: 54/100	(Kuehnen et al. 2012)
IL8, NOS3, PIK3CD, RXRA, SOD1	Fat mass and %fat mass	qt: 78	Umbilical cord tissue	SequenomMassARRAY	qt: 239	(Godfrey et al. 2011)
SLC6A4	BMI, weight, and waist circumference	qt: 168	Peripheral blood leukocytes	Bisulfite pyrosequencing	-	(Zhao, Goldberg, Vaccarino 2013)
TACSTD2	Fat mass	qt: 94	Whole blood	Bisulfite pyrosequencing	qt: 161	(Groom et al. 2012)
EWAS	BMI	qt: 64	Lymphocytes	Comparative array-based relative methylation (CHARM)	-	(Feinberg et al. 2010)
EWAS	Obesity	c/c: 7/7	Peripheral blood leukocytes	Infinium HumanMethylation27 BeadChip	c/c: 46/46	(Wang et al. 2010b)
EWAS	Obesity	c/c: 23/24	Whole blood	Infinium HumanMethylation27 BeadChip	-	(Almen et al. 2012)
EWAS	BMI	qt:479	whole blood	Infinium HumanMethylation450 BeadChip	qt:339, qt:1789	(Dick et al. 2014)
EWAS	Obesity	c/c: 48/48	Whole blood	Infinium HumanMethylation450 BeadChip	-	(Xu et al. 2013)
EWAS	Obesity	c/c: 24/22	Whole blood	lInfinium HumanMethylation27 BeadChip	-	(Almen et al. 2014)
EWAS	BMI	qt: 830	Whole blood	Infinium HumanMethylation450 BeadChip	qt:138	(Sun.D. et al. 2014)

Table 3.Candidate gene studies and epigenome-wide association studies of DNA methylation on obesity

1.10 Aims

Despite the alarming figures of obesity in the Arab world, the genetics of obesity remain understudied in Arab populations. Studies mainly performed on Caucasians have shown that single nucleotides polymorphisms, copy number variations as well as DNA methylation patterns contribute to the susceptibility to obesity in humans. Therefore, I sought to study these various contributors to obesity predisposition in an Arab population. This may identify novel contributions to obesity susceptibility and/or underscore conserved contributions found in other populations.

The specific steps to achieve this goal were to:

1- Start the collection of DNA from multigenerational Qatari families segregating obesity for the purpose of the study.
2- Perform genome-wide SNP genotyping (using the HumanOmni2.5-8 Beadchip), whole genome sequencing (using the Hiseq2500 platform) and genome-wide methylation profiling (using the Infinium Human Methylation450 Beadchip) through Illumina using whole-blood DNA.
3- Identify candidate obesity regions in the families through linkage and run of homozygosity analyses.
4- Identify novel or known rare variants possibility contributing to obesity risk in these families using variant prioritization analyses.
5- Identify copy number variants genome-wide in my sample set and screen predicted common and large rare CNVs for association with BMI.
6- Perform genome-wide association analysis of DNA methylation with BMI in my Qatari dataset to attempt replicating known associations.
7- Investigate the impact of differing genetic background and lifestyle on the replicated DNA methylation associations by comparing the effects observed in the Qatari dataset to the TwinsUK cohort.

Collectively, the work of this thesis represents the first genetic and epigenetic study on obesity in Qatar and the Arab countries, and it can serve as a pilot for future studies.

CHAPTER TWO

Materials and Methods

2. Materials and Methods

2.1 Ethical Approval of the Study

Ethical approval was obtained from Weill Cornell Medical College in Qatar (WCMC-Q) Institutional Review Board (IRB) on the 26thof September 2012 for the purpose of this study (IRB number: 2012-0025). This is in concordance with the Helsinki declaration of ethical principles for medical research involving human materials. The application forms, consents, and required documentations were prepared by the author. Ethical approval is provided in Appendix 1.1.

2.2 Selection of Study Subjects

Study subjects were identified through collaboration between WCMC-Q and Qatar Diabetes Association (QDA); a health centre that provides secondary care for diabetic patients as well as awareness programs on the management and prevention of diabetes and obesity in Qatar. Qatari obese probands (adults: BMI≥30, children: BMI>2SD above the mean for age-gender) with familial history of obesity were initially contacted to participate in the study, during their visits to the QDA. Probands were introduced to the study and asked to fill an information sheet with their personal information and contact details, for arranging home visits for family recruitment. A team consisting of the author and a nurse visited the families at home for recruitment. All family members including cases and controls were introduced to the study objectives and protocols, and a written informed consent was obtained from all participants (and their parents if participant's age was between 13-18 years). A total of 96 subjects from ten families participated in the study.

2.3 Samples Collection

Venous blood samples (~7ml) were drawn from participants by a specialized nurse and were kept in Ethylenediaminetetraacetic acid (EDTA) anticoagulant tubes for DNA extraction. Within 24 hours of collection, blood samples were aliquoted into 2ml tubes and stored at -80°C.

2.4 Phenotypic Data Collection

Anthropometric data like age, sex, height, and waist circumference were recorded for study participants. In addition, measurements of weight, BMI, fat mass (FM), fat mass percentage (FM %), fat-free mass (FFM), fat-free mass percentage (FFM %), muscle mass, muscle mass percentage, bone mass and basal metabolic rate (BMR) were obtained using a high capacity body composition analyser Bc-

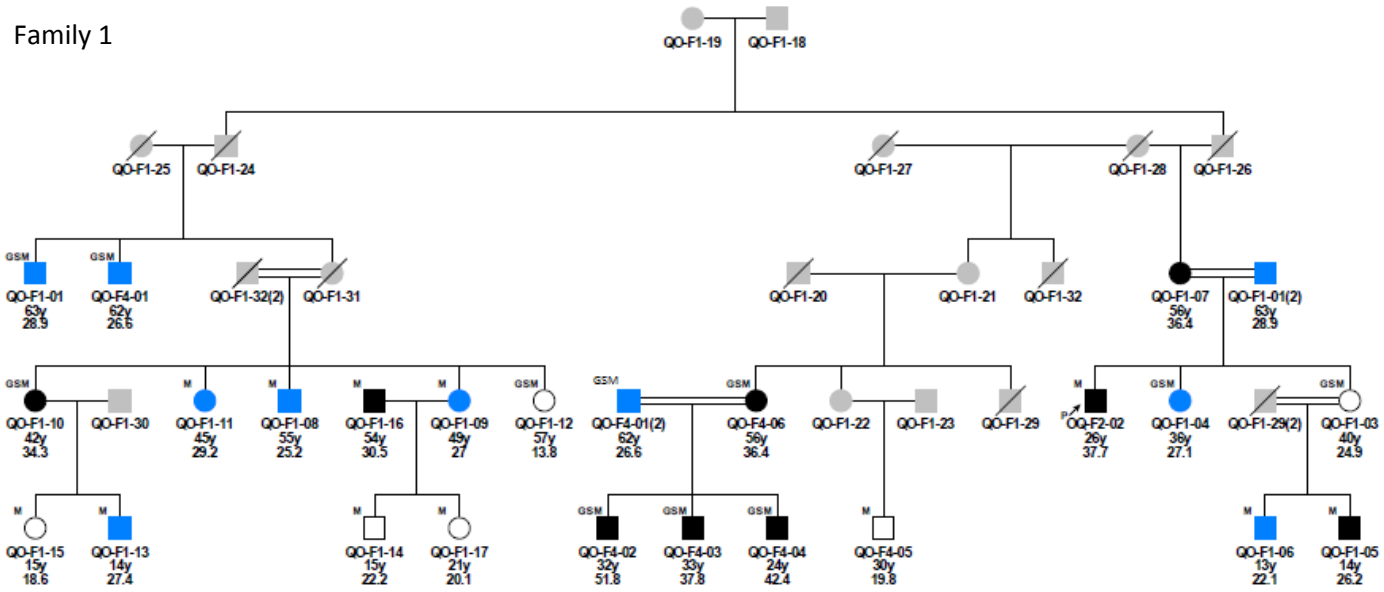
420MA device (Tanita, UK LTD.). BMI was calculated as weight in kilograms divided by the square of the height in meters squared. BMI is an inexpensive and simple measure for obesity where excess body weight relative to height indicates excessive fat mass. However, excess body weight that given by BMI measures could also indicate high bone or muscle mass, therefore in this study I also obtained the body composition measurements, which are based on the bio-impedance analysis.

The participants also filled a short questionnaire on health and lifestyle to get general information about their health status, and to also ensure that none of the affected subjects have a genetic syndrome. In addition, detailed information about the family structure was collected and checked with the participants during the visits, to generate family pedigrees, using Haplopainter (Thiele and Nurnberg 2005). The overall characteristics of the study participants are provided in Table 4, and the family pedigrees are shown in Figure 2.

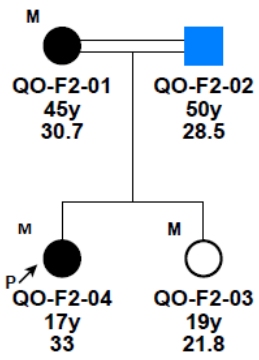
Characteristics	**Males N=39 (40.6%)**	**Females N=57 (59.4%)**
Age (years)	34.2±16.9	36.5±15.7
BMI(kg/m^2)	30.3±7.7	29.3±6.4
Phenotype	10 normal, 8 overweight, 21 obese	16 normal,13 overweight, 28 obese
Fat mass (kg)	26.8±13.1	25.7±13.6
Fat mass percentage (%)	28.4±17.7	33.2±11.0
Fat free mass (kg)	57.7±15.9	45.8±6.2
Fat free mass percentage (%)	70.1±9.7	66.6±11.2
Bone mass (kg)	3.1±0.4	3.3±0.3
Muscle mass (kg)	56.3±11.3	43.2±5.9
Muscle mass percentage (%)	68.0±0.4	62.9±11.0
Basal metabolic rate (kJ)	7843.2±1215.1	5944.3±759.7
Waist circumference (inches)	39.7±7.70	35±6.6
T2D status	3 diabetics, 34 nondiabetics, 2 unknowns	9 diabetics, 48 nondiabetics
Smoking status	8 smokers, 29 nonsmokers, 2 unknowns	57 nonsmokers

Table 4. Characteristics of the study participants.

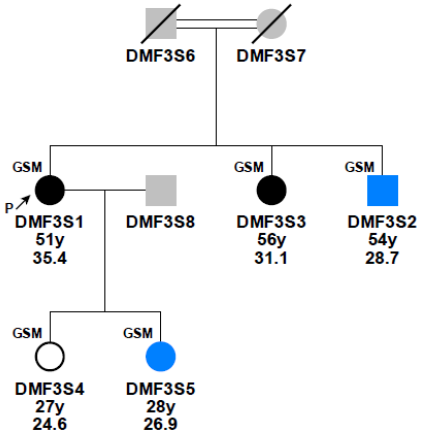
Family 1



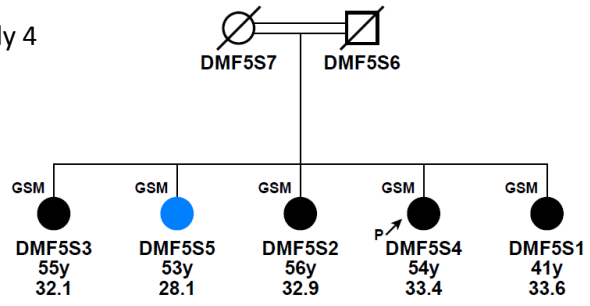
Family 2



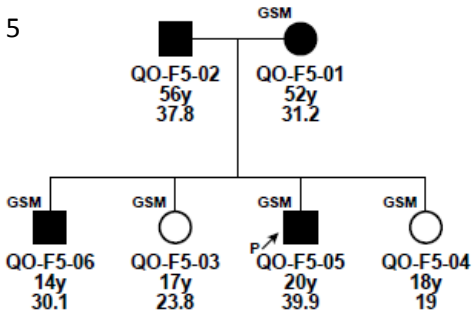
Family 3



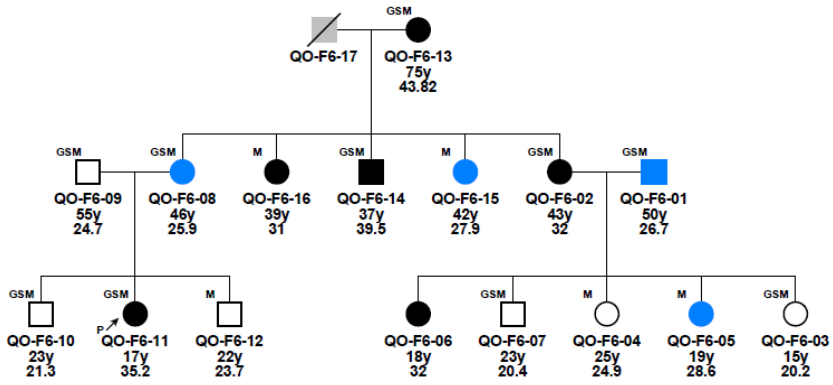
Family 4



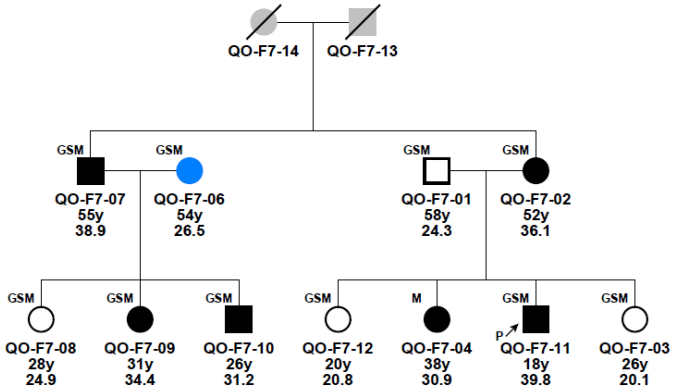
Family 5



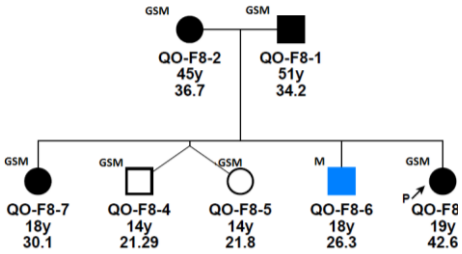
Family 6



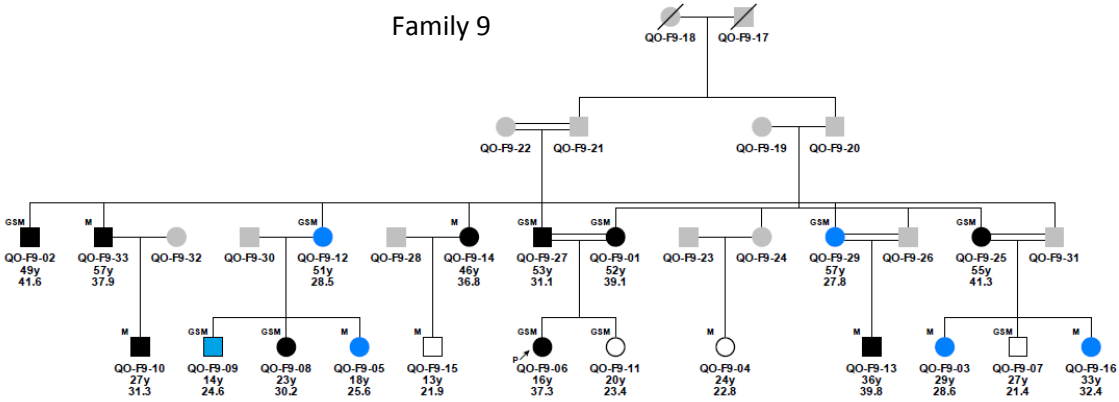
Family 7



Family 8



Family 9



Family 10

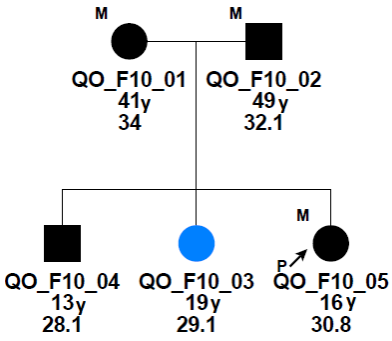


Figure 2. Pedigrees of ten Qatari families participated in the study. Obese subjects are shown in black, overweight in blue, normal in white and unknowns in grey. The probands are indicated with arrows. Sample ID, age and BMI are given below the individual symbol, respectively. The

label beside some individuals indicates the availability of genotyping, sequencing and methylation data for these subjects, while the M label indicates the availability of methylation data only.

2.5 Extraction of Genomic DNA

I performed the extraction of genomic DNA at WCMC-Q clinical laboratory, from 2ml of whole blood using QIAamp Blood Midi Kit (Qiagen, Crawley, UK). This procedure was carried out according to the manufacturer instruction including the recommended proteinase K and RNase A digestions. DNA concentration and purity were measured using NanDrop 2000c/2000 UV-Vis Spectrophotometer (NanoDrop, Thermo Fisher Scientific Inc.). In addition, the samples were quantified using Qubit 2.0 fluorometer from Invitrogen with the Qubit dsDNA BR Assay Kit (Invtrogen Life Technologies, USA) for accurate quantification. DNA samples were stored frozen at -80°C.

2.6 Generation of Genomic Data

This section will explain how SNP genotyping data, whole genome sequencing data and array-based methylation profiling data were generated. This was carried out at Illumina on a fee-for-service basis.

A total of 63 samples belonging to 8 families were selected for performing WGS. 5ug of DNA were prepared per sample in a total volume of 100ul (concentration= 50ng/ul based on Qubit readings) and shipped frozen in dry ice to Illumina Fast Track service to perform paired-end whole genome sequencing, using Hiseq2500 platform. In addition, these samples were also used to perform genome-wide SNP genotyping using HumanOmni2.5-8 arrays, as a quality control step by Illumina for their WGS service.

In addition, 96 samples belonging to 10 families were sent for genome-wide DNA methylation profiling, using the Infinium Human Methylation450 Beadchip. For that, 7ug of DNA were prepared per sample in a total volume of 100ul (concentration=70ng/ul based on Qubit readings), in a separate plate and sent frozen to Illumina. Table 5 summarizes the types of data generated for the purpose of this study.

Data type	Subjects	Method	Platform	Features	Assay	DNA required
SNP-array genotyping	8 families (n=63)	HumanOmni2.5-8BeadChip	iScan or HiScanSQ	2,379,855 SNPs (1kGP;MAF>2.5%)	Infinium LCG	200ng
Whole genome sequencing	8 families (n=63)	WGS from short-insert paired end reads	HiSeq 2500	Minimum of 35× average coverage	Truseq DNA (PCR –free)	3-5ug
DNA methylation	10 families (n=96)	Infinium HumanMethylation450 BeadChip	iScan or HiScanSQ	> 485,577 methylation sites within and outside CpG islands	Infinium HD (PCR free)	500ng (1 µg with automation)

Table 5.Summary of the genome-wide datasets generated in this study.

2.6.1 Genome-Wide SNP Genotyping using the HumanOmni2.5-8 Beadchip

The Beadarray® technology forms the fundamental basis of the high throughput SNP genotyping arrays developed by Illumina. This technology involves the assembly of hundreds of thousands of copies of 50-mer oligonucleotides that are complementary in their sequence to the sequences immediately adjacent to target SNPs onto 3-micron silica beads (Illumina 2015). These beads are self-assembled in microwells of fiber optic bundles with uniform spacing of ~5.7 microns and each bead is designed to be specific for tagging a particular SNP (Illumina 2015).

The HumanOmni2.5-8 Beadchip features 2,379,855 SNPs across the genome to provide comprehensive coverage of infrequent and common variants from the 1000 genome project (1kGP) with MAF > 2.5%. These SNP markers serve as tag SNPs for other highly correlated markers within their haplotype blocks, and they were selected to be maximally informative for diverse world populations (Illumina 2013). In addition, the high density of the HumanOmni2.5-8 Beadchip, with average distance of 1.19 kb between SNPs, makes it a powerful platform for studying copy number variations (Illumina 2013).

Due to the high efficiency, accuracy and robustness of the array genotyping technology (Steemers et al. 2006), Illumina performed genome-wide SNP genotyping, as a quality control step for their WGS service, to validate sequencing accuracy and sample identity. Figure 3 shows the workflow for performing the SNP genotyping using Illumina Infinium Assay. Briefly, DNA samples were whole-genome amplified, fragmented, precipitated and resuspended according to the Infinium LCG Assay Guide (Illumina). The samples were then loaded on the HumanOmni2.5-8 arrays (eight samples per array) for hybridisation to the oligonucleotide probes, followed by a washing step to remove the unhybridised or the non-

specifically hybridised DNA. A two-colour single-base extension step was then carried out in which differentially labelled nucleotides were added to the oligonucleotide probes in a complementary manner to the nucleotide present in the hybridised sample DNA. The HiScan or iScan system then scanned the beadchip by laser to excite the fluorophore of the extended base, and high resolution images of the fluorophores emitted light were recorded. The output images and intensity files were imported into GenomeStudio software (Illumina) to extract the genotype information.

The SNP array genotyping data generated will be utilized in this study to perform linkage analysis and ROH analysis as well as CNV analysis.

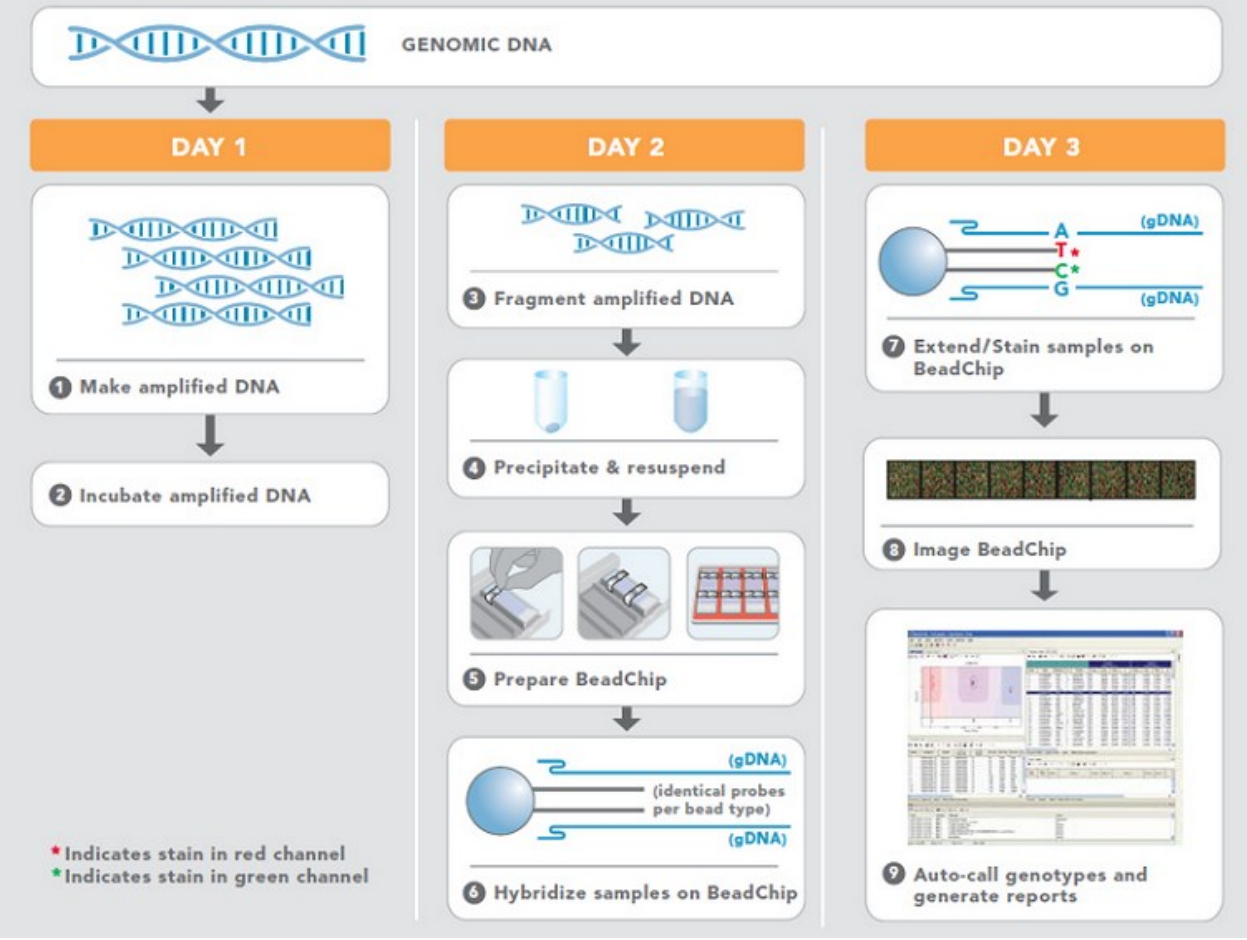


Figure 3. Workflow for genome-wide SNP genotyping using the Illumina Infinium assay. Figure is reproduced from Illumina.com (see Appendix 2.2).

2.6.2 Whole Genome Sequencing (WGS) from Paired-End Reads

Paired-end sequencing is an advanced technology that allows rapid and parallel sequencing of genomic segments which produces massive number of short sequencing reads that represents together the sequence of the whole genome. The sequencing technology implemented by Illumina is based on the fundamental idea of sequencing by synthesis (SBS) technology, in which single bases are sequenced as they are incorporated into the growing DNA strand, as illustrated in Figure 4.

In this study, 63 samples were whole genome sequenced by Illumina using the Hiseq 2500 platform with a minimum of 35x average coverage. The service process involves quantification of DNA samples using PicoGreen (Quant-iT PicoGreen dsDNA Reagent, Invitrogen, Catalogue #: P11496) on Spectromax Gemini XPS (Molecular Devices), to ensure good sample quantity, prior to library construction. Then, paired-end libraries were generated with 1ug of genomic DNA using the Illumina TruSeq DNA Sample Preparation Kit (Catalogue #: FC-121-2001), according to the TruSeq DNA PCR-Free Sample Preparation Guide. Libraries were then subjected to quality control step using gel electrophoresis to check their sizes and were then quantified. This was followed by denaturation, dilution and clustering of libraries onto v4 flow cells using the Illumina cBot™ system, according to cBot User Guide.

The clustered flow cells were processed using HiSeq 2500 instruments and sequenced at a minimum of 100bp paired-end runs, following the instructions in HiSeq 2500 User Guide. Real-time image analysis and base calling for the Hiseq runs was performed using Illumina HiSeq Control Software (HCS) and Real-Time Analysis (RTA). The sequences generated were then processed through Illumina's bioinformatics pipeline to obtain a full range of called variants for each sample. In this study, the raw sequencing data, generated by Illumina, was processed using the analysis pipeline implemented in the Bioinformatics core at WCMC-Q as will be discussed in Section 2.8.1.3, and was used for the purpose of identifying rare obesity variants as will be discussed in Chapter 3.

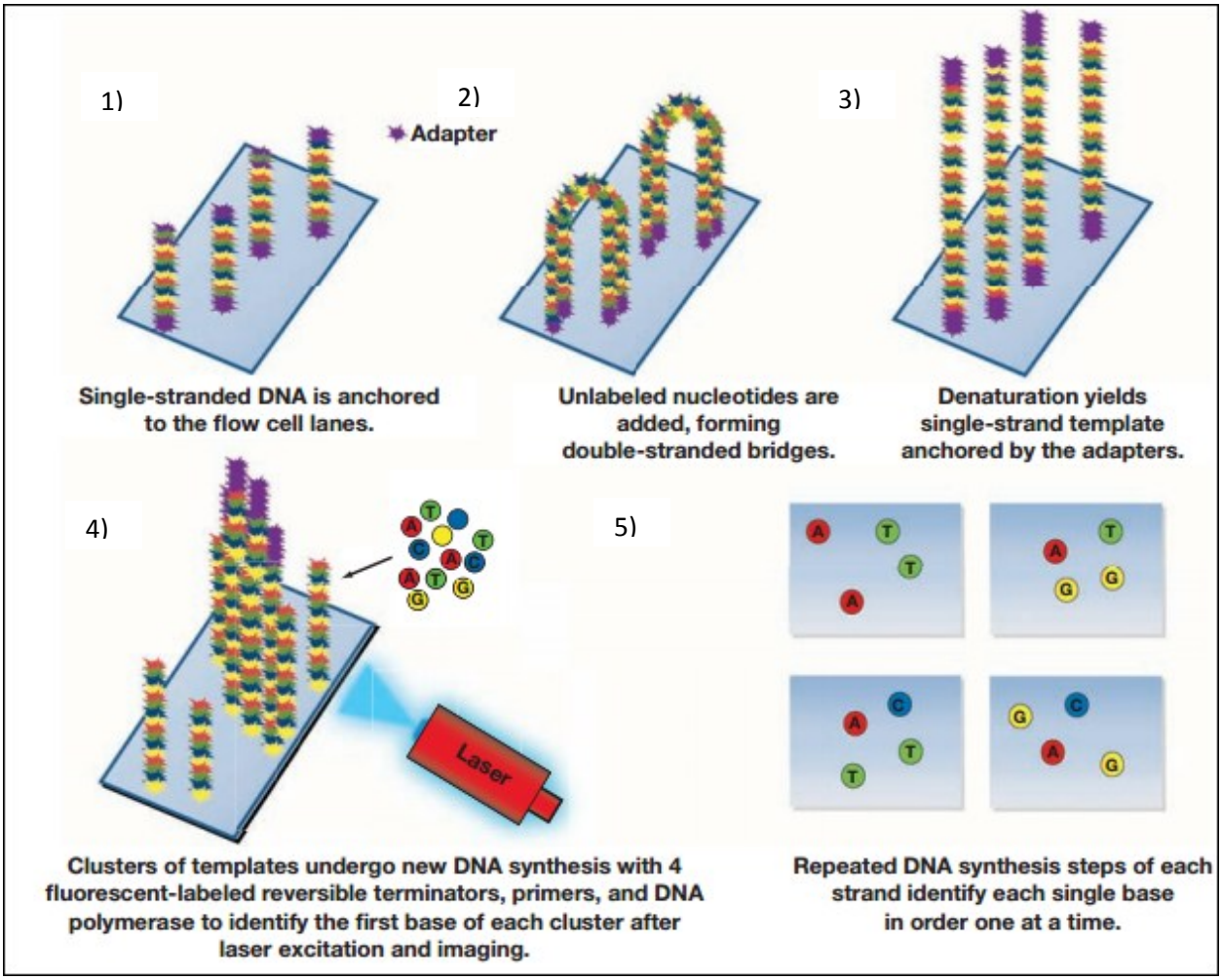


Figure 4. Schematic diagram of Illumina sequence by synthesis next-generation sequencing platform. Figure was reproduced with permission from Ross *et al.* (2011) (Ross and Cronin 2011), as given in Appendix 2.3.

2.6.3 Genome-Wide Methylation Profiling using Infinium HumanMethylation450 BeadChip

The Infinium HumanMethylation450 BeadChip offers a cost-effective analysis of DNA methylation at genome-wide level. This technology features more than 485,577 CpG sites covering 99% of RefSeq genes with an average of 17 CpG sites per gene region (Bibikova et al. 2009; Bibikova et al. 2011). These CpG sites targets 96% of CpG island regions in the genome and are distributed across promoter regions, 5'UTR, 3'UTR as well as gene body (Illumina 2012a).

The Infinium assay technology relies on quantifying sodium bisulfite induced C to T conversions at defined genomic positions using a combination of two assays; Infinium I and Infinium II (Illumina 2012b).

As illustrated in Figure 5, Infinium I assay uses especially designed two bead type probes per CpG locus; one is complementary to the unconverted (methylated) sequence and the other is complementary to the C-to-U-converted (unmethylated) sequence and base extension is the same for both alleles. The design of this assay is based on the co-methylation assumption within a 50 bp span from CpG query sites (Illumina 2012b). On the other hand, Infinium II assay uses a single bead type per CpG locus, and the methylation status of the query site is determined by a single base extension step after hybridisation. With Infinium II assay, the underlying CpG sites (up to 3) are usually represented by "degenerate" R-base, allowing the assessment of the methylation status at the query sites independently from neighbouring sites (Illumina 2012b). These two assays were employed in the HumanMethylation450 BeadChip to enhance breadth and coverage.

In this study, 96 samples were submitted to Illumina to perform the methylation profiling using the Infinium HumanMethylation450 BeadChip, as a service. The procedure was carried out according to the Infinium HD Assay Methylation Protocol Guide. Briefly, DNA was bisulfite converted using bisulfite conversion kit (Illumina). The bisulfite-converted DNA was then denatured; neutralised, whole genome amplified and then fragmented using an enzymatic process. The fragmented DNA was precipitated, resuspended and loaded into the Infinium HumanMethylation450 BeadChip for hybridisation overnight. This beadchip has an 8-sample format that allowed the processing of eight samples per chip. The beadchip was then washed and as single-base extension step using fluorescently labelled bases was carried out to distinguish between a C or a T nucleotide in the converted DNA. The HiScan or iScan system was used to scan and record fluorescent images of the beadchip, which were then processed through the methylation module of GenomeStudio to extract the methylation information, as will be discussed in Section 2.8.3.1.

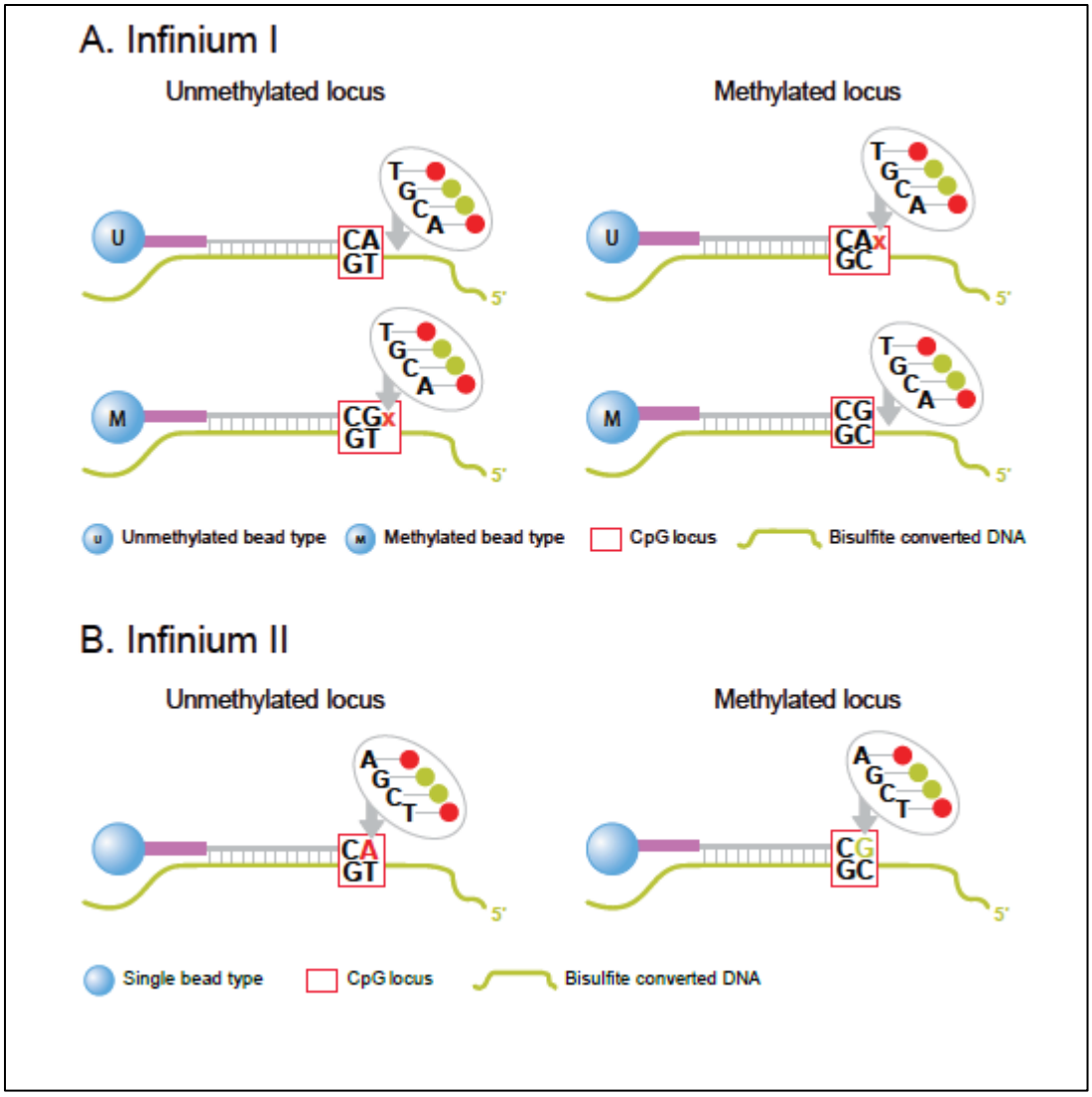


Figure 5. The two assays employed in the Infinium HumanMethylation450 BeadChip. A) Infinium I assay employs two bead types per CpG site, one for the methylated and the unmethylated states. B) Infinium II assay uses one bead type per CpG site, and the methylation state is determined by single base extension step after hybridisation. Figure is reproduced from Illumina.com (see Appendix 2.2).

2.7 Additional Genomic Data Sets

2.7.1 Data for 108 Qatari Genomes

In this study, I utilized existing data from 108 unrelated Qatari subjects recruited by the team of Drs Ronald Crystal and Khalid Fakhro at WCMC-Q. The purpose was to replicate findings from Chapters 3 and 4 of this thesis. The dataset for the 108 Qatari genomes was initially collected to conduct population studies on the Qatari population. The dataset consists of genome-wide SNP genotyping data (using the HumanOmni2.5 SNP array) and WGS data (using the Hiseq 2500 platform) that were generated by Illumina on fee-for-service basis. The sample set includes 87 adults with BMI data (52 obese, 27 overweight and 8 normal). The CNV analysis of the 108 Qatari genomes was carried out using PennCNV by the author (data are not shown).

2.7.2 Data from Adult Obesity and Childhood Obesity Case-Control Cohorts

I tried to replicate CNV findings from Chapter 4 using the CNV results from the PhD work of my colleague Julia Sarah El-Sayed Moustafa available in her thesis entitled “Genomic Structural Variation and its Impact on the Human Phenotype”. In particular I used the CNV information on the child and adult obesity cases and controls which consisted of 646 child obesity cases and 589 child controls, and 709 adult obesity cases and 197 adult controls. These samples were genotyped on the Illumina HumanCNV370-Duo chip and the CNVs were called using PennCNV (Wang et al. 2007) and cnvHap (Coin et al. 2010).

2.7.3 Data from the Saudi Extreme Obesity and Mental Retardation Case-Control Sample Set

I aimed to replicate CNV findings from Chapter 4 using the CNV calls given in the PhD thesis of my colleague Haya Al-Saud, which is entitled “The Genetics of obesity in Saudi Arabian Population”. The sample set consisted of adults and children with a total of 85 cases and 72 controls. These samples were genotyped using Affymetrix arrays (Affymetrix 6.0, Axiom GWH-96 and CytoScan HD), and the CNVs were called using PennCNV.

2.7.4 Data from Qatari Families with T2D

My DNA methylation data was analysed together with another DNA methylation dataset that was generated by my colleague Wadha Al Muftah to study DNA methylation patterns in relation to T2D. The dataset consists of methylation and WGS data for a total of 27 subjects from 6 Qatari Families. Both

datasets have matching sample ethnicity, sample type (whole blood), DNA extraction method (QIAamp blood midi kit), and methylation profiling technology (Infinium HumanMethylation450 Beadchip). Therefore, the two datasets were combined to increase sample size. The combined dataset consists of a total of 123 subjects from 16 Qatari families. This dataset will be analysed for the purpose of identifying differentially methylated sites with BMI as will be discussed in Chapter 5 of this thesis, and for the purpose of identifying differentially methylated sites in T2D as discussed by my colleague Wadha Al Muftah in her PhD thesis, which is entitled "The Genetics and Epigenetics of T2D among Qatari Families".

2.7.5 Data from TwinsUK Cohort

The TwinsUK cohort was established in 1992 (Moayyeri et al. 2013) and it currently consists of more than 13,000 MZ and DZ twin participants (with same gender), the majority (82.7%) of whom are healthy Caucasian females. The genome-wide methylation data (using the Infinium HumanMethylation450 BeadChip) for 810 twin participants (with BMI information) was used to replicate methylation loci associated with BMI in my Qatari dataset. In addition, the data was used to investigate the impact of genetic background and lifestyle/environmental pressures on these associations by comparing the effects observed in the Qatari and TwinsUK datasets in a meta-analysis framework, as will be discussed in Chapter 5. Details on the experimental approaches carried on the TwinsUK methylation data was previously described by Tsaprouni (Tsaprouni et al. 2014).

2.8 Data Processing and Analysis

This section will explain how SNP genotyping data, whole genome sequencing data and array-based methylation profiling data were processed and analysed to achieve the study objectives. Therefore, this section is divided into three major sub-sections: identification of rare obesity mutations, identification of common and rare obesity CNVs, and identification of differentially methylated sites with BMI.

2.8.1 Identification of Rare Obesity Mutations

As introduced earlier, one of the aims of this study was to identify rare obesity variants. The strategy I applied to achieve this goal involved two major steps. The aim of the first step was to identify candidate obesity regions that may harbour rare variants through linkage analysis, and to prioritize the candidate linkage regions further through run of homozygosity (ROH) analysis as it might help identifying homozygous obesity variants. The rationale behind identifying candidate obesity regions as a first step

before searching for rare obesity variants was to narrow down the search to a reduced number of genomic regions and thus variants to those that are most promising. Linkage and ROH analyses were carried out using the high density SNP genotyping data. The aim of the second step was to look for potential rare obesity variants within monogenic obesity genes as well as the candidate obesity regions identified through linkage and ROH analyses. This second step was carried out using the WGS data.

2.8.1.1 Linkage Analysis

2.8.1.1.1 Extraction of SNP Genotyping Data

Genotypes were extracted from the raw HumanOmni2.5-8 SNP array intensity profiles using the genotyping module of GenomeStudio software (Illumina), and were exported in Plink format using Plink Input Report Plug-in tool. This generated a pedigree (.ped) file containing family ID, individual ID, paternal ID, maternal ID, gender and affection status followed by the genotypes for all markers, and a map file (.map) providing information about the SNP markers including chromosome, marker name, genetic position in centimorgan (CM) and physical position in base pairs (bp). The SNP genotyping dataset contained a total of 63 subjects from 8 families.

2.8.1.1.2 Quality Control of SNP Genotyping Data

Before performing linkage analysis, quality control was performed on the SNP genotyping data using Plink (Purcell et al. 2007), Pedstats (Abecasis and Wigginton 2005), Pedwipe and Merlin (Abecasis et al. 2002).

The first group of QC steps was performed with Plink (available at http://pngu.mgh.harvard.edu/purcell/PLINK/). First, gender check was carried out to ensure that the gender inferred from the genotyping data matches the clinical records, using the option --check-sex. All genotyped samples had matching gender information, therefore none were excluded. Second, samples with more than 5% missing genotypes were supposed to be excluded, as this can be an indication of sample genotyping failure, however, all samples had call rates > 95% and thus none were dropped out based on this criterion. Third, SNPs with more than 5% missing genotypes were excluded (31,404 SNPs) as they indicate SNP genotyping failure. Fourth, SNPs with minor allele frequency (MAF) of <5% (calculated on the founders only) were excluded (1,049,751 SNPs) because the small size of the homozygous and the heterozygous clusters makes calling these variants difficult and thus they might present as false positives. Fifth, SNPs that failed the Hardy-Weinberg equilibrium (HWE) test at a

significance level of $1X10^{-6}$ were removed (9 SNPs) as they might be indicative of genotyping error. Sixth, SNPs with more than 5% Mendelian errors were excluded (2 SNPs) as they may also indicate genotyping error.

Next, Pedstats was used to identify genotypes showing Mendelian inconsistencies and Pedwipe was used to remove those genotypes from the dataset. Since Merlin can only analyse markers with unique genetic positions for performing linkage analysis, markers that map to the same genetic position were excluded (248,057 SNPs). The number of SNP markers was then reduced further to be suitable for performing linkage analysis. The reason for that was to avoid the bias in the linkage statistics that may arise from the existence of high linkage disequilibrium (LD) between markers (Abecasis and Wigginton 2005).

The first method applied to do that (Method A) was to select markers based on information content (IC), which is also known as information extraction, and LD measures. Information content is an estimation of the amount of information that can be derived from family linkage using a particular set of markers and it is the probability of heterozygosity expected under HWE (Kruglyak 1997). Herein, IC was calculated, using the --information option in Merlin (Abecasis et al. 2002), and then the 25% least informative SNPs were excluded, using Plink --exclude option, in order to ensure high informativeness of linkage markers. In addition, LD was calculated between pairs of markers in a window using the --independent-pairwise option in Plink, and one of a pair of SNPs was pruned out, using the --exclude option, if LD was greater than 0.1 in order to reduce marker-marker LD bias in the marker set.

The data was also checked for hidden non-Mendelian errors, which are genotypes that imply the occurrence of excessive unlikely recombination events between closely linked markers as they may reflect genotyping errors. The hidden errors were detected using Merlin and were wiped with Pedwipe (Abecasis et al. 2002). This method gave a total of 20,872 SNPs to be used for linkage analysis. Figure 6 summarizes the quality control and markers' selection steps using Method A.

In an effort to maximize the IC of linkage markers and hence the power of linkage detection, I selected a second set of markers using a different method (Method B) and tested the IC compared to Method A. Method B markers were selected based on Illumina's V panel linkage marker set, since they were optimized to ensure high IC level and thus higher likelihood for detecting recombinant events (Murray et al. 2004). The high information content given by those markers throughout the genome was obtained through the selection of markers with high heterozygosity and with appropriate density (Murray et al.

2004). According to Illumina, the average heterozygosity of the linkage V panel markers is 43% in Caucasians, 36% in the Han/Japanese population and 33% in the Yoruba population and the average genetic distance between the markers is 0.62 cM (Illumina 2007).

Therefore, after performing the fifth QC step (Mendelian errors detection), I extracted the V panel linkage markers from my dataset, using Plink. A total of 2,521 markers out of the published 6,056 linkage V panel markers were present in the dataset. These markers were used with their genetic and physical positions provided by Illumina linkage V panel map, which gives accurate physical and genetic positions. Similar to the steps applied for Method A, markers that map to same genetic positions and genotypes presenting hidden non-Mendelian errors were excluded prior to running linkage analysis. This gave a total of 63 subjects and 2,456 SNP markers for linkage analysis. Figure 7 summarizes the quality control and markers' selection steps using Method B.

For each of the 22 autosomes, Method A marker set (20,872 SNP markers) gave higher density and was slightly more informative than Method B (2,456 SNP markers), as shown in Figures 8 and 9. The average genome-wide information content for Method A markers was 0.80 while for Method B was 0.72, with average distance between markers of 0.13 and 1.70 cM for Method A and Method B respectively. Since Method A marker set was more informative, and it also provided higher marker density through the genome, I used Method A markers to carry out the linkage analysis.

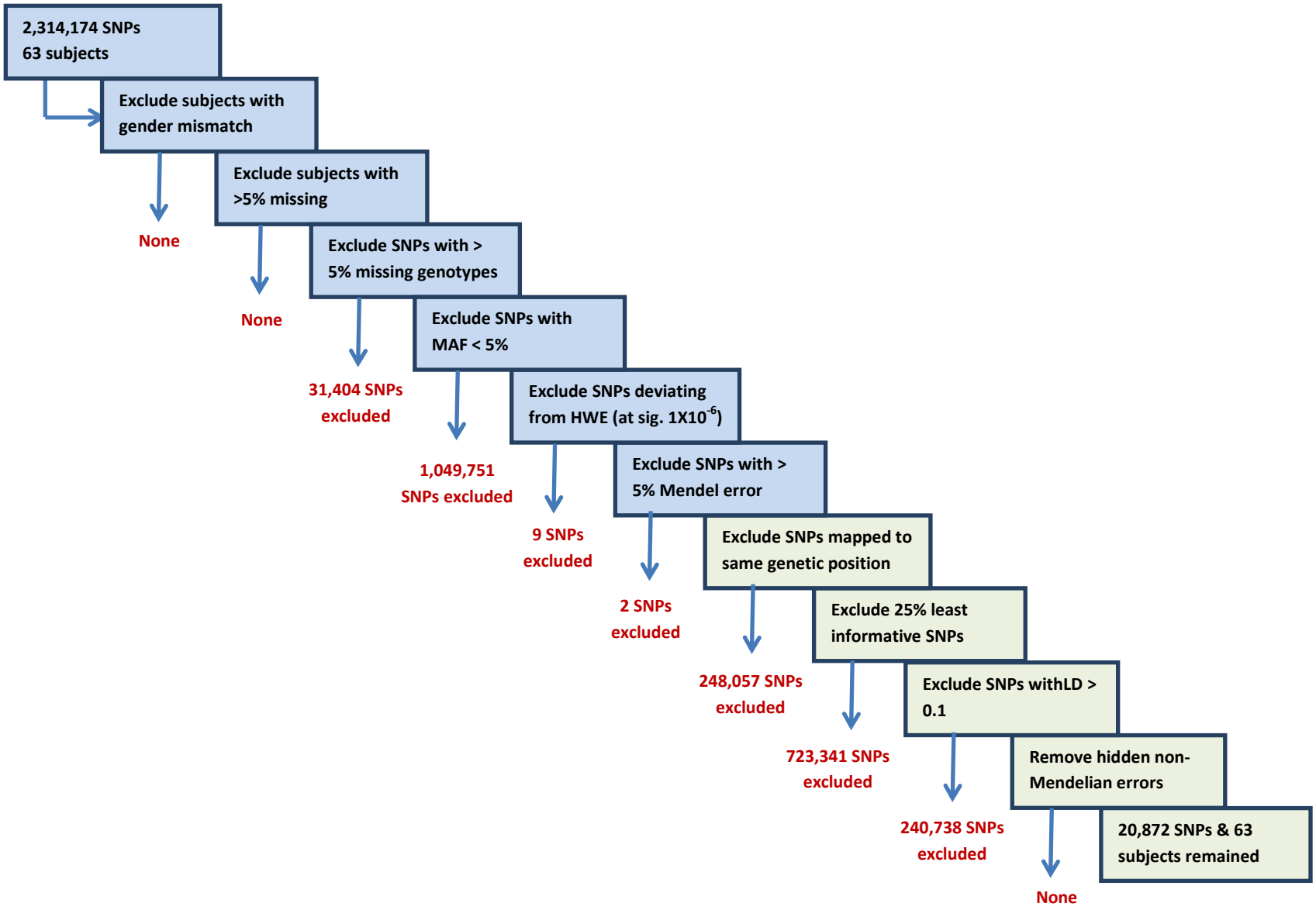


Figure 6. A summary of quality control and marker selection steps using Method A. Blue boxes shows steps common between Method A and Method B (also given in Figure 7), while green boxes shows steps specific to Method A.

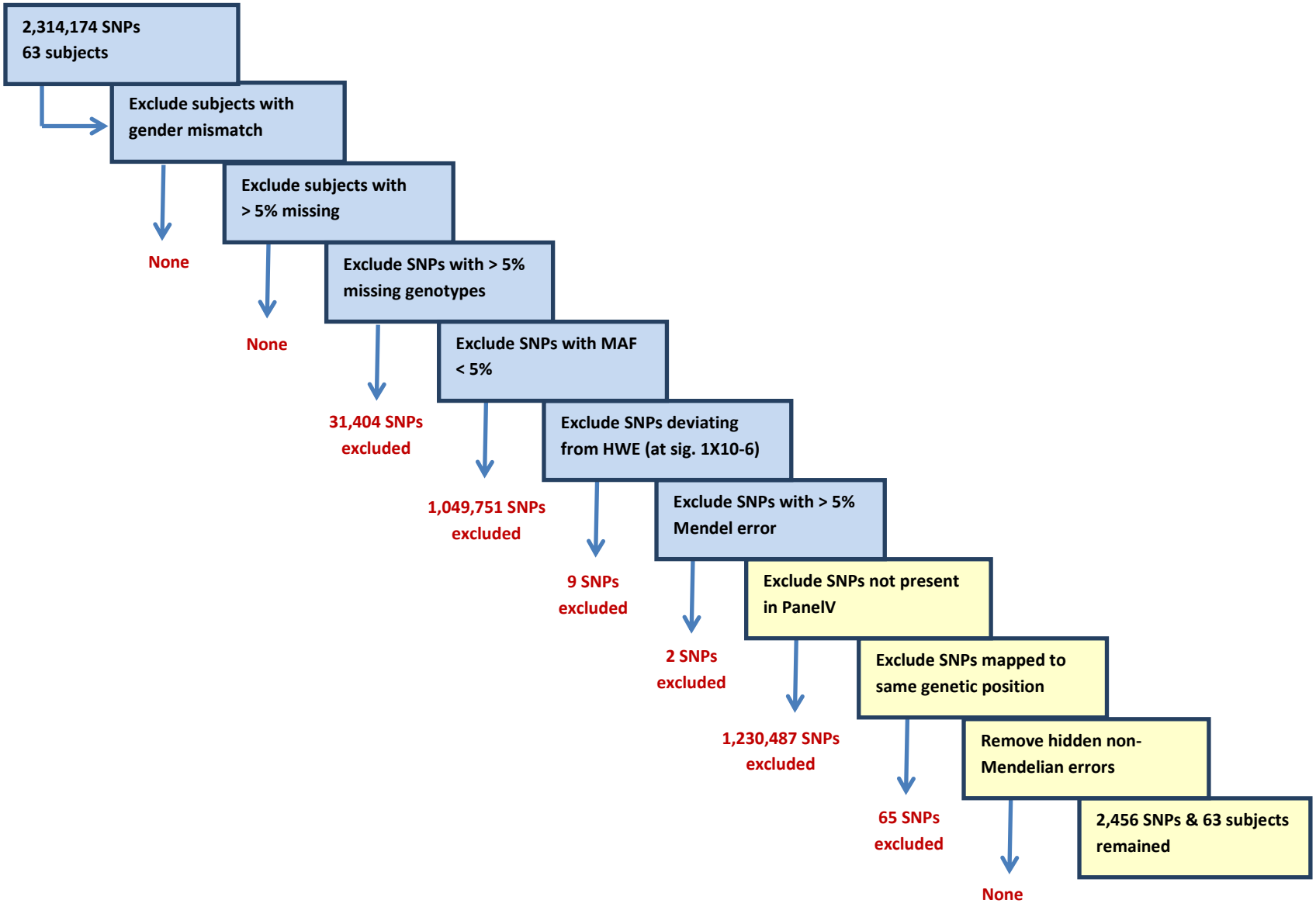


Figure 7. A summary of quality control and marker selection steps using Method B. Blue boxes shows steps common between Method A (also given in Figure 6) and Method B, while yellow boxes shows steps specific to Method B.

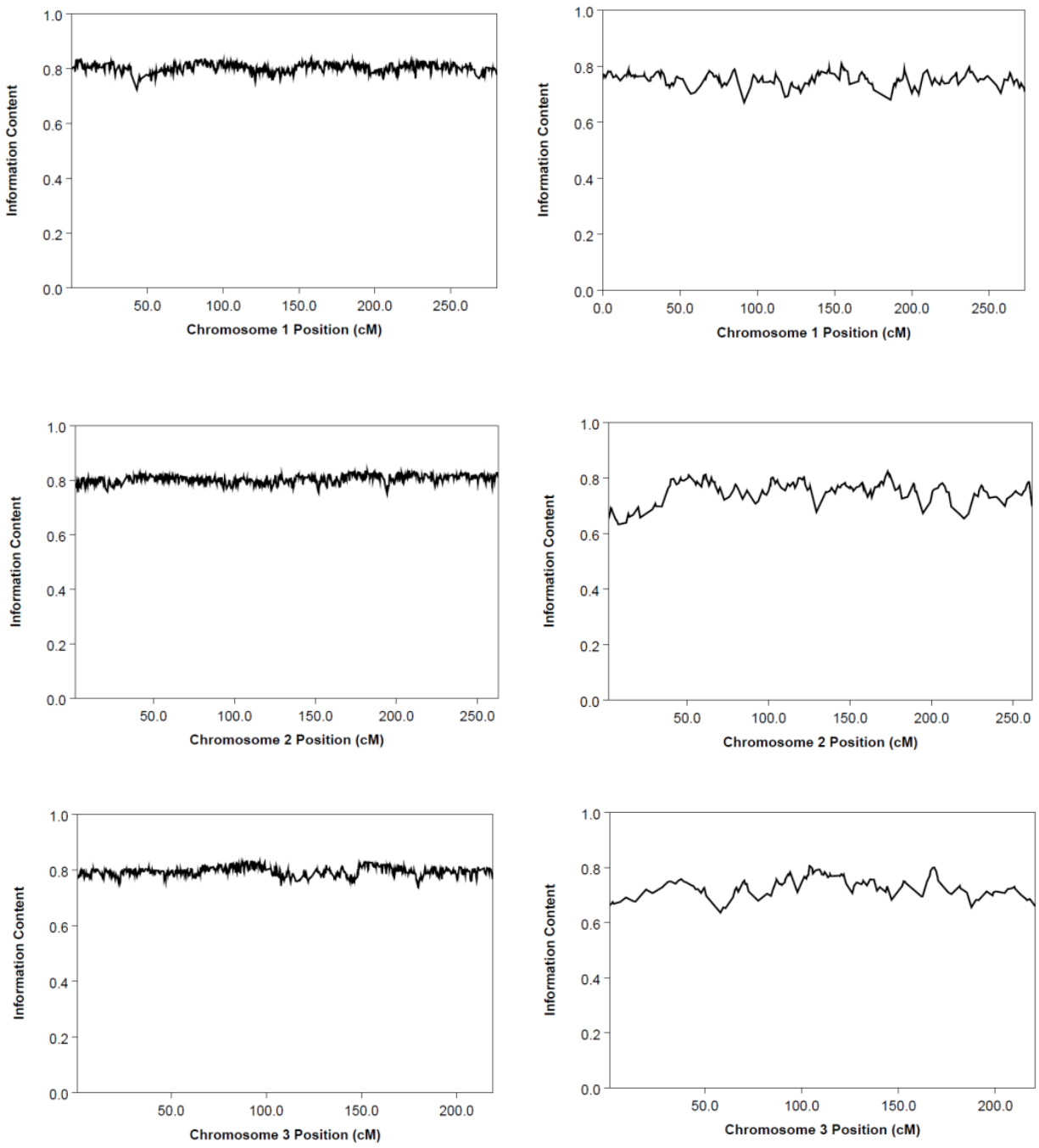


Figure 8. Information content for Method A compared to Method B for chromosomes 1, 2 and 3. Plots on the left side are for Method A while plots on the right are for Method B. The plots for all chromosomes are given in Appendices 1.2 and 1.3.

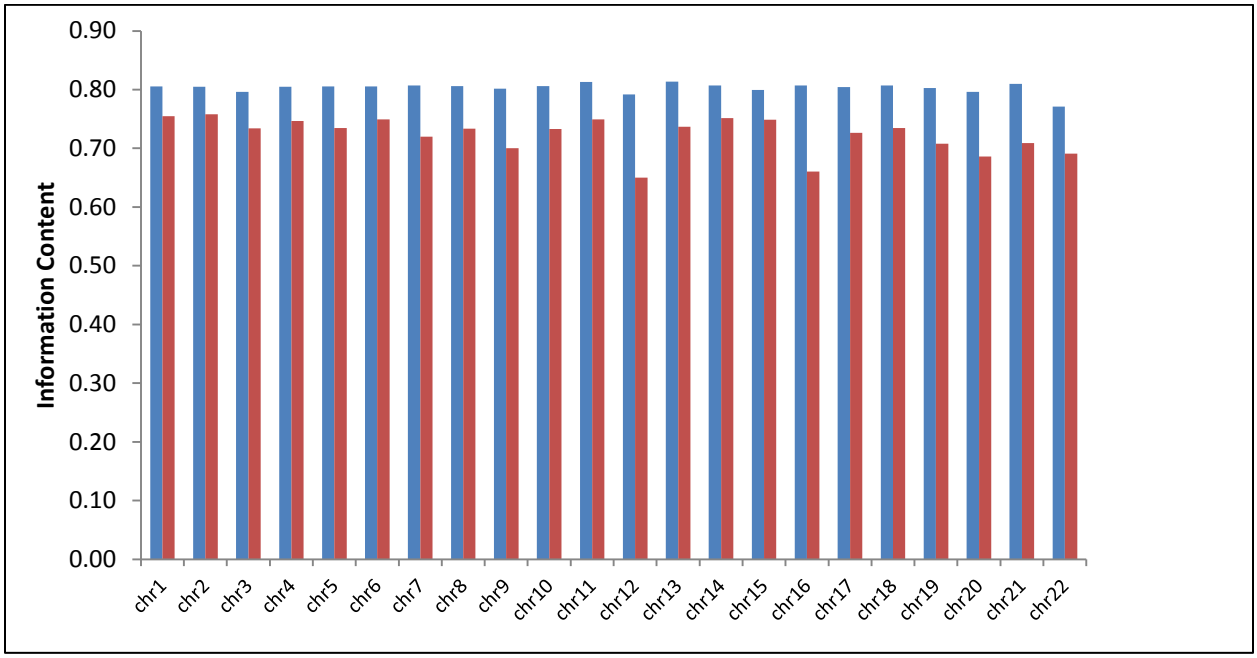


Figure 9. Information Content (IC) for Method A compared to Method B. Method A is shown in blue while Method B is shown in red.

2.8.1.1.3 Non-parametric Linkage Analysis

To test for co-segregation of marker loci with obesity, as a discrete trait, nonparametric multipoint linkage analysis (NPL) was carried out using Merlin. NPL evaluates allele sharing among affected individuals without making any assumptions about the mode of inheritance of the disease trait and thus it is a suitable approach for pedigree studies of complex diseases like obesity, in which a number of genetic and environmental factors affect disease etiology (Abecasis et al. 2002). Subjects with BMI of ≥30 were considered as affected (n=34) and subjects with BMI of <25 were considered as unaffected (n=18) while overweight subjects (30>BMI≥ 25) were considered as unknowns (n=11). NPL was carried out using the –npl option for nonparametric linkage and the –rsq of 0.1, which calculates pairwise correlation coefficient (r^2) between neighbouring markers, and creates a cluster of markers with r^2 >0.1 and all intervening markers, since LD between closely spaced markers can inflate linkage calculations.

Age and gender were included as covariates in the analysis and linkage results were expressed as LOD scores, logarithm of the odds (to the base 10). The LOD score provides an estimate of the likelihood that

two loci lie near each other on a chromosome and are thus co-inherited. LOD scores ≥3.3 are considered as significant linkage signals as it means that the odds are more than a thousand to one in favour of genetic linkage (Lander and Kruglyak 1995). The per family analysis option --perFamily in Merlin was applied to obtain information on the contribution of each family to linkage signals. In addition, the simulation option --simulate was used to generate datasets with simulated genotypes but preserving the actual family structure, allele frequencies, marker spacing and patterns of missing data. Linkage analysis on the simulated datasets indicates how often the real linkage peaks might arise by chance, and that is given as empirical P-values. Linkage regions were defined as the region within a drop of 1 LOD score on both sides of the detected linkage peaks.

2.8.1.1.4 Variance Component Linkage Analysis

In addition to NPL, I performed variance component (VC) linkage analysis with ten obesity-related quantitative traits in order to identify quantitative trait loci related to obesity using BMI, fat mass (FM), fat mass percentage (FM%), fat-free mass (FFM), fat-free mass percentage (FFM%), muscle mass, muscle mass percentage, basal metabolic rate (BMR), bone mass and waist circumference. The analysis was performed using the 63 subjects with genotyping data including 34 obese subjects, 18 lean subjects as well as the 11 overweight subjects who were considered as unknowns in the NPL. VC linkage was carried out using the option --vc in Merlin (Abecasis et al. 2002). As VC linkage analysis are designed for normally distributed traits, the normalisation option --inverseNormal in Merlin was applied to perform quantile normalisation on each of the quantitative traits prior to running VC linkage (Abecasis et al. 2002).

2.8.1.2 Run of Homozygosity (ROH) Analysis

Since this study was performed in families from an inbred population, it is expected that some of the obesity susceptibility variants to be in homozygous state. Therefore, ROH analysis was carried out to identify shared ROH regions between affected subjects to help prioritizing candidate linkage regions, for the purpose of identifying potential obesity variants. ROH analysis was performed using the SNP array data since the homozygosity regions are expected to be long enough to be detected by SNP arrays (Kirin et al. 2010) and as there are several established methods for ROH detection from array data.

ROH analysis was performed with Plink using the SNP array genotyping data after being cleaned for SNPs with: >5% missing genotypes, MAF <0.05, deviation from HWE (at significance level of $1X10^{-6}$) and Mendelian errors, giving a total of 1,233,008 SNPss. The parameters for ROH detection was set to 50

SNPs as sliding window size (--homozyg-window-snp 50), allowing 3 heterozygotes (--homozyg-window-het 3) and 15 missing SNPs (--homozyg-window-missing 15) within a window. Homozygosity window threshold of 0.05 (--homozyg-window-threshold 0.05) was used to define a given SNP as in homozygous segment, and a minimum of 300 SNPs (--homozyg-snp 300) with a minimum density of 1 SNP/50kb (--homozyg-density 50) was used to call a final segment as homozygous. In addition, the option (--homozyg-gap 1000) was applied to split a segment in two if two SNPs within a window were too far apart, and the option --homozyg-group was used to obtain pools of overlapping segments. The ROH pools with more obese subjects than lean were then cross checked with the candidate linkage regions. The overlapping linkage and homozygosity regions were prioritized as top candidates but also non-overlapping linkage and ROH regions were still considered as candidate regions of interest.

2.8.1.3 Processing of WGS Data

The raw sequencing reads for the samples were delivered by Illumina along with the complete set of variants in each sample. The average genomic coverage for the 63 samples in my dataset was 39.5. Coverage is the number of time a particular base in the genome has been covered by sequencing reads, and usually multiple reads per base are needed to generate reliable base call. The Illumina pipeline for processing the raw sequencing reads (average of 659,841,529 reads per genome) consisted mainly of the Illumina tool ELAND for alignment of sequence reads against the human reference genome (NCBI build 37) and CASAVA -1.9.0al_110909 pipeline (Illumina) for SNP calling. With CASAVA, the SNPs and genotypes were called for each sample individually (single-sample calling). In this study, the raw sequencing data were processed using the in-house WGS pipeline implemented in the Bioinformatics core at WCMC-Q, which consists mainly of Bowtie2 (Langmead and Salzberg 2012) for alignment and Genome Analysis Toolkit (GATK) (McKenna et al. 2010), developed by the Broad Institute, for variant calling.

The GATK pipeline is being used extensively in academia and in big research projects such as the 1000 genome project (1kGP) (Clarke et al. 2012). The advantage of using an open source WGS pipeline like GATK over commercial WGS pipelines like CASAVA is that it is highly customizable. However, it requires expertise to set-up optimally. In addition, the use of the GATK pipeline allows performing multi-sample SNP calling rather than single-sample calling, in order to distinguish whether non-variant genotypes are homozygous reference genotypes or are missing genotypes (Kumar et al. 2014).

Here, I used the multi-sample calling approach which was also used in the 1kGP (Clarke et al. 2012). This was done in collaboration with my colleague Dr Pankaj Kumar who implemented the GATK pipeline at the Bioinformatics core at WCMC-Q. Based on my families' data and data from other research groups at WCMC-Q, we found that the GATK multi-sample calling identifies more variants than CASAVA single-sample calling (Kumar et al. 2014). However, both pipelines had highly similar performance in terms of calling putatively causative variants (Kumar et al. 2014). The downstream analyses presented in this thesis are using the in-house WGS pipeline which implements the GATK multi-sample calling approach.

2.8.1.3.1 In-house WGS Pipeline

The Fastq files, which contain the raw short sequence reads, were aligned to the human reference Genome Build 37 using Bowtie2 (Langmead and Salzberg 2012). This generated the sequence alignment map (SAM) files which were compressed into binary alignment map (BAM) files. The BAM files were sorted and indexed using Novosort tool for multi-threaded sorting and indexing. This was followed by removal of duplicate reads, multiple read pairs with identical coordinates that might represent PCR bias, and retaining the pair with the highest mapping quality, using Picard tools (available at http://picard.sourceforge.net/command-live-overview.shtml). The aligned BAM files were then locally realigned around insertions and deletions (indels) to correct for any misalignment that may occur due to their presence. This was done using the GATK walker Realigner Target Creator and Indel Realigner (McKenna et al. 2010). The base quality Phred scores, which evaluate the quality of individual bases in DNA sequences, were recalibrated in the realigned BAM files using GATK Base Score Recalibrator and Print Reads (McKenna et al. 2010). This recalibration step aims to correct for variation in Phred scores with machine sequencing cycle and sequence context giving more accurate scores. After that, the recalibrated BAM files were used as input for performing genotype calling with GATK Unified Genotyper (McKenna et al. 2010).

The obtained variant call format (VCF) files were recalibrated using GATK Variant Score Recalibrator to assign a well-calibrated probability to each variant call (McKenna et al. 2010). The average number of called SNPs per genome was 3,884,771 SNPs. After recalibration, the VCF files were annotated using VariantAnnotator (McKenna et al. 2010) and SNPeff (Cingolani et al. 2012) to get information about the variants including expected impact of the variants, amino acid change, encoded gene, transcript information and others. For storage purposes, the aligned BAM file before calibration, the recalibrated BAM file and the annotated VCF files were stored while the other intermediates files were deleted. The whole genome sequencing analysis pipeline is shown in Figure 10.

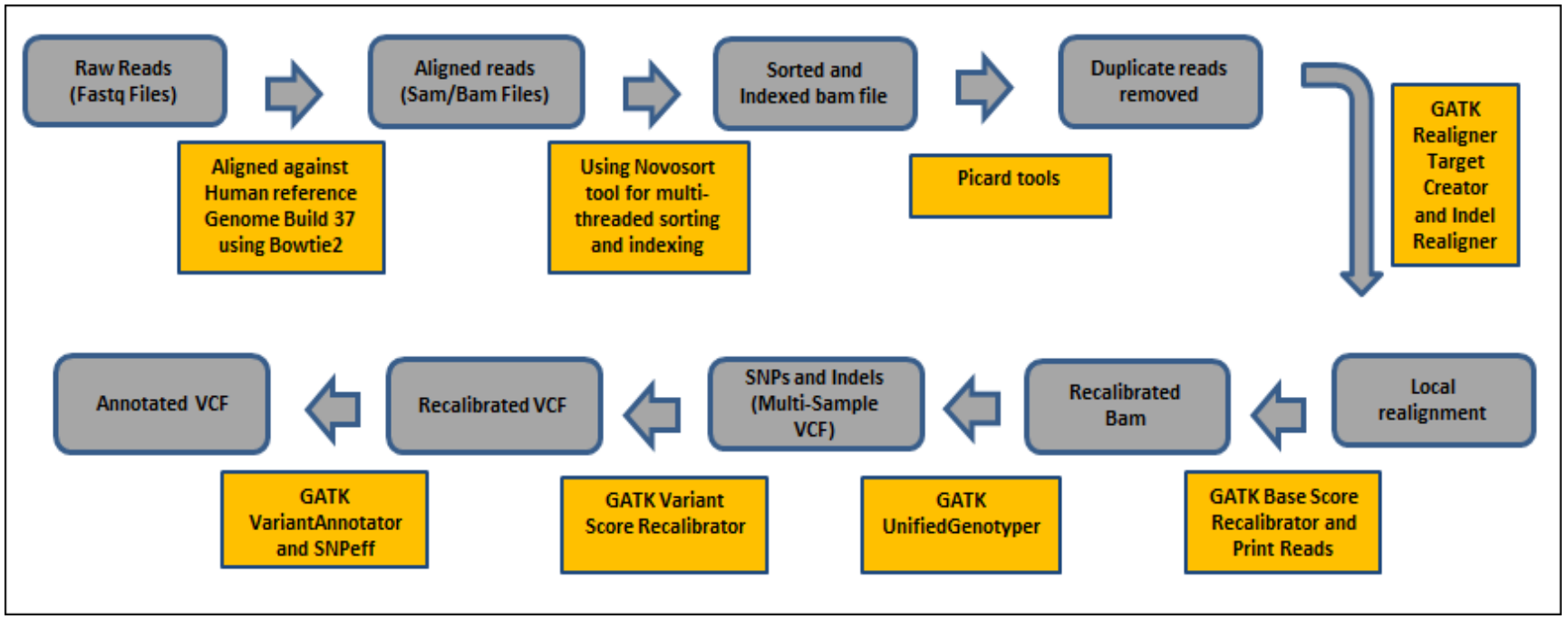


Figure 10. Whole genome sequencing analysis pipeline. Grey boxes shows the input and output files from each step, while orange boxes shows tools used in each step.

2.8.1.3.2 Quality Control of WGS Data

Three major QC steps were performed on the WGS data to ensure good quality of data before using them in downstream analysis to draw biological conclusions.

The first QC was carried out at the Fastq level using FastQC, which is a java application available at (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). FastQC was used to investigate the Phred scores across all bases for each of my study subjects, using the Fastq files as input. All samples had good overall Phred scores. An example of the plots generated by FastQC to assess Phred scores is provided in Figure 11. In addition, FastQC was used to check the guanine-cytosine (GC) content of the sequenced genomes since the deviation from the expected average of around 40% can be indicative of sample sequencing failure. All of the samples had GC distribution close to the theoretical distribution. Figure 12 shows that the GC distribution for one of my study subjects, as an example.

The second QC step was performed at the Bam files level, using QPLOT (Li et al. 2013), to assess the quality of sequencing and alignment data. QPLOT software is developed using C++ language (available at http://genome.sph.umich.edu/wiki/QPLOT#Where_to_Find_It) and generates a series of diagnostic metrics for run quality and produces graphical summaries for these metrics. As illustrated in Figure 13, the distribution of Phred scores has improved after bam file recalibration.

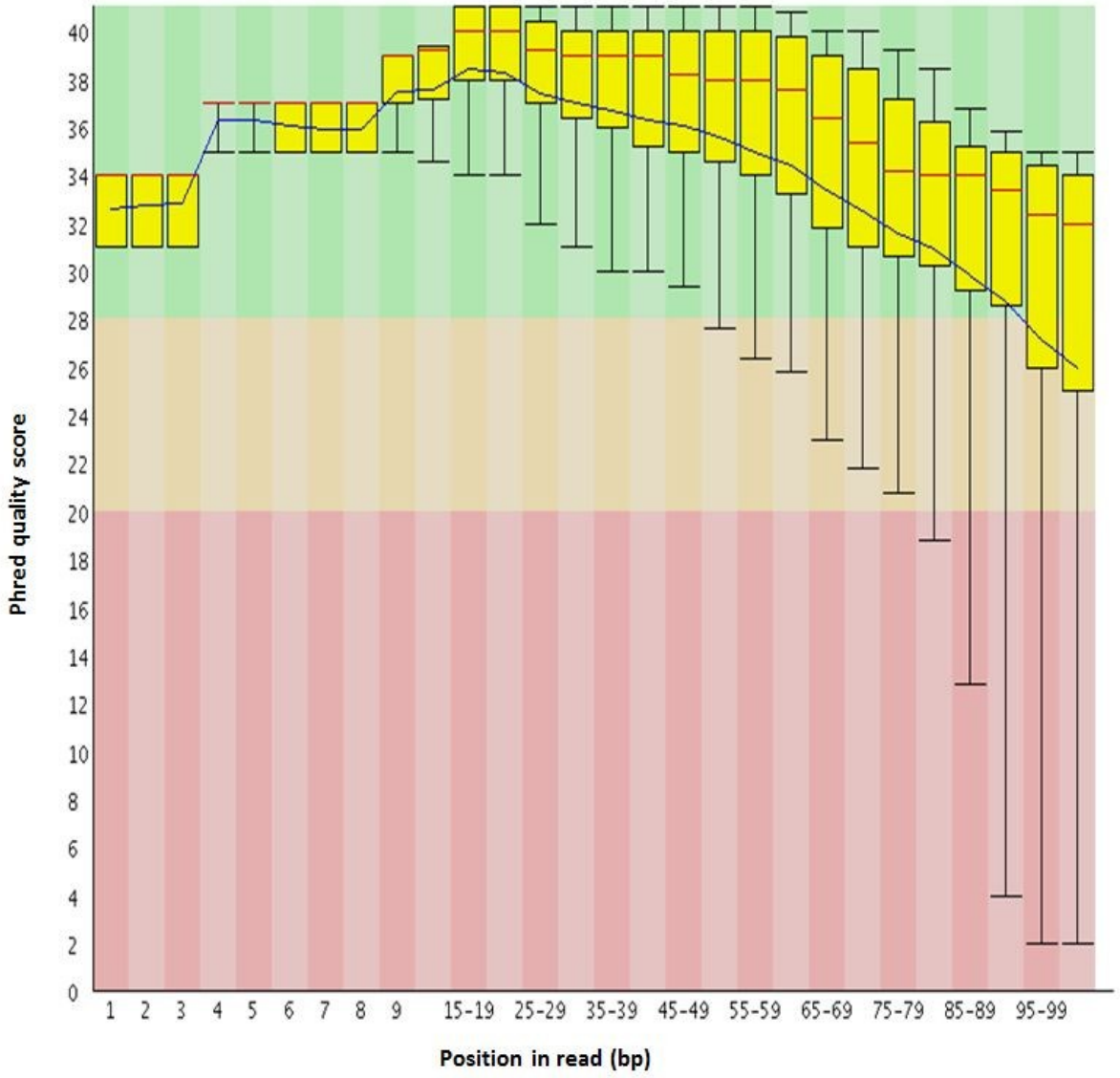


Figure 11. Phred scores across all bases for one sample (QO_F9_01). X-axis shows position of base in the sequencing reads. Y-axis determines the Phred score. The yellow boxes represent the inter-quartile range (25-75%) of Phred scores, the central red lines represent the median scores, the upper and lower whiskers represent the 10% and 90% points and the blue line represents the mean scores.

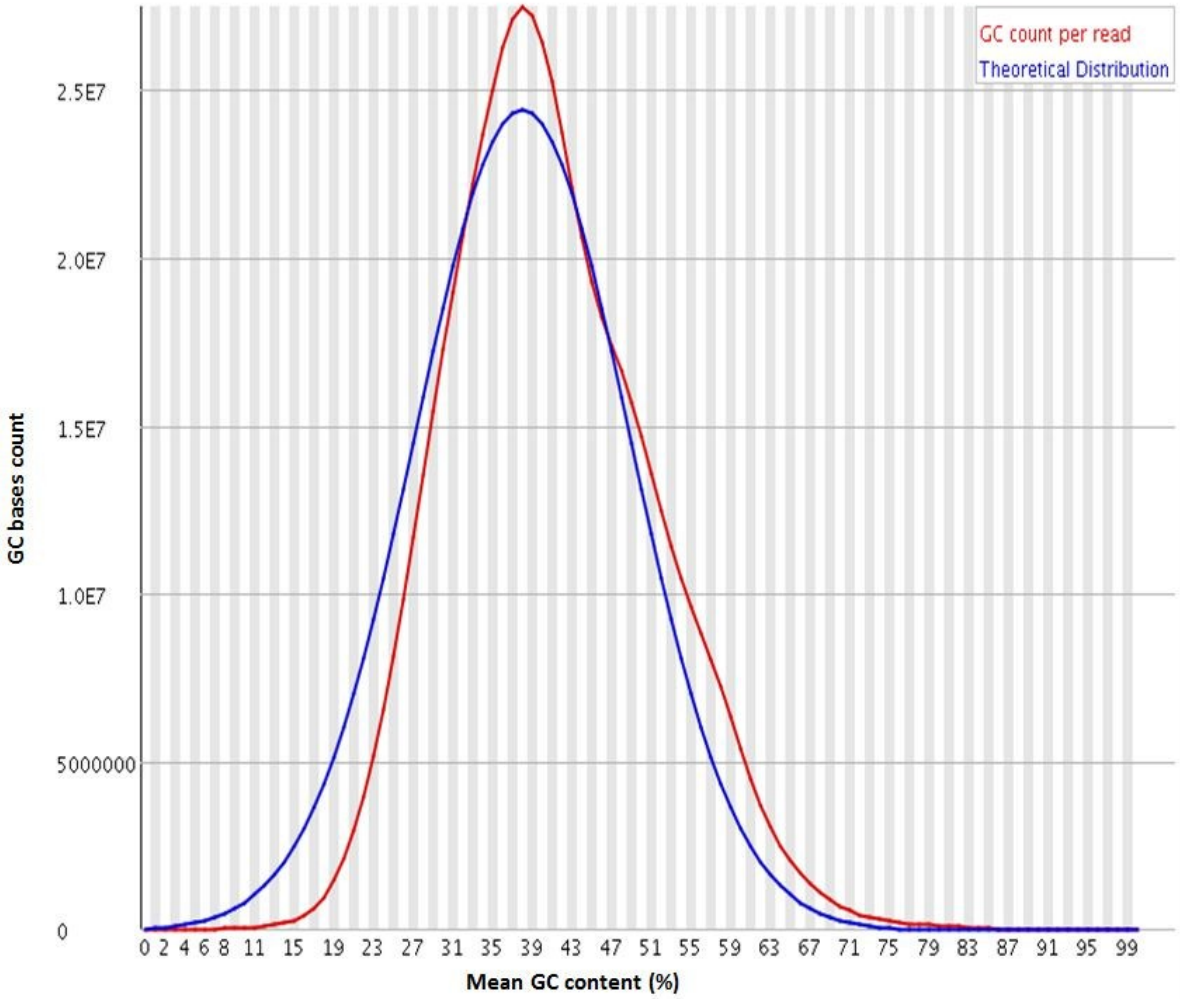


Figure 12. GC distribution over all sequences for one sample (QO_F9_01). Red line represents the distribution of GC count per read while the blue line represents the theoretical GC distribution. Little shift from the theoretical distribution is observed due to systematic bias. The peak shows the overall GC content which is about 38 to 39%.

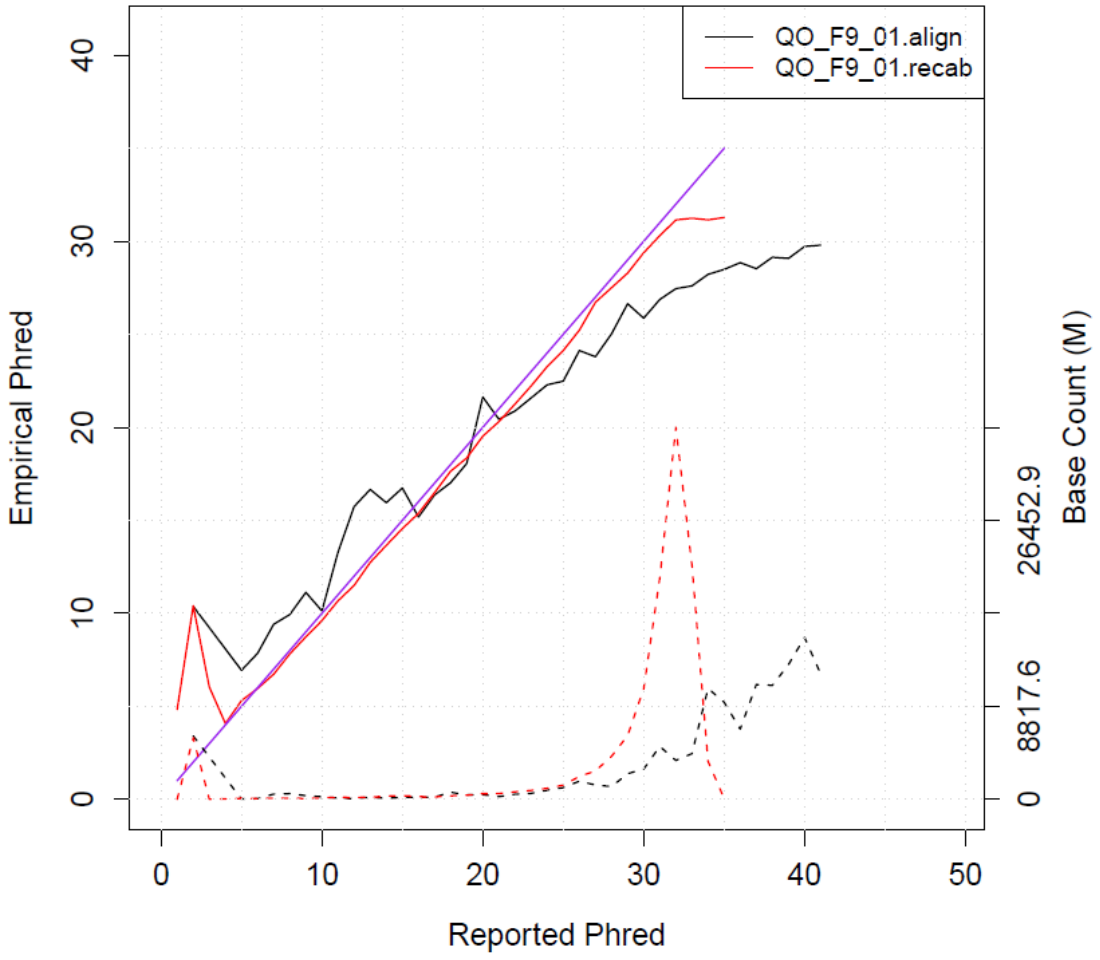


Figure 13. Improvement in reported Phred scores after performing BAM file recalibration, for one sample (QO_F9_01). The purple line shows the empirical Phred scores through sequencing runs. Reported Phred scores after recalibration (red line) are closer to the empirical Phred scores than before recalibration (black line). Also, the distribution of Phred scores for sequenced bases after recalibration (dashed red line) peaked around 30, while it was more scattered before recalibration (dashed black line).

The third QC was to estimate the concordance between the genotyping calls obtained from the HumanOmni2.5-8 arrays and the genotypes obtained from WGS for my study subjects. For the array genotyping data, genotypes with a GenCall score ≥0.15 were used for the concordance test. The GenCall score is a quality score calculated for each genotype in GenomeStudio and it ranges between 0 to 1 with a lower value indicating further deviation of a sample from its genotyping cluster (Illumina 2014). I used the concordance (calculated as a percentage of matching number of genotypes out of the total number of genotypes for each subject) to estimate the overall percentage of concordance per chromosome, as shown in Figure 14. The high percentage of concordance obtained reflects the accuracy of the WGS pipeline and the genotypes obtained using these two genotyping methods that are used in the downstream analyses of this study.

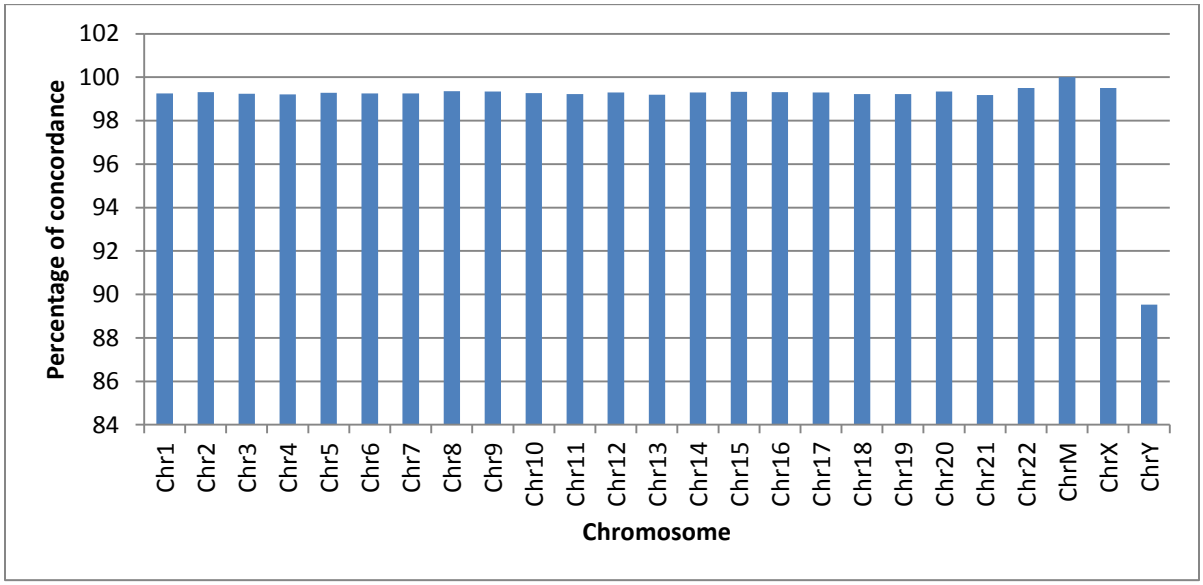


Figure 14. High concordance between the genotypes obtained from SNP-array genotyping and WGS.

2.8.1.3.3 Ingenuity Variant Analysis

Ingenuity variant analysis (IVA) (Ingenuity® Systems, www.ingenuity.com) was used to facilitate the search for potential rare obesity variants with large effect on the development of obesity. IVA is a commercially available web-based application that allows interactive filtering of variants detected through WGS or WES to subsets of most compelling variants in order to be further studied. IVA incorporates evidence from Ingenuity Knowledge Base, which integrates information about biological interactions at various levels including proteins, genes, cells, drugs and diseases. IVA also integrates some external analytical tools such as Sorting intolerant from tolerant (SIFT) (Ng and Henikoff 2003) and Polymorphism Phenotyping (POLYPHEN-2) (Adzhubei et al. 2010). Both tools predict the effect of nonsynonymous variants based on the degree of sequence homology and amino acid conservation cross species. POLYPHEN2 also combines three dimensions structural feature of proteins when available. The predictions given by these tools can help prioritizing variants of interest.

As discussed in Section 2.8.1.3, VCFs were generated through the mutli-sample GATK calling pipeline. A single multi-sample VCF was then created using VCFtools (available at http://vcftools.sourceforge.net/) and uploaded into IVA. The samples were classified as cases (n=34) and controls (n=17), and the pedigree information of the samples were also uploaded. IVA automatically used NCBI37/hg19 as the reference genome since the variants in the VCF were derived from aligned sequence reads to this reference genome. A total of 16,470,959 variants present in the VCF were annotated in IVA.

Four major filters were applied on the 16,470,959 variants (present in 21,999) to narrow them down to the most promising potential variants. The first filter was the confidence filter, by which variants that have at least a Phred score of 20 in at least one case or control were only considered. By applying this default setting, none of the variants were dropped off. The second filter was a frequency-based filter to keep only rare variants that have MAF <1% in the 1000 Genome project, the public Complete Genomics genomes or the NHLBIESP exomes, or variants that are not available in these datasets, using the built-in frequency filter in IVA. A total of 6,786,327 variants in 20,966 genes remained after applying this filter. Although the frequency of the variants in these datasets might not necessarily reflect their frequency in the Qatari population, I used these figures due to the small sample size and the family structure in my dataset. The third filter was the predicted deleterious filter, which was applied to keep variants that are expected to have phenotypic effect, or associated with gain of function of a gene, or associated with loss of function of a gene including missense, frameshift, in-frame indel, and start/stop codon. These filtering steps gave a total of 2,699 variants in 1,877 genes. These variants were further prioritized using the

"biological context" filter, which allows the identification of variants that affect genes relevant to a particular phenotype or disease such as obesity. In IVA, the list for obesity relevant genes contains 719 genes so far. Applying this filter gave 89 variants in 53 obesity-relevant genes.

These 89 variants were further analysed using two approaches to identify variant of potential interest. The first approach was to focus on variants in known monogenic obesity genes, which are *LEP, LEPR, POMC, MC4R, SIM1 and PCSK1*. While, the second approach was to focus on variants located within the candidate obesity regions identified, including: 1) overlapping regions between linkage and ROH, 2) linkage regions, and 3) ROH regions. The number of cases and controls with a particular variant, the gene function and the relevant literature were used to prioritize most promising variants.

2.8.1.3.4 PCR-based Validation of Candidate Obesity Variants

2.8.1.3.4.1 Primer Design

Primers for polymerase chain reaction (PCR) were designed using Primer3 (Rozen and Skaletsky 2000) (available at http://bioinfo.ut.ee/primer3-0.4.0/), to validate four candidate obesity variants. The default primer design parameters were used, with primer size between 18 and 23bp, the annealing temperature (Tm) between 57 and 62°C, and GC content of 30-70%. Primers were checked for specificity the UCSC In-Silico PCR (available at http://genome.ucsc.edu/cgi-bin/hgPcr). Primers were ordered from Integrated DNA Technologies (IDT) and suspended in sterile ddH_2O to make 100uM stock primers. PCR primers are shown in Table 6 and PCR reagents mix is given in Table 7.

Gene	Primer ID	Primer sequence (5' to 3')	Product size (bp)	Annealing temperature (°C)	Family
PCSK1	PCSK1_F	CAGATACCAGGAATGGGCTTA	420	59	3
	PCSK1_R	TGCCATTATAATTTTATGGAAAGAAA			
NMUR2	NMUR2_F	CTACGTGGCCATCCTACACC	235	60	9
	NMUR2_R	GAAGGAGGTGACCTGGATGA			
CLOCK	CLOCK_F	TTTGCTTGAAAATAATGTTCTGATCT	500	60	7
	CLOCK_R	CAGAGCAGCTATGACCAAATC			
RESTSAT	RETSAT_F	GCCTCTGGGCAGTTTTCTTA	396	60	4
	RETSAT_R	ACTCTGGGTCCTCAGCATGT			

Table 6. Primers used for PCR-based validation of candidate obesity variants.

Reagent	Volume per reaction (ul)
Forward primer (10 uM) (Integrated DNA Technologies)	0.5
Reverse primer (10 uM) (Integrated DNA Technologies)	0.5
AmpliTaq Gold (Applied Biosystems, cat no: 4398881)	12.5
DNA sample	1ul (of 200ng)
Sterile ddH_2O	10.5

Table 7. PCR reagents mix for the validation of candidate obesity variants.

2.8.1.3.4.2 Touchdown PCR

When performing PCR-based validation of the candidate variants, touchdown PCR was used in order to avoid non-specific PCR amplifications. The basic idea of touch down PCR is to have two rounds of PCR cycles (Korbie and Mattick 2008). The first round has a high Tm (10°C degrees above the initial primers Tm) that is decreased by 1°C in each subsequent cycle (n= 10 cycles). This round will ensure the specific binding of the primers at highest temperature and thus the amplification of the target of interest. The second round of PCR cycles has a lower temperature, which is equal to the initial primers Tm. This will allow further amplifications (n= 35 cycles) of the PCR products generated during the first round.

Touchdown thermo-cycler program:

The 95 °C for 15 minutes
95 °C for 45 seconds
Tm+10 °C for 45 seconds -1 °C /cycle
72 °C for 1 minute
Go to 2 for 45 seconds
95 °C for 45 seconds
Tm for 45 seconds
72 °C for 1 minute
Go to 6 for 29 cycles
72 °C for 10 minutes
4 °C Hold

2.8.1.3.4.3 Gel Electrophoresis

The success of the PCR reactions was assessed using gel electrophoresis. 1% agarose gel was prepared with 1% tris-acetate EDTA (TAE) and 0.1ng/ul of ethidium bromide. 8ul of PCR product mixed with 2ul of loading dye (New England Biolabs, cat:B7024S) were loaded on the gel along with 7ul of 100bp DNA ladder (New England Biolabs, cat: N3231S) to assist determining the size of the PCR product. The gel was run for 45 minutes to 1 hour at a voltage of 90, and was then examined using a gel documentation device (Bio-Rad, France).

2.8.1.3.4.4 Sanger Sequencing

Before performing Sanger sequencing, PCR products were cleaned-up using an enzymatic reaction that eliminates unincorporated primers and Deoxynucleotide Triphosphate (dNTPs). This was done by adding 6ul of ExoSAP-IT (Affymetrix, cat: 78200) to 15ul of PCR product, followed by thermo-cycler incubation at 37°C for 15 minutes and at 80°C for 15 minutes. For each variant, sequencing was performed in both

forward and reverse directions. Samples were prepared by adding 4ul of cleaned PCR product, and 1ul of 25uM primer (forward or reverse). The samples were submitted to the sequencing facility at WCMC-Q. The sequencing was performed using an ABI 373 automated sequencer, and the sequencing data file (*.ab1) were analysed using mutation surveyor software (available at http://www.softgenetics.com/mutationSurveyor.html) to validate the candidate variants.

2.8.2 Identification of Common and Rare Obesity CNVs

The HumanOmni2.5-8 SNP genotyping data that were used for linkage and ROH were also utilized for performing CNV analyses. CNV prediction was carried out using PennCNV (Wang et al. 2007) (available at http://www.openbioinformatics.org/penncnv/) as will be described below. All CNV analyses were limited to autosomes.

2.8.2.1 CNV Prediction using PennCNV

PennCNV (Wang et al. 2007) is a free software tool that employs a Hidden Markov Model (HMM) to predict CNVs from genome-wide SNP genotyping data at kilobase resolution. This program integrates information on LRR and BAF at each SNP, the distance between neighbouring SNPs, frequency of SNPs and optionally family information to generate CNV calls. This section will describe the various steps for CNV prediction using PennCNV, which are summarized in Figure 15.

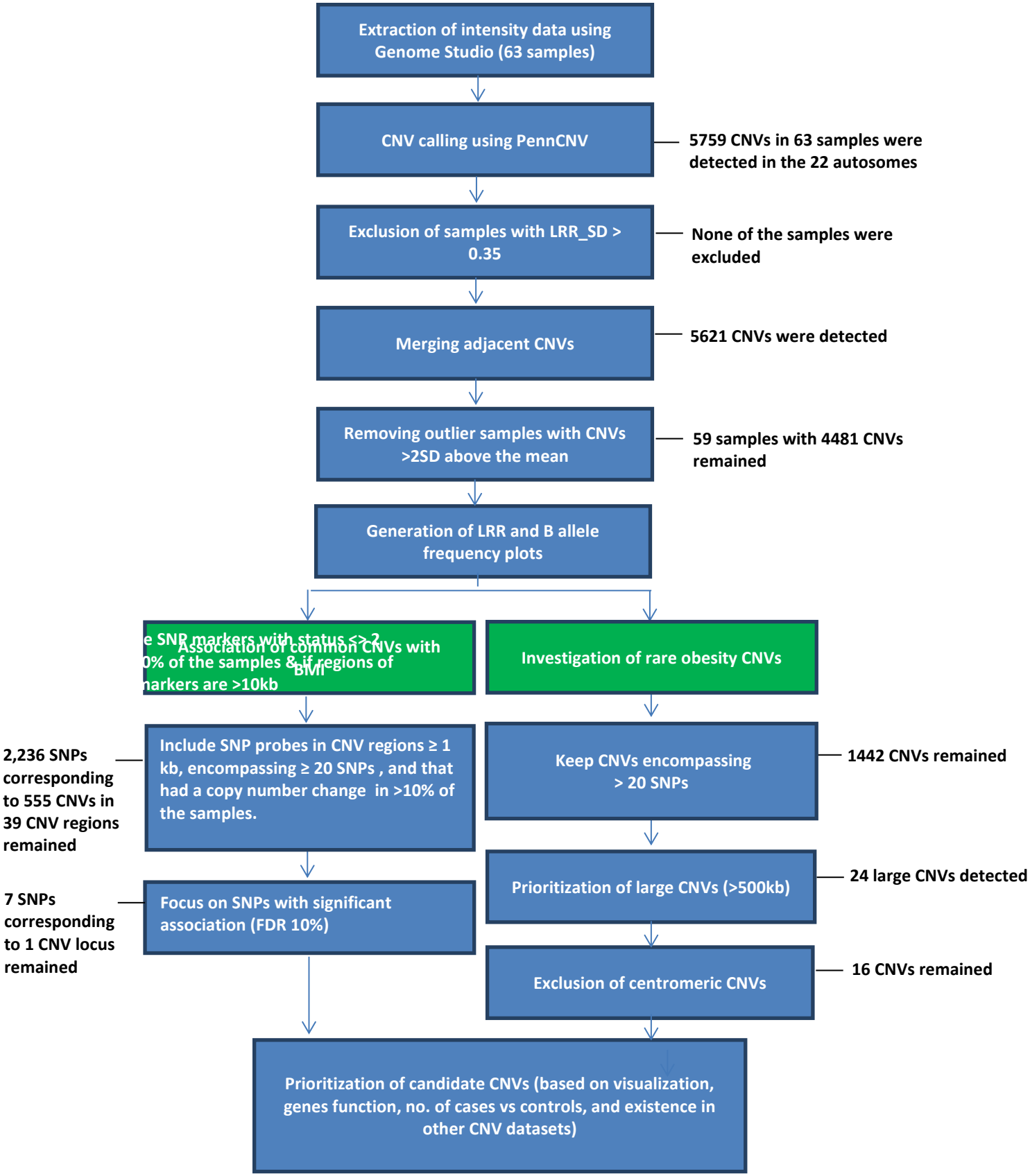


Figure 15. CNV analysis workflow for the identification of common and rare obesity CNVs.

2.8.2.1.1 Extraction of SNP-arrays Intensity Data

The Genotyping Module of GenomeStudio was used to generate and export the intensity file required for performing CNV analyses in PennCNV. This intensity file contained SNP ID, chromosome number, position, genotype, B-allele frequency (BAF) and log R ratio (LRR) information for all samples. LRR is a measure proportional to the total signal intensity, while the BAF is a measure of the relative contribution of the B allele to the total signal intensity. Therefore, LRR reflects the copy number at a particular position while BAF reflects the allelic composition at each SNP. LRR and BAF are measured in GenomeStudio using the formulas: LRR= $\log_2$ (X+Y), and BAF= Y/ (X+Y), where X = signal from the allele A, and Y=signal from allele B. SNP probes within normal copy number states are expected to have an LRR of zero and a BAF of 0, 0.5 or 1, corresponding to the three possible genotypes at each SNP. Analysis of these measures for the probes across the genome is used to identify CNVs (Peiffer et al. 2006).

2.8.2.1.2 CNV Calling

Prior to CNV calling, the intensity file exported from GenomeStudio was split using kcolumn.pl script available in PennCNV package in order to obtain a separate intensity file for each sample, and a signal list file containing the names of the 63 separate intensity files obtained. In addition, a Population Frequency of B allele (PFB) file was generated for the sample set using compile_pfb.pl script available in PennCNV. These files together with the HMM file available in PennCNV, which provides the program with information on the expected signal intensity values for different copy number states, were used for CNV calling.

The calling was performed with a Perl script (detect_cnv.pl) available in PennCNV, which calculates the frequencies of loss and gain of each genetic marker and counts the total number of copies at each marker assuming two copies as the normal state. The calling step generated a report of the raw CNVs (deletions and duplications) detected in the 22 autosomal chromosomes of the samples. Copy number (CN) was presented as CN=0 for homozygous deletion, CN=1 for heterozygous deletion, CN=3 for heterozygous duplication and CN=4 for homozygous duplication. Two copies (CN=2) were assumed for each marker as the normal state and were not included in the CNV report. The total number of raw CNVs detected in the 63 subjects was 5,759 CNVs.

2.8.2.1.3 Quality Controls and Filtering Steps

The first quality control step performed on the raw CNV calls was to check the LRR_SD for each subject in order to identify and exclude low quality samples. According to PennCNV, the LRR threshold applied can be set to 0.24, 0.30 or 0.35 based on the user preference. In my dataset, the LRR_SD ranged between 0.127 and 0.315. The threshold applied was 0.35, therefore, none of the samples (n=63) were excluded based on this QC step. Figure 16 shows the standard deviation of LRR for each of the samples.

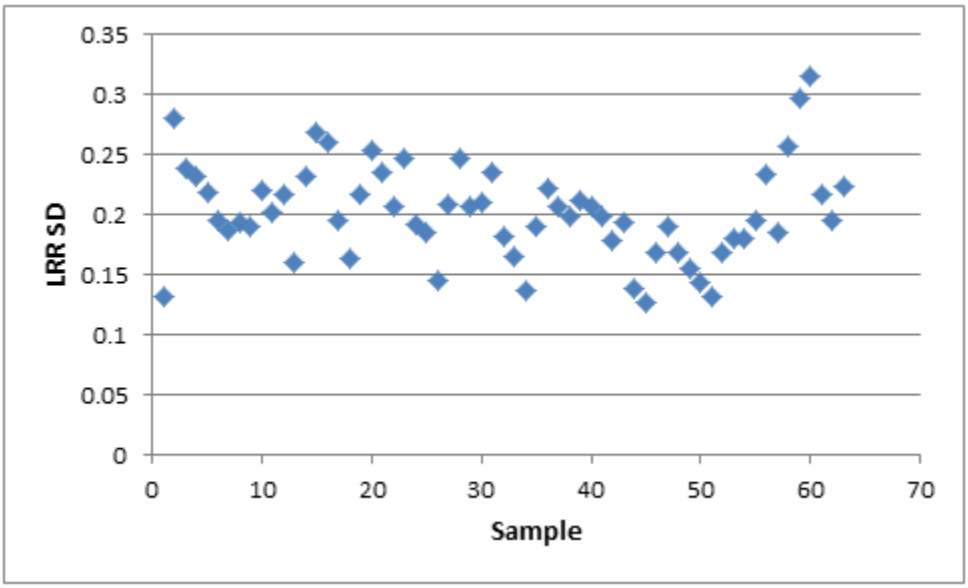


Figure 16. The LRR standard deviation for each of the study samples

Second, the perl script clean_cnv.pl was applied to the called CNVs in order to merge large adjacent CNVs (>500kb) that tend to be split automatically by PennCNV into smaller CNVs (ranged between 100-150kb). I applied the default merging threshold which merges CNVs if the gap between them is less than 20% of the total CNV length. CNV merging gave a total of 5,621 CNVs from the 5,759 CNVs that were originally called.

Third, outlier samples were identified based on the total number of CNVs detected. The number of CNVs detected in each of the samples is shown in Figure 17. The average number of detected CNVs per sample was 89.2 and the standard deviation was 61. Outliers were defined as samples carrying total number of CNVs more than 2SD above the mean. A total of four outliers were detected from the 63 samples and were excluded from further analyses, giving a total of 59 samples with 4,481 CNVs.

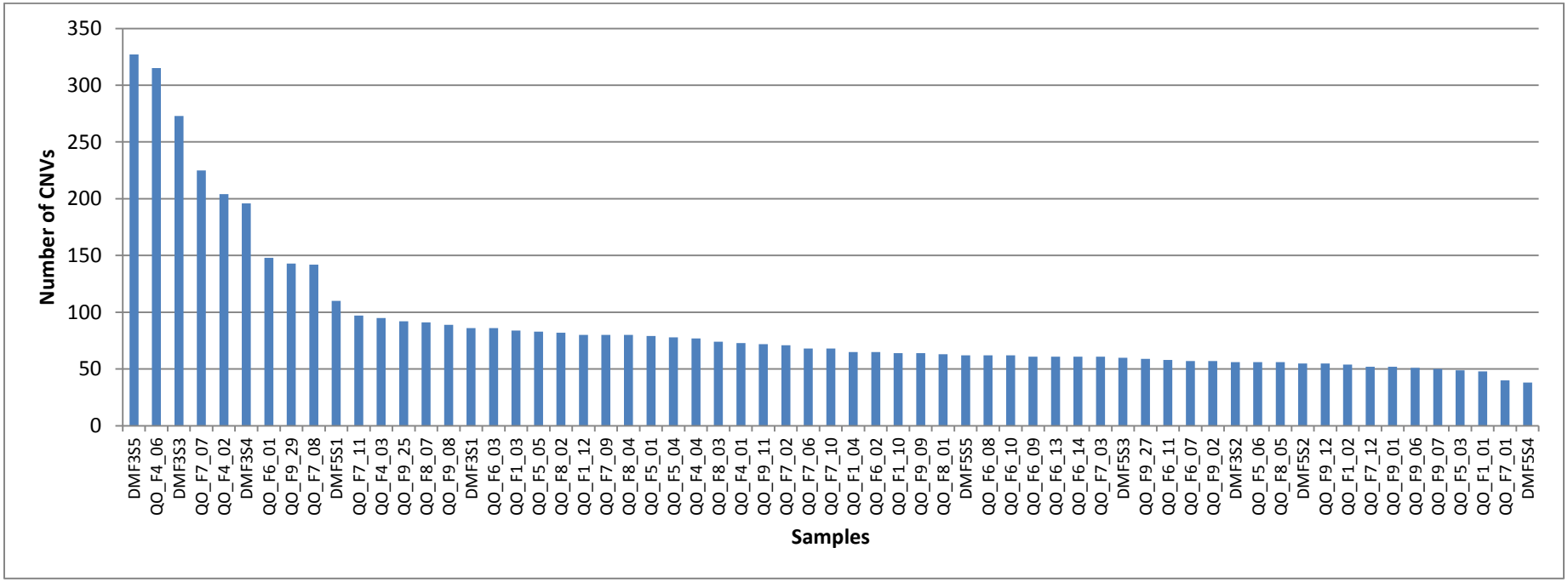


Figure 17. Number of CNVs detected in each of the study samples.

2.8.2.1.4 Identification of Genes Overlapping CNVs

RefGene annotation was used to identify which genes encompassed the detected CNVs. This was done using the scan_region.pl script from PennCNV package and the RefGene annotation files refGene.txt and refLink.txt for build hg19 of the human genome reference, which are available at the University of California, Santa Cruz (UCSC) genome browser (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/).

2.8.2.1.5 Visualization of CNVs

The script visualize_cnv.pl in PennCNV was used to generate LRR and BAF plots of the CNVs. These plots allowed the visual examination of CNV calls for candidate CNVs (at later stages of the analysis) to judge whether these candidate CNVs were reliable or not.

2.8.2.2 Analysis of the 4,481 CNVs Remained after Quality Control

The remaining 4,481 CNVs were analysed further using two approaches as will be discussed below. The first approach was applied for the identification of common CNVs associated with BMI. While, the second approach was to identify potential large rare obesity CNVs, as was summarized by the workflow given in Figure 15.

2.8.2.2.1 Association of Common CNVs with BMI

The first approach was to focus on the identification of common obesity CNVs. SNP probes included in CNV regions ≥ 1kb and encompassing a minimum of 25 SNPs were retained for further analysis. This gave a total of 19,939 SNP markers corresponding to a total of 1,098 CNVs detected in the 59 samples. I then selected SNP markers that that had copy number change (i.e with CN status <>2) in more than 10% of the samples. This filter gave a total of 2,236 SNPs that correspond to 555 CNVs (and 39 CNV regions) detected in the 59 samples. The CNV states of these 2,236 SNP probes were tested for association with BMI. Since PennCNV does not handle extended pedigrees and cannot test for the association of CNVs with quantitative traits, I tested the association between the CNV calls generated by PennCNV and BMI using a linear mixed model that partitions the total phenotypic variance into polygenic and environmental variance. A matrix of the expected proportion of alleles shared IBD across the genome between each pair of individuals in the sample set was used to model the phenotypic variance/covariance. Copy number state was modelled as fixed effect whilst the polygenic effect

between family members was modelled as random effect. Covariates (age and gender) were also included in the model as fixed effects.

A likelihood ratio test was used to calculate significance of association. The test compares the likelihood of the described full model and the likelihood of a null model where the effect of CNVs is constrained to zero. The obtained P-values were then corrected using False Discovery Rate (FDR), which is a widely used statistical method that corrects for Type I error when performing an analysis that involves multiple testing or comparisons (Benjamini et al. 2001; Storey 2002). I applied FDR of 5% as the significance threshold and significant associations were then further investigated.

2.8.2.2.2 Investigation of Rare Obesity CNVs

The second approach was to investigate rare CNVs, CNVs that are present in less than 1% of the population and are more than 500kb in size, for the purpose of identifying rare obesity CNVs. Due to small size of samples (n=63) in my study, I could not assess the rarity of the predicted CNVs in the Qatari population based on the observed frequencies. Therefore, I prioritized CNVs based on the number of SNPs they encompass and their sizes. CNVs that encompass less than 20 SNPs were excluded and this gave a total of 1,442 remaining CNVs. Performing this step can help reducing false positive CNVs in downstream steps and is likely to also retain larger CNVs. After that, large CNVs with a minimum size of 500kb were prioritized, as they are more likely to represent rare CNVs, since large CNVs are likely to affect many genes and thus can be intolerated. A total of 24 large CNVs were detected, and eight of them that were located in centromeric regions were excluded (according to the centromeric coordinates of each chromosome). Centromeric CNVs were excluded because of the very low coverage of SNP probes in the centromeres, due to their highly repetitive DNA sequence, which means that CNVs called in these regions are likely to be false positives. After applying this step, a total of 16 large non-centromeric CNVs remained.

These CNVs were visually examined using their LRR and B allele frequency plots to investigate how reliable they are and were checked in other CNV datasets, including the childhood obesity (CHO) data and the adulthood obesity (ADO) data, the Saudi obesity and Mental retardation data and the 108 Qatari genome data, which were described in Section 2.7, to select potential candidate obesity CNVs.

2.8.2.2.3 Validation of the 16p11.2 deletion using Multiplex Ligation-Dependent Probe Amplification

Multiplex ligation-dependent probe amplification (MLPA) is a widely used technology for the validation of CNVs. The basic idea of MLPA is shown in Figure 18. Briefly, multiple probes are designed to target the sequence overlapping with the CNV region of interest. Each primer consists of two halves, a left and a right half, each carrying a different universal primer sequence (X and Y), that is fluorescently labelled. The right half also contains a stuffer sequence with a unique size for each probe. After denaturation, the two halves of the probe hybridise to immediately adjacent genomic segments, and a ligation reaction is carried out to join the two halves of the probe. Then, the ligated probes are amplified using the universal pair of primers and subjected to fragment separation by capillary electrophoresis. The signal intensities are then normalised against a group of control probes outside the CNV region to infer the copy number state at the region of interest.

In this study, MLPA was carried out to validate the presence/absence of the 16p11.2 deletion in Family 1 members. This was done using the SALSA MLPA P343-B1 Autism-1 probemix kit (MRC Holland, cat: P343-050R), which is designed to target three chromosomal regions at the 15q11-q13, 16p11 and 22q13. The probemix contains 49 probes including 11 test probes that target the 16p11.2 region and nine control probes that target non-copy number regions, given in Table 8. MLPA was conducted according to the MRC Holland protocol. Briefly, DNA Samples were diluted at a concentration of 20ng/ul using TE0.1. Then, 5ul of DNA were used in the MLPA reaction starting with DNA denaturation followed by hybridisation, ligation, PCR amplification and fragment analysis, as described in Table 9.

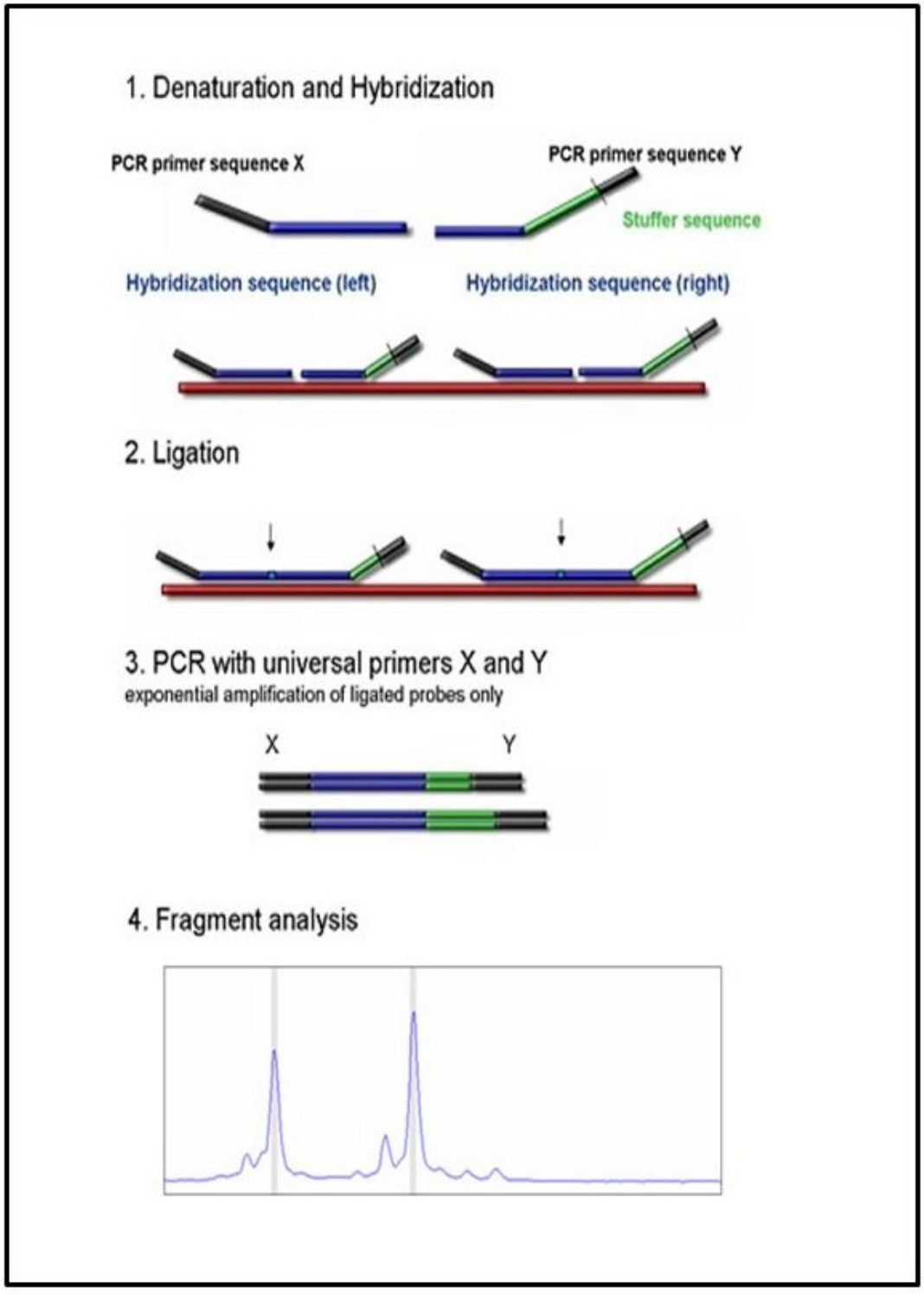


Figure 18. Schematic representation of the Multiplex Ligation-dependent Probe Amplification (MLPA) technology. The figure is modified from MRC Holland website with permission (see Appendix 2.4).

Probe ID	Length (bp)	Target gene	5' half of the primer	3' half of the primer	Start (hg18)	End (hg18)
16p11.2 probes						
11677-L12448	481	*LAT*	CCAAGCTGTGTCTCCTTTACCAGTTTGTAT	CCAAGGGGCATCCAGTTCAAACGGCCTCGTGAGTACAAGG	28,904,653	28,904,722
11672-L12443	364	*SPN*	CAACCTACCAGGAAGTTTCCATCAAGATGT	CATCAGTGCCCCAGGAAACCCCTCATGCAAC	29,582,828	29,582,888
11669-L12440	238	*MAZ*	CAATAGGCTTCACCACGGCAGCAT	ACCTGCGCATCCACGCGGTGAAGGACCACGG	29,728,356	29,728,410
11673-L12444	418	*MAZ*	CCAGAAGGAAAGGAGGAAGAAATGTTT	TCTTAGGGGAATTCGCTAGGTTTTAACGATTTGTTTCTCCTGC	29,729,166	29,729,235
00550-L22423	344	*MVP*	CGGGAAGTGGAGGTCGTGGAGATC	ATTCAGGCCACCATCATCAGGCAGAACCAGGC	29,752,775	29,752,829
11668-L12439	208	*SEZ6L2*	GGAGGAAAGGTGGCAGCCAGATTA	CTTAGAGAGGCACAGAGGAGAGAGATCGGGGTGAGTC	29,817,845	29,817,905
11674-L12445	454	*HIRIP3*	CCAAAGAGGAGAATCCAAGGCGAGCCTCAA	AGGCAGTTGAGGAGAGCAGTGATGAGGAACGGCAGAGGGA	29,913,518	29,913,587
11667-L14670	172	*HIRIP3*	GTGGATGAAGCCGCTTCCAGGGAAGACA	AACTGGACCTTACCAAGAAGGGCAAGAGGCCTCCCAC	29,913,975	29,914,038
13162-L12447	226	*DOC2A*	CCTACGTCAAGCTGCACTTGCTGCCT	GGAGCCTGTAAGGTAACAGCCTCCTCTCGCCC	29,928,004	29,928,061
11675-L12446	465	*MAPK3*	CCTGGCAAGCACTACCTGGATCAGCTC	AACCACATTCTGGGTGAGGGAGGCCCGGCTG	30,036,474	30,036,531
11671-L12442	337	*CD2BP2*	CAGGAGGAGATGGAGGAAGGCCACTT	TGATGCCGATGGCAACTACTTCCTGAACCGGGATG	30,272,761	30,272,821
Control probes						
09176-L09350	124	*TRAF4*	TGCCAAGGTGAGTCCACACTGCCAGGA	AGAAGCCCAAGCACAGTCTGTTGCCCTTGGA	24,098,394	24,098,460
09979-L10438	190	*SMARCA4*	TGGCGGCATTGTCCAAGGGTGAGA	AGCTTCCCAACTGGATGGGGTGGGCAGGTGG	10,998,005	10,998,059
07208-L06858	409	*GHRHR*	CCATAGCATCTCTATTGTAGCCCTCTTCGT	GGCCATCACCATCCTGGTTGCTCTCAGGTTTGTCATC	30,977,299	30,977,367
09431-L09680	154	*RBM14*	TGCCACCTTCGACTACCAGCAGGCTTT	TGGCAACAGCACTGGTGGCTTTGATGGGCAA	66,148,469	66,148,526
05297-L04685	328	*PROS1*	CAGGAGATTTTGAATGTGAATGCCCC	GAAGGCTACAGATATAATCTCAAATCAAAGTCTTGTGAAGGTAGGAT	95,102,326	95,102,403
02658-L02125	250	*ATM*	CTATATGTAGAGGCTGTTGGAAGCTGCTT	GGGAGAAGTGGGTCCTATAGATTTCTCTACCATAGCTATACAACATA	107,675,640	107,675,716
09727-L10077	130	*NOS1*	TTGTAAGCGCGGACAGACACTGCTGAACCT	TTCCTCTGGGACCCCCTGTGGCCCTCGCTCT	116,137,405	116,137,465
10243-L02149	265	*SH3BP5L*	GCTGCACAACGCTGCTCGAGAAATGGTGT	TTGTGGCTGAGCAGGGCGTCATGGCTGACAAGAACCGACTGGA	247,075,309	247,075,380

Table 8. The 16p11.2 probes and the control probes of the SALSA MLPA P343 Autism-1 probemix kit.
Information source: MRC Holland, SALSA MLPA P343 Autism-1 probemix product description.

	Reaction (X1)	**Thermo-cycler condition**
Denaturation	5ul of DNA (at 20ng/ul)	95°C for 5 minutes 25°C – pause
Hybridisation	Hybridisation mix: 1.5ul of SALSA MLPA buffer 1.5ul of SALSA MLPA Probemix	95°C for 1 minute 60°C – pause
Ligation	Ligation mix: 25ul ddH_2O 3ul of Ligase buffer A 3ul Ligase buffer B 1ul SALA Ligase-65	45°C –pause **Add 32ul of ligation mix** 54°C for 15 minutes 98°C for 5 minutes 20°C – pause
PCR amplification	PCR mix: 7.5 ddH_2O 2ul of SALSA PCR primer mix 0.5 ul SALSA Polymerase	35 cycles of: 95°C for 30 seconds 60°C for 30 seconds 72°C for 60 seconds 72°C for 20 minutes 15°C – pause
Fragment analysis	Sample: 0.7ul of PCR product (undiluted products and diluted products of 1:4, 1:8 and 1:16 dilutions) 0.3 ul of Rox 500 10 ul of HiDi fromamide	**Incubate for 3 min at 86°C, cool for 2 min at 4°C.** Injection voltage:1.6 kV Injection time: 15 seconds Run voltage: 15 kV Run time: 1800 seconds Oven temperature: 60°C

Table 9. MLPA reaction mix and thermo-cycler conditions

Fragment analysis was performed by the Genomic Core Facility in the faculty of medicine at Imperial College. The samples were run on an ABI 3730xl capillary electrophoresis instrument (Applied Biosystems) that uses DS-30 as the standard dye set. The analysis of the MLPA data was performed using Coffalyser.Net (available at http://coffalyser.wordpress.com/) by following the instructions given in their manual. Briefly, the default parameters for the capillary system ABI 3730xl were selected and the filter set, which defines the fluorescent dye which the machine recognizes, was set to 'D' since in the experiment a FAM label was used for probe labelling and ROX500 as size marker (GS500-250). The experiment type was set to "DNA-MLPA" analysis and the Coffalyser work sheet for the SALSA MLPA P343-B1 Autism-1 probemix kit, which contains the probes information, was selected. Then, the fragment data traces files (*.fsa), which contain the raw signal data, were imported and the sample type was defined either as sample (for five test samples), reference sample (for three samples expected to have normal copy number state), or no DNA (for three no DNA samples that were processed to ensure that no reagents were contaminated). Thereupon, the fragment analysis was run using the default "auto-bin" as the probe recognition method. This method creates a common probe length reference vector across the samples based on the predicted probe product lengths, and matches each peak to

probe ID. Comparative analysis was then carried-out involving two normalisation steps: 1) intra-individual normalisation by dividing the peak height of each of the test probes by the mean peak height of the nine control probes. This step is important to control for probe bias due to probe properties, and 2) inter-individual normalisation by dividing the normalised peak heights (from step 1) by the mean of the corresponding probes in the reference samples. The final ratios for each test sample were visualized to facilitate the detection of the presence or absence of the 16p11.2 deletion.

2.8.3 Identification of Differentially Methylated Sites with BMI

As discussed in Section 2.6.3, genome-wide methylation profiling data was generated for 96 subjects using the Infinium HumanMethylation450 Beadchip. My DNA methylation data was analysed together with another DNA methylation dataset on Qatari families that was generated by my colleague Wadha Al Muftah, as described Section 2.7.4. The combined dataset consists of a total of 123 subjects from 16 Qatari families. The methylation data was extracted, quality controlled and then analysed for differential methylation with BMI, as described below.

2.8.3.1 Extraction of DNA Methylation Data

First, the methylation module of GenomeStudio (Illumina) was used to pre-process the raw DNA methylation intensity data, according the manufacturer's recommendations. The total number of subjects was initially 123 and total number of methylation probes was 485,577 (including 482,421 CpG probes, 3091 CpH probes (H = A/C/T) and 65 SNP genotyping probes).

For each interrogated CpG site, GenomeStudio assigns a beta-value (b-value), which is a widely used quantitative measure for DNA methylation levels calculated as the ratio of the methylated probe intensity to the total intensity. B-values range between 0 and 1, where 0 shows absence of methylation and 1 shows complete methylation (Du et al. 2010; Marabita et al. 2013). Another method for measuring methylation levels is using the M-value which is simply the logit transform of the b-value (Du et al. 2010). The range of M-values may vary across different datasets and they exhibit bimodal distribution, where the positive mode represents the methylated status and the negative mode represents the unmethylated status (Du et al. 2010). Both methods can be used interchangeably in methylation studies (Du et al. 2010).

2.8.3.2 Quality Control of DNA Methylation Data

Quality control (QC) was carried out on the extracted DNA methylation data in order to reduce technical variability before performing differential methylation analysis and drawing biological conclusions. The QC steps were performed based on the lumi: quantile normalisation (QN) + beta mixture quantile dilation (BMIQ) normalisation pipeline. This pipeline was selected as the most optimal pipeline for pre-processing of Infinium HumanMethylation450 Beadchip data by Marabita *et al.* based on their evaluation of six different analysis pipelines (Marabita et al. 2013).

2.8.3.2.1 Sample Filtering

The first quality step aimed to exclude samples with low percentage of detected sites or with abnormal distributions of intensities and/or b-values. All samples had high percentage of detected sites with an average of 99.5% (with P-value <0.01) and followed the expected data distributions. Therefore, no samples were excluded based on these criteria.

2.8.3.2.2 Probe Filtering

The methylation dataset contained a total of 11,135 probes on the X chromosome and 416 probes on the Y chromosome. These probes were filtered out from the dataset to avoid any potential bias in the downstream differential methylation analysis that may arise due to differences in the proportion of males and females in the dataset. Therefore, the total number of remaining probes became 470,870 probes.

After that, the detection P-value for each probe was calculated in GenomeStudio by comparing the intensity of the probe to the average intensity of the negative control probes (600 probes), as an estimate of the strength of DNA hybridisation over the background. Probes with high detection P-value of >0.01 in more than 5% of samples were discarded as the values obtained from those probes appeared not reliable. A total of 2,398 probes were excluded leaving 468,472 probes.

The existence of SNPs within the query site or probe region has a potential risk of interference with DNA methylation measurements. Since the Infinium assay is based on the quantitative genotyping of C/T SNPs at the CpG query sites after bisulfite conversion, the presence of a C/T polymorphism at the query site may confound methylation measurements (Dedeurwaerder et al. 2014). This means that in the case of a fully methylated site, for example, methylation measurement can approach 1 in subjects carrying C/C genotypes while it will be close to 0 in subjects with T/T genotype and will be around 0.5 in subjects

with C/T genotype (Dedeurwaerder et al. 2014). Thus, the methylation measure is not a true reflection of methylation status of the query site but rather the genotype. SNPs within probe regions might also affect methylation measurements by interfering with probe binding. Therefore, in the Human Infinium HumanMethylation450 Beadchip assay design, Illumina avoided assays that included known SNPs positions (from dbSNP). In 65 instances, regions of interest were covered by assays that included SNP positions and were described in Illumina manifest. These probes were excluded from my dataset giving a total of 468,407 remaining probes. Since WGS data were also available in my study, the methylation values were set to missing if a genetic variant was present within ±110 base pairs of the CpG query sites to avoid potential bias due to SNP existence, and these represented 0.5% of the methylation sites.

2.8.3.2.3 Colour-Bias Adjustment

As discussed in Section 2.6.3, Infinium HumanMethylation450 Beadchip contains two different types of assays. Infinium I assay has two bead types (methylated and unmethylated) for each CpG site and the signals generated are reported by the same colour channel. Infinium I assay represent 30% of the 450k array probes (20% are reported via the red channel and 10% via the green channel). On the other hand, Infinuim II assay has single bead type per CpG locus and the signals are reported by two different colour channels depending on the label colour of the nucleotide added at the final extension step (A or T nucleotides measured on the red channel while C or G nucleotides on the green channel) (Bibikova et al. 2011). Differences in labelling and scanning properties of the two colour channels may result in some inconsistencies in the intensities measured especially in the case of Infinium II assay (Marabita et al. 2013). To correct for this potential colour-bias, a normalisation step of the two colours was performed on the intensity data using the smoothQuantileNormalisation method available in the R package lumi (Du, Kibbe, Lin 2008). Figure 19 shows the reduction in colour-bias between the two channels after performing the colour-bias adjustment step.

2.8.3.2.4 Quantile Normalisation (QN)

Variations in intensity measurements between samples might arise from technical issues such probe type and sample position on the slide. Quantile normalisation (QN), which is available in the lumi package (Du, Kibbe, Lin 2008), was applied to normalise the intensities after colour-bias adjustment, assuming that the distribution of intensities is similar across samples. Figure 20 shows the reduction in between-sample variation after quantile normalisation.

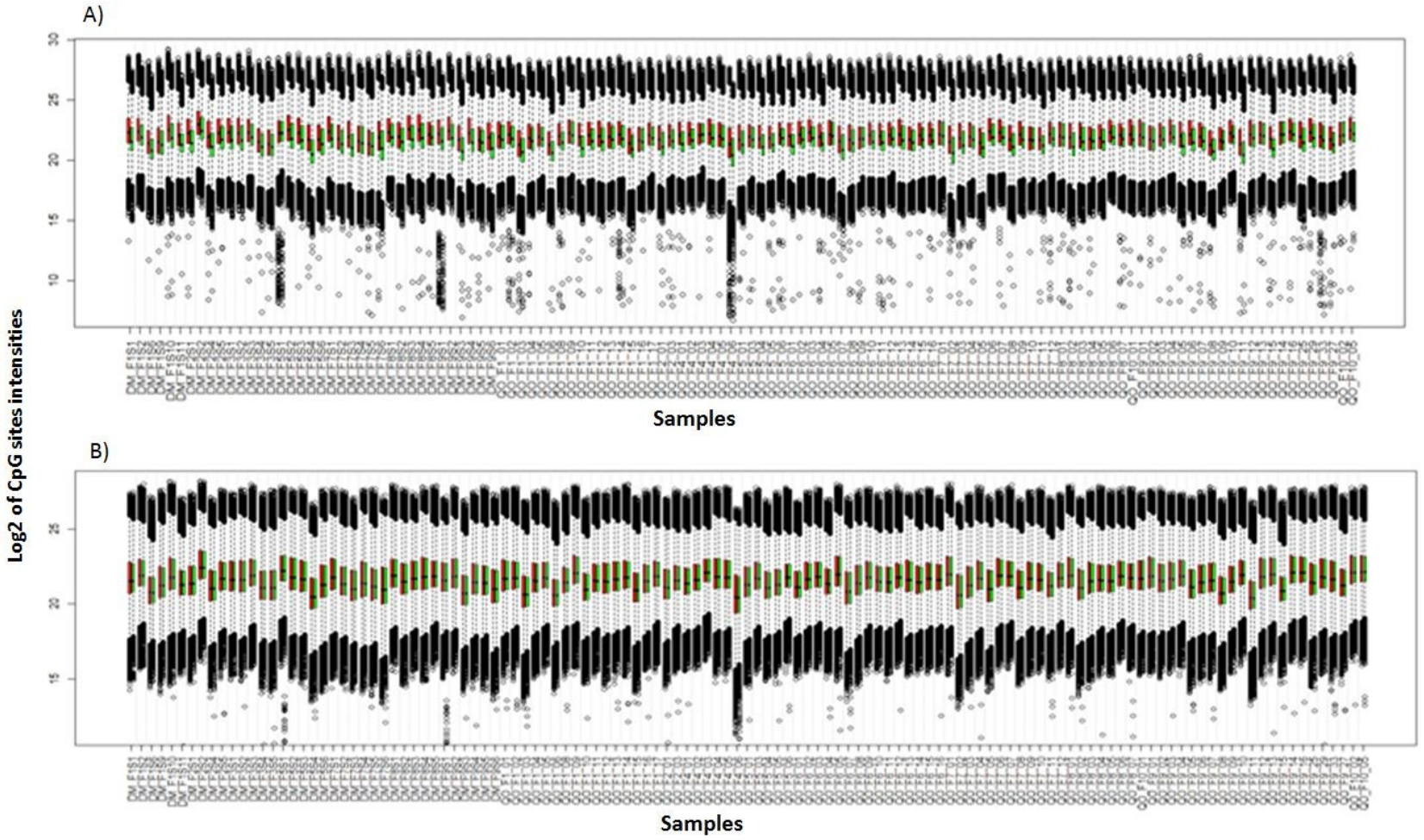


Figure 19. Boxplots of CpG sites intensities for the two colour channels per sample. A) Before colour-bias adjustment, and B) After colour-bias adjustment. The circles represent outliers, and the red and green boxes represent the red colour channel and the green colour channel respectively.

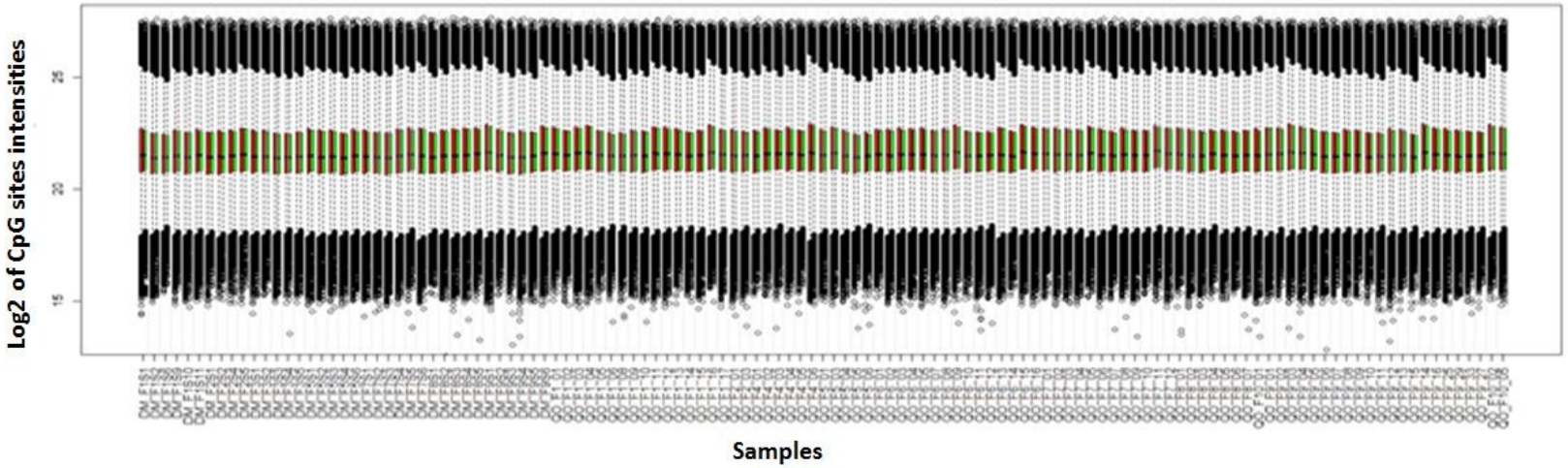


Figure 20. Boxplots of the colour adjusted CpG sites intensities per sample after quantile normalisation. The circles represent outliers, and the red and green boxes represent the red colour channel and the green colour channel, respectively.

2.8.3.2.5 Beta Mixture Quantile Dilation (BMIQ) Normalisation

Due to their different chemistries, Infinium I and Infinium II assays exhibit distinct distributions of b-values. In general, the peak of Infinium I assay at the extreme ends of methylation are usually further out compared to Infinium II, showing that Infinium I assay has an extended dynamic range of measured methylation b-values compared to Infinium II assay. This is due to the fact that Infinium I assay is more likely to be the assay of choice for targeting CpG islands (Zaghlool et al. 2015).

Based on previous experience (Marabita et al. 2013), assay type bias does not sufficiently reduce with quantile normalisation. Therefore, beta mixture quantile dilation (BMIQ) normalisation was applied on the b-values of the quantile-normalised intensity data in order to match the distribution of the two assay types, especially at the extreme ends (Teschendorff et al. 2013). BMIQ is a quantile normalisation-based method that decomposes the density profile of Infinium I and Infinium II assays into two groups of three b-value distributions; unmethylated (close to 0), methylated (close to 1) and hemi-methylated (close to 0.5), and then it applies quantile normalisation to fit each of the b-value distributions of Infinium II to its corresponding b-value distribution of the Infinium I (Dedeurwaerder et al. 2014; Teschendorff et al. 2013). Figure 21 shows the effect of BMIQ normalisation on the b-values distribution.

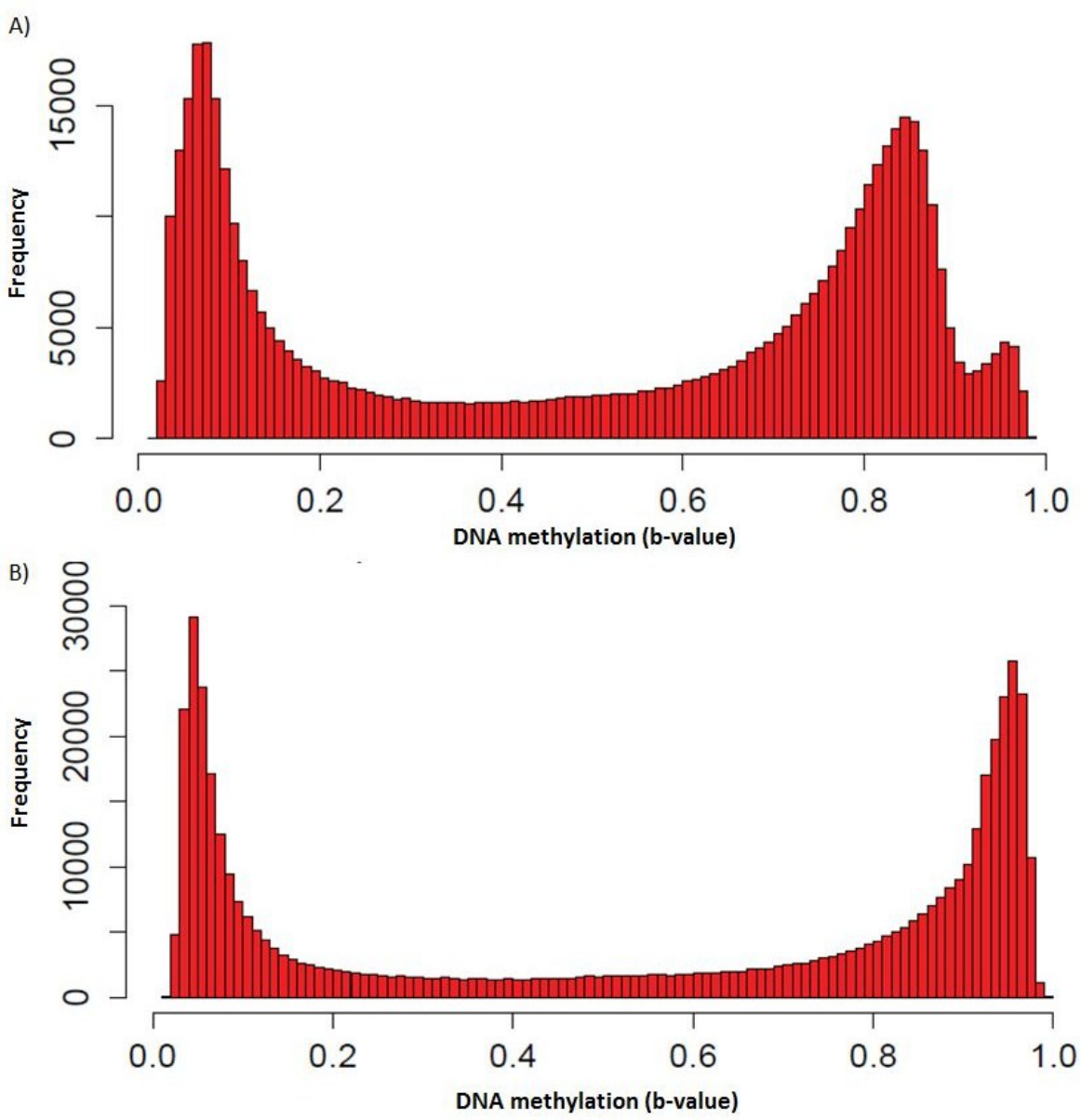


Figure 21. **Histograms of DNA methylation profiles (b-values) for one of the study subjects (QO_F7_09) showing the effect of BMIQ normalisation.** A) Before BMIQ normalisation, and B) after BMIQ normalisation. The extra peak in A) is due to the probe bias. Figure is reproduced from Zaglool *et al.* (Zaghlool et al. 2015).

2.8.3.2.6 Adjustment for Cellular Heterogeneity

Whole blood is composed of a cocktail of different cell subtypes in uneven fractions, and these subtypes carry unique DNA methylation profiles (Houseman et al. 2012; Reinius et al. 2012). This means that differential methylation at some CpG sites between cases and controls might not reflect real differences in DNA methylation patterns but rather reflects changes in the proportion of these subtypes as a consequence of disease for example.

Due to the fact that the methylation profiles in this study were generated from whole blood-DNA, it was necessary to take the potential confounding effect of cellular heterogeneity into account (Michels et al. 2013). One way of quantifying cell type composition is using flow cytometry approach. However, it was difficult to implement this approach in the study since it requires large volumes of fresh blood sample, labour-intensive antibody tagging and timely cell analysis (Zaghlool et al. 2015). Housman *et al.* (Houseman et al. 2012) have recently developed and validated a statistical method called MethylSpectrum implemented in the MethylSpectrum software that permits the estimation of cell type composition at data level rather than blood sample level. This method uses reference information on unique DNA methylation signatures for each of the principle immune components in blood (monocytes, granulocytes, NK-cells, B-cells, $CD8^+$-T-cells, and $CD4^+$-T-cells) to estimate the cell-type distribution in study samples using their individual genome-wide methylation profiles.

Liu *et al.* applied the MethylSpectrum method in their methylation study on rheumatoid arthritis to obtain cell type information (cell type coefficients) from the HumanMethylation450 array data derived from whole blood-DNA (Liu et al. 2013). They showed that adjusting for cell type information in differential methylation analysis has a notable reduction in the association signals obtained, showing the importance of such adjustment in removing false positive signals. Therefore, I applied the MethylSpectrum method, using the MethylSpectrum software (version 2.0), to the methylation data and included the cell type coefficients calculated as covariates in the association model between DNA methylation and BMI.

2.8.3.3 Association of DNA methylation with BMI

Association of DNA methylation levels with BMI was performed using a Java package called Population and Pedigree Association Testing of the Epigenome (PopPANTE), available at FIXME (https://sites.google.com/site/populationgenomics/poppante). PopPANTE uses the DNA methylation data in a variance component framework to model the resemblance among family members. The

association between the phenotypic traits of interest (BMI) and DNA methylation levels was evaluated using linear mixed model, where DNA methylation level was modelled as a fixed effect while the phenotypic variance (environmental and polygenic variance) was modelled as a random effect. A matrix of the expected proportion of alleles shared IBD across the genome between each pair of individuals in the sample set was used to model the phenotypic variance/covariance.

The association model can also be extended to include covariates as fixed effects. In my study age, gender, T2D status and the six cell type coefficients were included as covariates. A likelihood ratio test between the full model and the null model where the effect of DNA methylation values is constrained to zero, was used to evaluate the significance of association. The differentially methylated CpGs were then annotated using the IlluminaHumanMethylation450k.db package in R. Quantile-Quantile (QQ) plot was generated to ensure that the association statistics were not inflated.

2.8.3.4 Selection of CpG Associations with BMI for Replication

I aimed to replicate the major associations from two large EWASs that were published at the time of this study (Dick et al. 2014; Sun.D. et al. 2014). This includes three CpG sites at ATP-binding cassette, sub-family G (WHITE), member 1 (*ABCG1*)**,** carnitine palmitoyltransferase 1A (*CPT1A*) and ring finger protein 145 (*RNF145*) that were reported by Sun *et al.* (Sun.D. et al. 2014), and three CpG sites at the hypoxia inducible factor 3, alpha subunit (*HIF3A*)*,* clustered mitochondria (*cluA/CLU1*) homolog (*CLUH*)*,* Kruppel-like factor 13 (*KLF13*) that were reported by Dick *et al.* (Dick et al. 2014). I attempted to replicate the strongest association for each reported locus. I did not carry out the association only at the selected probes in order to avoid missing any potential inflation of the association statistics that might be due to the limited size of the sample set. Instead, I first performed an EWAS and assessed inflation through QQ-plot and calculation of the lambda inflation coefficient of the EWAS. To correct for multiple testing, associations were considered replicated if the association P-value was below 8.33×10^{-3} (0.05/6).

2.8.3.5 Association of DNA Methylation with BMI in the TwinsUK Cohort

The association of two CpG sites at *ABCG1* and *CPT1A* with BMI was performed on the TwinsUK data. This was carried out by my supervisor Dr Mario Falchi and my colleague Dr Alessia Visconti, who are currently working at Kings College, and have access to the data. For this analysis, probes with detection P-value >0.05 or missing values were removed and blood cell type coefficients were estimated from the methylation data using the Houseman method (Houseman et al. 2012). The association was then

assessed with a linear mixed effect model using also PopPANTE to correct for family and zygosity in the TwinsUK cohort.

Similar to the association model in the Qatari data, the model included DNA methylation level as fixed effect and the phenotypic variance as random effect. A matrix of the expected proportion of alleles shared IBD across the genome between each pair of individuals in the sample set was created using the R package kinship2 (Therneau et al. 2014) to model the phenotypic variance/covariance. The covariates T2D status and the blood cell types coefficients for monocytes, granulocytes, NK-cells, B-cells, $CD8^+$-T-cells, and $CD4^+$-T-cells were included as fixed effects.

2.8.3.6 Meta-analysis of the DNA Methylation association studies with BMI using the Qatari and the TwinsUK data

A meta-analysis between the Qatari and the TwinsUK sample sets was performed for the associations of the two CpG sites at *ABCG1* and at *CPT1A* with BMI. This was carried out using the GWAMA (Genome-Wide Association Meta-Analysis) software (Magi and Morris 2010). The regression coefficients with their standard errors form the Qatar and the UK studies were combined using a fixed-effect model with inverse variance, and the heterogeneity between the two studies was estimated using the Cochran's Q test and by measuring the proportion of variability that is explained by between-trial heterogeneity both implemented in GWAMA.

CHAPTER THREE

Identification of Rare Obesity Mutations

3. Identification of Rare Obesity Mutations

As introduced in Section 1.7.2, there is a growing evidence that rare variants play a role in the susceptibility of obesity and other complex diseases (Blakemore et al. 2009; Bodmer and Bonilla 2008; Creemers et al. 2012; Ichimura et al. 2012; Nejentsev et al. 2009; Pearce et al. 2013; Pritchard 2001; Schork et al. 2009). Therefore, one of the aims of this study was to identify rare obesity mutations that may confer susceptibility to obesity in my sample set.

As discussed in Section 1.7.1, family-based designs offer an opportunity for enrichment of rare disease variants with increased locus specific heritability, which can facilitate the detection of rare disease variants. The strategy I applied was to perform whole genome sequencing in families with multiple affected individuals from the Qatari population and to combine linkage and ROH evidence in these families to narrow down the regions and thus variants of interest.

The linkage and ROH analyses were performed using the SNP genotyping data available for a total of 63 subjects (34 obese, 17 lean and 12 overweight). These subjects are indicated by the "GSM" label in their pedigrees which are given in Figure 2 in Section 2.4. I then selected potential obesity variants based on variant prioritization analyses through Ingenuity Variant Analysis, gene functions and the available relevant literature using the WGS data for the 34 obese and 17 lean subjects. In addition, I validated the candidate variants in the families in whom they were detected using Sanger, and attempted to replicate the identified variants in a population-based cohort from Qatar that is described in Section 2.7.1. Methods that are relevant to this chapter are given Section 2.8.1 while the results and discussion are presented below.

3.1 Identification of Candidate Obesity Regions

3.1.1 SNP-Array Genotyping and Quality Control

SNP array genotyping was performed using the HumanOmni2.5-8 SNP array for 63 samples (34 obese, 17 lean and 12 overweight). The genotypes were extracted from the raw intensities using GenomeStudio (Illumina). All samples had genotyping call rates > 95% and inferred gender was consistent with clinical information. The genotyping data were subjected to a group of quality control steps using Plink (Purcell et al. 2007), Pedstats (Abecasis and Wigginton 2005), Pedwipe and Merlin (Abecasis et al. 2002), for the purpose of identifying candidate obesity regions through linkage and ROH analyses.

3.1.2 Linkage Analysis

As discussed in Section 2.8.1.1, linkage analysis was conducted with 20,872 SNP markers selected using Method A, which is based on markers informativeness and LD pruning, since they provided higher information content (IC) (average IC= 0.80) and higher marker's density (average genetic distance between markers= 0.13cM) compared to Method B, which was based on Illumina's V panel linkage marker set (average IC= 0.72, and average genetic distance 1.7cM).

For the purpose of the NPL for obesity as adiscrete trait, the obese (n=34) and the lean (n=18) subjects were considered while the overweight subjects were set as unknowns (n=11). For the VC analysis for obesity-related traits, all 63 subjects were considered. The results of NPL VC linkage were expressed as LOD scores. LOD scores ≥3.3 are usually considered as significant linkage signals while LOD scores ≥1.9 as suggestive (Lander and Kruglyak 1995), though these criteria may be too strict for this small sample size. None of the signals detected for NPL and VC showed significant linkage. The highest signal for NPL was on 3p14.1 with a LOD score of 1.4. For the variance component analyses, two suggestive signals were obtained at same position on 16p12.3 with a LOD score of 2.3 and 2.1 for muscle mass percentage and fat-free mass (FFM) respectively. In addition, another suggestive signal with LOD score of 2.1 was detected on 10q23.1 with muscle mass percentage. Linkage signals with LOD scores ≥1 were considered as regions of linkage, and the linkage regions were defined as the region within a drop of 1 LOD score on both sides of the obtained linkage signals, as shown in Table 10. The NPL plots are given in Figure 22, and the VC linkage plots for the signals with LOD scores ≥1.5 are given in Figure 23. NPL and VC linkage plots for all of the 22 autosomes are shown in Appendices 1.4 and 1.5 respectively.

Chr.	Trait	Peak SNP	Peak SNP position (cM)	Peak SNP position (bp)	LOD score	Region start position (bp)	Region end position (bp)	Region cytoband	Empirical P-value	Most contributing family	LOD of most contributing family
Nonparametric linkage											
3	Obesity	rs9809725	92.866	69,541,799	1.4	54,807,240	76,126,167	p14.3-p12.3	6.3X10^{-}	6	0.4
13	Obesity	rs259730	51.79	46,845,976	1.3	41,820,714	67,752,445	q14.11-q21.32	6.7X10^{-}	4	0.6
14	Obesity	rs11623206	77.50	76750175	1.1	91,871,744	80,182,967	q22.2-q31.1	1.2X10^{-}	9	0.9
Variance component linkage											
16	Muscle mass	rs1457907	38.034	17,346,236	2.3	13,616,086	25,248,442	p13.12-p12.1	5.6X10^{-4}	7	1.4
16	FFM	rs1457907	38.034	17,346,236	2.1	13,616,086	25,248,442	p13.12-p12.1	9.8X10^{-4}	7	1.2
16	Bone mass	rs1457907	37.793	17,346,236	1.8	13,364,121	26,564,980	p13.12-p12.1	1.8X10^{-3}	1	1.2
16	BMR	rs1457907	37.793	17,346,236	1.3	740,466	49,767,921	p13.12-q12.1	7.6X10^{-3}	7	0.9
10	Muscle mass %	rs5029676	107.377	87,189,235	2.1	77,379,512	93,347,297	q22.2-q23.32	1.0X10^{-}	9	1.2
10	Waist	rs1655155	170.916	87,189,235	1.6	13,074,4348	134,316,622	q26.3-q26.3	3.8X10^{-}	1	0.7
10	FFM%	rs2286741	131.279	115,388,588	1.4	105,517,583	118,359,548	q24.33-q25.3	5.6X10^{-3}	1	0.8
10	FFM%	rs7086375	61.307	35,246,612	1.4	27,851,166	37,066,341	p12.1-p11.21	6.1X10^{-3}	7	0.8
10	Bone mass	rs2817683	117.44	98,838,694	1.0	63,043,987	121,922,699	q21.2-q26.12	1.5X10^{-2}	1	0.6
10	BMR	rs902467	117.739	98,923,531	1.0	63,043,987	121,480,161	q21.2-q26.11	1.5X10^{-2}	1	0.6
13	Waist	rs9550629	51.79	20,881,255	1.4	34,637,818	63,816,136	q32.33-q11.2	5X10^{-3}	1	0.8
14	FFM%	rs10131138	13.193	21,061,167	1.5	107,287,663	23,244,138	q32.33-q11.2	4.4X10^{-3}	1	1.0
14	Fat mass	rs1766142	40.346	36,696,327	1.3	33,442,259	56,418,830	q13.1-q22.3	7.8X10^{-3}	9	1.1
14	FFM%	rs3736824	14.753	21,821,821	1.2	107,287,663	23,244,138	q32.33-q11.2	1.0X10^{-2}	1	1.0
14	Fat%	rs1766142	40.346	36,696,327	1.1	33,442,259	54,081,493	q13.1-q22.1	1.2X10^{-2}	9	1.5
5	FFM%	rs1175072	156.841	150,887,925	1.3	140,372,222	158,131,574	q31.3-q33.3	7.8X10^{-3}	9	0.8
5	Bone mass	rs6556877	108.369	94,952,136	1.1	81,710,048	123,449,266	q14.2-q23.2	1.2X10^{-2}	1	0.7
5	FFM%	rs349682	146.803	143,849,358	1.1	134,769,827	162,600,918	q31.1-q 34	1.3X10^{-2}	9	1.0
18	FFM%	rs6506399	22.387	6,489,772	1.1	4,318,407	33,015,239	p11.31-q12.2	1.2X10^{-2}	9	0.6
18	Muscle mass %	rs4362456	25.144	7,416,263	1.1	4,318,407	33,015,239	p11.31-q13.33	1.2X10^{-2}	9	0.9
20	FFM%	rs6015458	100.463	57,807,095	1.1	52,677,959	59,491,673	q13.2-q13.33	1.1X10^{-2}	1	0.7
20	Bone mass	rs2427386	113.273	61,323,883	1.0	55,555,203	217,281	q13.31-p13	1.6X10^{-2}	1	0.4

Table 10. Summary of the signals with LOD score of ≥1 for NPL and VC linkage. Coordinates are according to GRCh37/hg19 genome assembly.

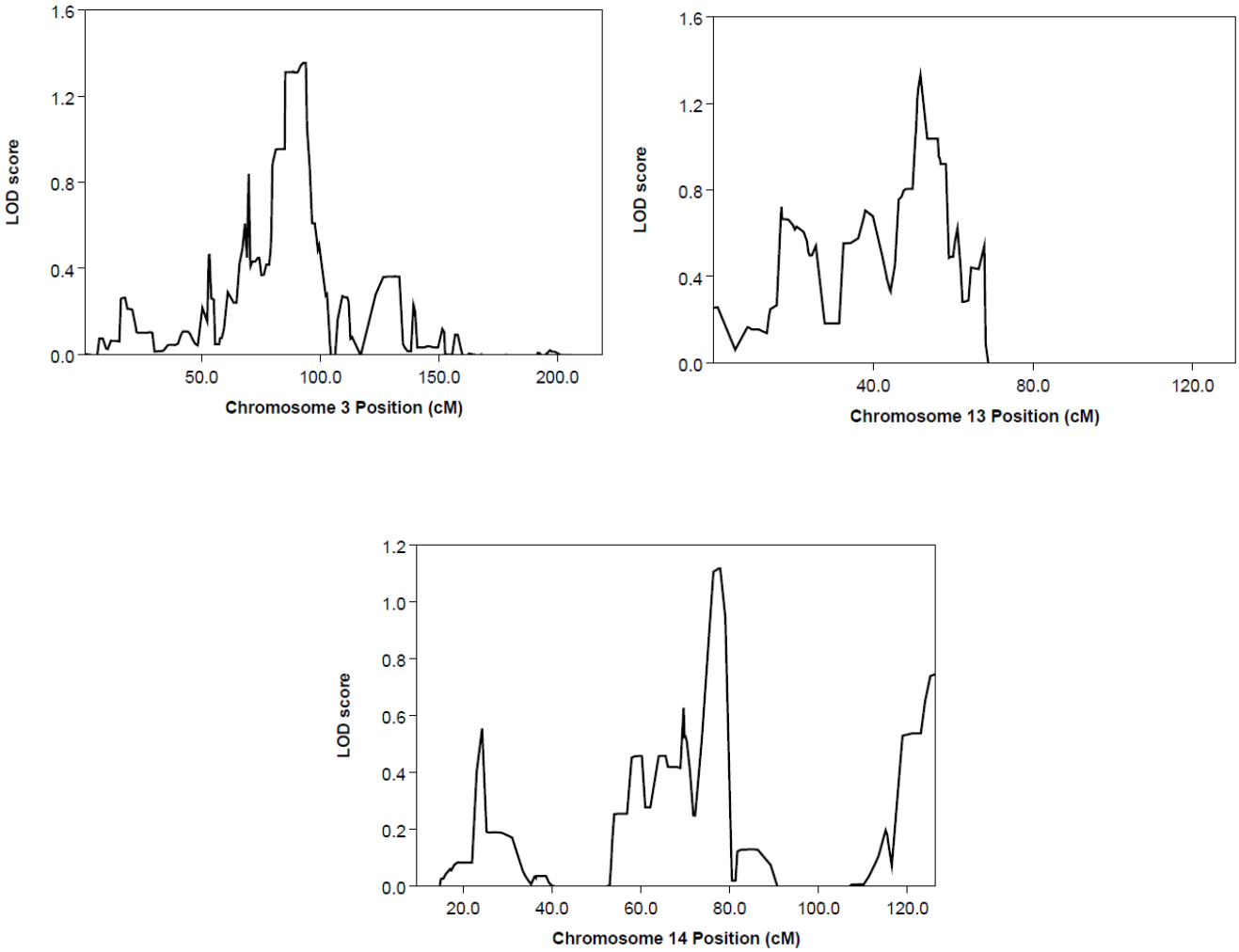


Figure 22. The NPL signals with LOD scores ≥1.0 on chromosomes 3, 13 and 14.

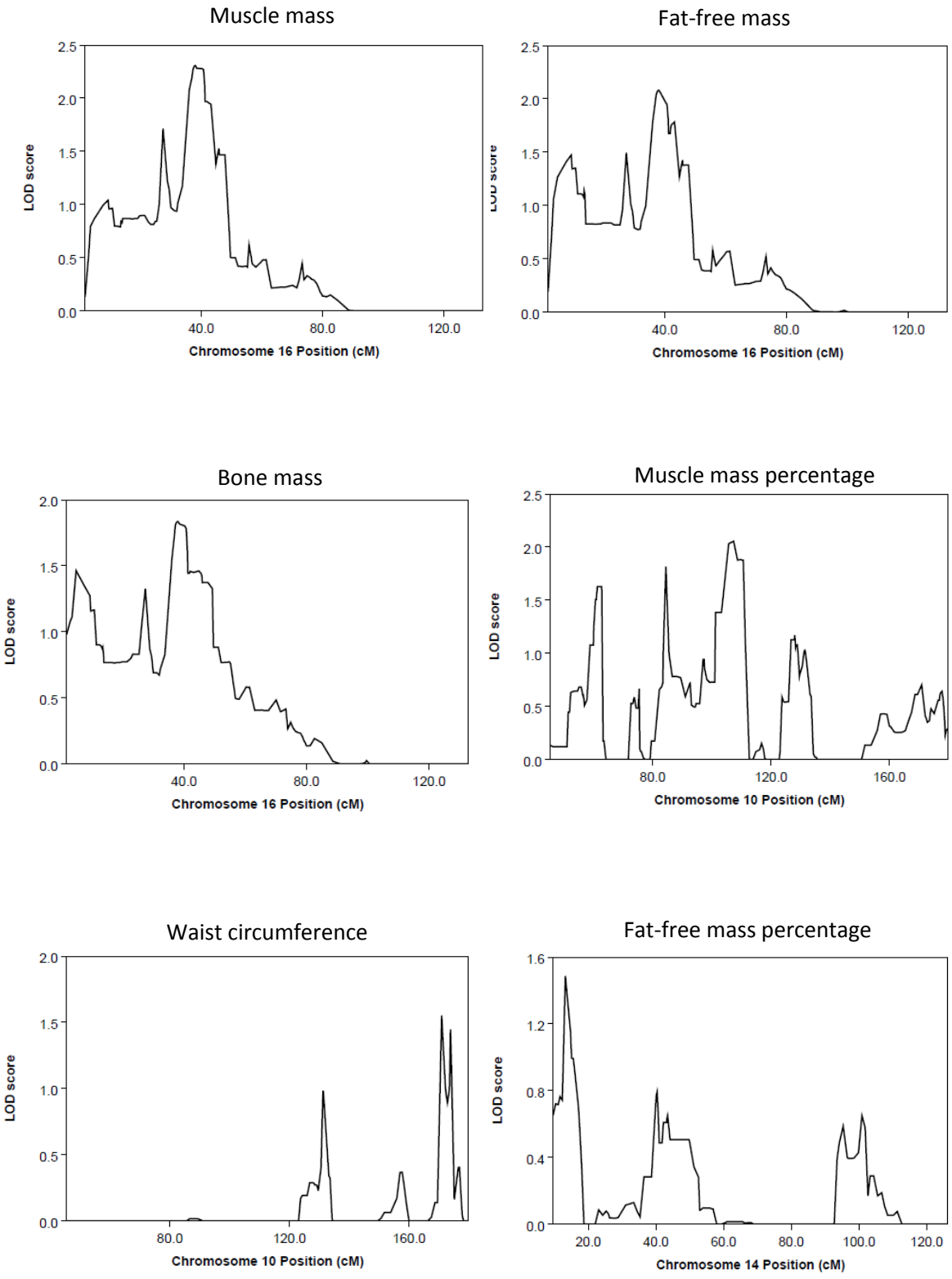


Figure 23. The VC linkage signals with LOD ≥1.5 on chromosomes 16, 10 and 14.

3.1.3 Run of Homozygosity Analysis

As discussed in Section 2.8.1.2, ROH analysis was performed for the 63 samples (34 obese, 17 lean and 12 overweight) using the SNP array genotyping data after the exclusion of SNPs with: >5% missing genotypes, MAF<0.05, deviation from HWE and Mendelian errors. Overlapping ROH regions detected in the samples were given by Plink as ROH pools. When considering ROH regions of all subjects in the dataset (34 obese, 17 lean and 12 overweight), 279 ROH pools were identified in the 22 autosomal chromosomes. When considering the ROH regions of the obese and lean subjects only, a total of 248 ROH pools were identified. Out of these 248 ROH pools, 168 pools contained more obese than lean subjects with enrichment >1. Enrichment was calculated as (number of cases with ROH * number of controls without ROH)/ (number of cases without ROH * number of controls with ROH). These pools were considered as candidate regions and were checked for overlap with candidate linkage regions. A total of 47 ROH pools were overlapping with 14 candidate linkage regions, as given in Table 11.

Chr.	Linkage trait	Linkage region start (hg19)	Linkage region end (hg19)	ROH pools overlapping with linkage regions (number of subjects showing overlap)
3	Obesity	54,807,240	73,964,127	S1 (6), S4 (2), S5 (3)
5	FFM %	140,372,222	158,131,574	S12 (2), S13 (2)
5	Bone mass	82,904,528	123,449,266	S6 (2), S7 (2), S8 (1), S9 (1), S17 (2), S20 (2), S35 (1)
5	FFM%	134,769,827	162,385,429	S12 (2), S13 (2)
10	Muscle mass %	77,379,512	93,347,297	S1 (5), S35 (3)
10	Waist circumference	130,744,348	133,198,982	S5 (3), S11 (3)
10	FFM %	105,727,659	117,718,290	S20 (1), S1 (5), S35 (3), S14 (2), S20 (1)
10	BMR	63,043,987	121,480,161	S1 (5), S35 (3), S14 (1), S20 (5)
13	Waist circumference	34,637,818	61,201,257	S8 (5)
14	Obesity	54,221,757	78,870,408	S1 (4), S2 (1)
18	FFM %	4,468,352	33,015,239	S35 (3), S14 (1)
18	Muscle mass %	4,468,352	33,015,239	S35 (3), S14 (1)
20	FFM%	53,647,394	58,795,536	S10 (1), S11 (3), S24 (2)
20	Bone mass	217,281	56,511,689	S1 (4), S2 (4), S3 (4), S4 (4), S7 (3), S8 (3), S9 (2), S10 (1), S11 (2), S24 (2)

Table 11. Summary of overlapping regions between linkage and ROH regions. Column 5 of the table provides the IDs of the ROH pools that overlap with the linkage region and the number of subjects within the pool that show the overlap with the linkage region is given between brackets. Expanded table providing lists of all genes located within these regions is provided in the Appendix 1.6.

3.2 Identification of Candidate Obesity Variants

3.2.1 Whole Genome Sequencing and Quality Control

Whole genome sequencing was performed for 63 subjects (34 obese, 17 lean and 12 overweight) with the Illumina's HiSeq 2500 platform. As discussed in Section 2.8.1.3, the data were analysed using the pipeline implemented at WCMC-Q shown in Figure 10. All samples had good overall Phred scores and had average GC content close to the expected (about 40%). The average depth of sequencing across the samples was 39.5 and the average number of sequencing reads was 659,841,529. The average number of called SNPs per genome (using GATK multi-sample calling) was 3,884,771 SNPs. In addition, high concordance was observed between the genotyping calls obtained from the HumanOmni2.5-8 genotyping arrays and the WGS pipeline, which demonstrated the accuracy of these two genotyping methods that were used in this study.

3.2.2 Ingenuity Variant Analysis (IVA)

A multi-sample VCF containing the variants information for the obese (n=34) and the lean (n=17) subjects was uploaded into IVA for the purpose of identifying potential rare obesity variants in the dataset. A total of 16,470,959 SNP variants present in the VCF were annotated in IVA. Table 12 provides an overview on these variants.

Four major filters were applied on the 16,470,959 variants that were identified in the 51 genomes, including confidence filter, frequency based filter, predicted-deleterious filter, and biological context filter in order to obtain a short list of potential obesity-relevant variants. This gave a total of 89 variants in 53 genes (see Section 3.22 for details). These variants were analysed further using two approaches. The first approach was to prioritize variants within the known monogenic obesity genes (*LEP, LEPR, POMC, MC4R, SIM1 and PCSK1*). While, the second approach for studying the 89 variants was to focus on variants located within the candidate obesity regions identified, including 1) overlapping regions between linkage and ROH, 2) linkage regions, and 3) ROH regions, as summarized by the workflow in Figure 24. These analyses gave four candidate variants given in Table 13.

Total number of variants	16,470,959
Total number of genes	21,999
Rare variants (MAF < 1%)*	
Total number of variants	8,564,188 variants
Non-synonymous variants	19,164 variants
Synonymous variants	2,123 variants
Noncoding variants	8,542,910 variants
Variants present in dbSNP	6,918,484 variants
Variants not present in dbSNP	1645,704 variants
Common variants (MAF ≥ 0.5%)*	
Total number of variants	6,384,161 variants
Non-synonymous variants	14,709 variants
Synonymous variants	451 variants
Noncoding variants	6,369,010 variants
Variants present in dbSNP	6,370,757 variants
Variants not present in dbSNP	13,404 variants

Table 12. Overview of the variants uploaded into IVA.

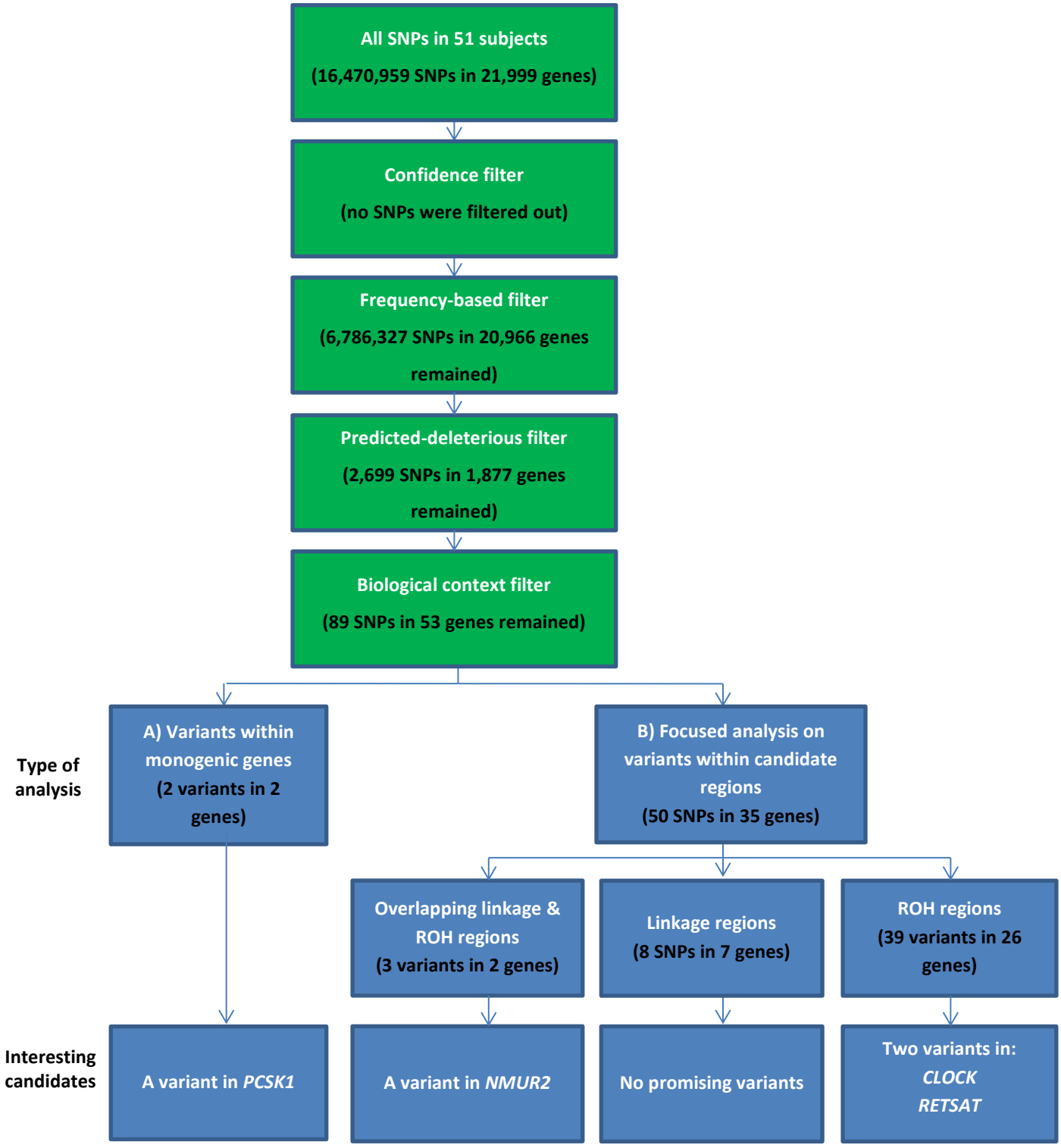


Figure 24. Variant prioritization workflow for the identification of potential rare obesity variants. Green boxes show the major filtering steps performed in IVA while blue boxes how the focused analysis within the monogenic obesity genes and the candidate obesity regions.

Chr.	Position	Ref allele	Alt allele	dbSNP ID	Region	Gene symbol	Protein variant	Cases	Controls	Impact	SIFT prediction	SIFT score	PolyPhen-2 prediction	Conservation phyloP P-value	Family
5	95,764,963	C	T	rs1799904	Exonic	*PCSK1*	p.R80Q; p.R33Q	1	0	Missense	Tolerated	0.12	Possibly damaging	$1.74X10^{-4}$	3
5	151,784,043	C	T	-	Exonic	*NMUR2*	p.W211*	2	0	Stop gain	-	-	-	$1.46X10^{-6}$	9
4	56,348,952	T	C	-	Exonic	*CLOCK*	p.M1V	2	1	Start loss	Damaging	0	Benign	-	7
2	85,578,069	A	C	-	Exonic	*RETSAT*	p.L144R	2	0	Missense	Damaging	0	Probably damaging	$1.60X10^{-5}$	4

Table 13. Candidate obesity variants identified in this study. Coordinates are given according to the GRCh36/hg18 genome assembly and allele coding based on the positive strand.

3.2.3.1 The Rare Variants in *LEPR* and *PCSK1* Genes

The analysis of variants within the monogenic obesity genes gave two variants only; one in leptin receptor (*LEPR*) and the other in proprotein convertase subtilisin/kexin type 1 (*PCSK1*). The *LEPR* variant was a heterozygous missense mutation (p.S1007C) that has been previously reported in dbSNP (rs145651189) with a frequency of 0.09% in the 1000 genome project data and 0.04% in the NHLBIESP exomes data. The variant was present in 1 case and two controls in my study, and was predicted to be tolerated and benign by SIFT (SIFT score > 0.05) and polyphen-2, respectively. Thus, this variant did not seem promising.

The variant in *PCSK1* was a heterozygous missense variant (p.R80Q) that has been previously reported in dbSNP (rs1799904) with a frequency of 0.87% in the 1000 genome projecdataset. This rare variant was detected in one case only in my study (BMI=30.1kg/m^2) who was also diabetic. The variant was predicted to be tolerated by SIFT but possible damaging by Polyphen-2. Therefore, the variant seemed to be of potential interest. Sanger sequencing confirmed the presence of the *PCSK1* variant in the subject in whom it has been detected and also confirmed its absence in the other family 3 members, as illustrated in Figures 25 and 26.

3.2.3.2 The Rare Variant in *NMUR2* Gene

The most promising variant within the overlapping regions between linkage and ROH was a novel stop gain variant (p.W211*) in neuromedin U receptor 2 (*NMUR2*). The overlapping region was on chr5:141,912,239-155,615,071 (build: hg19). The linkage region was identified on chr5:134,769,827-162,385,429 (build: hg19) with fat free mass percentage (FFM %), and the most contributing family to the signal was family 9. The ROH was identified in two cases, one from family 9 and one from family 8. The variant was detected in two cases from family 9. In one case (QO_F9_02, BMI= 46.1kg/m^2), the variant was in homozygous state (and it is one of the cases in whom the ROH has been detected), while in the other case (QO_F9_08, BMI= 30.2kg/m^2), the variant was seen in heterozygous state. The variant was not present in any of the controls, and was not also seen in data from the 1000 Genome project, the public Complete Genomics genomes or the NHLBIESP exomes.

Sanger sequencing was performed to confirm the presence of the variant in the two cases and also to check the genotypes of all Family 9 members with available DNA samples (n=17) including the overweight subjects, in order to obtain more information about the segregation of the variant in the family. As illustrated in Figures 25 and 26, Sanger sequencing has shown that the *NMUR2* variant is

present in a total of 3 cases (1 homozygous and 2 heterozygotes) out of 10 cases, and in 4 overweight subjects (1 homozygous and 3 heterozygous) out of 6 overweight subjects. None of the four controls in the family had the variant.

3.2.3.3 The Rare Variant in *CLOCK* Gene

One of the candidate obesity variants is the novel missense variant (p.M1V) in the circadian locomotor output cycles kaput (*CLOCK*) gene, detected in a candidate ROH region. The ROH region was identified in 2 cases (from family 1 and 9) and 1 control from family 1. While the variant was detected in two cases from family 7 as compound heterozygous with a G>A substitution on the 3'UTR of the gene (chr4:56,295,141; hg19) and in one control also from family 7 in heterozygous state. Clock consists of 846 amino acids and the mutation identified (p.M1V) result in loss of the translation start site, therefore the variant was considered as a potential candidate, despite the fact that the ROH was not detected in family 7 members. Sanger sequencing confirmed the presence of the variant in these subjects in addition to one overweight subject that was not analysed in IVA, as illustrated in as illustrated in Figures 25 and 26.

3.2.3.4 The Rare Variant in *RETSAT* Gene

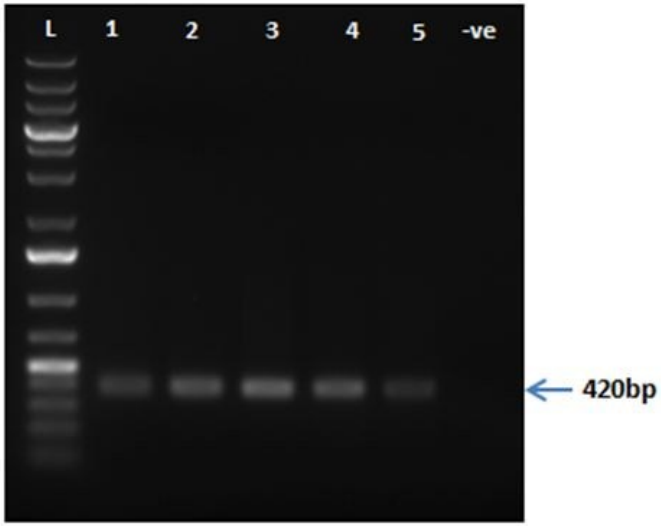
Another candidate obesity variant is the novel missense variant (p.L144R) in retinol saturase (*RETSAT*) gene, detected in an ROH region. The ROH region was seen in three cases from family 6, family 8 and family 3. The variant was seen in two cases from family 4 (compound heterozygous) and none of the controls. Interestingly, the variant was predicted to be damaging and probably damaging by SIFT and Polyphen2, respectively. Therefore, the variant was considered as a potential candidate although the ROH and the variants were seen in different families. Sanger sequencing was performed on family 4 members for validation. Sanger sequencing confirmed the presence of the variants in the two cases as well as an additional overweight subject, as illustrated in Figures 25 and 26. No control was available from this family.

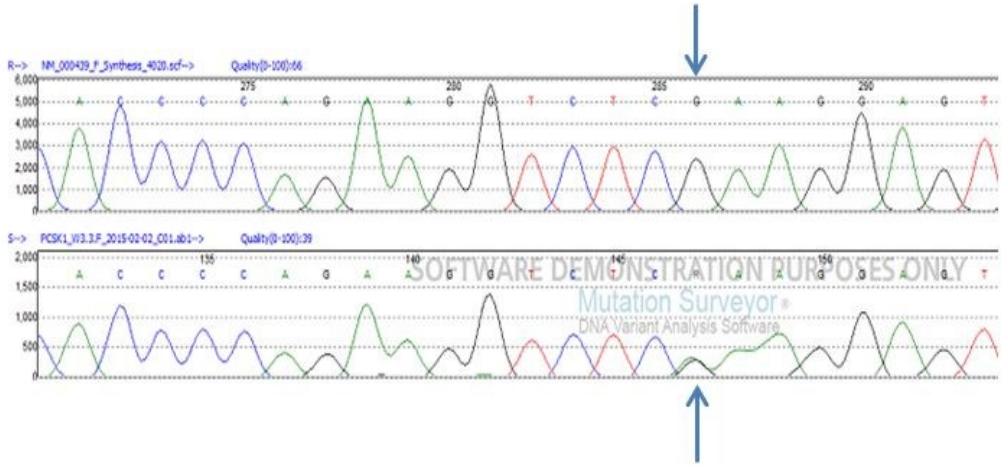
3.2.4 Replication of the Candidate Rare Obesity Variants

I checked the candidate variants in the WGS of the 108 Qatari genomes, which are described in Section 2.7.1, using an internal web application at WCMC-Q called Jbrowse. I aimed to find out whether any of the four potential candidate variants identified in my study is present in cases and/or controls from this population-based cohort consisting of 108 Qatari genomes. None of the four candidate variants were

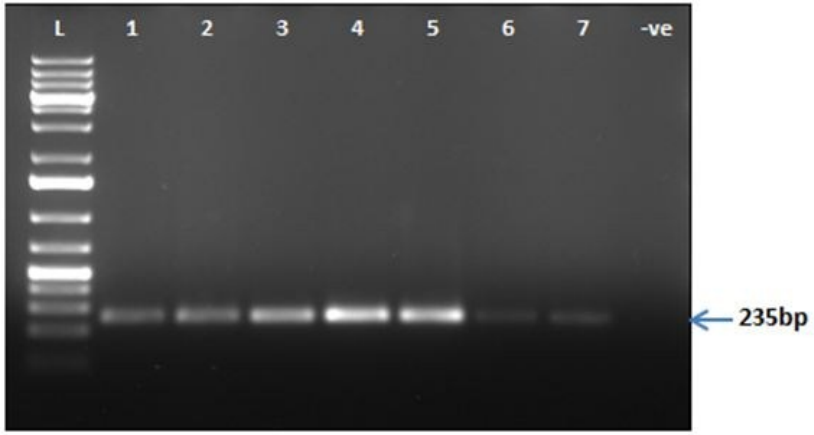
seen in the 108 Qatari genomes data. Although 108 is not a large number of samples, the absence of these variants may support the rarity of theses variants in the Qatari population.

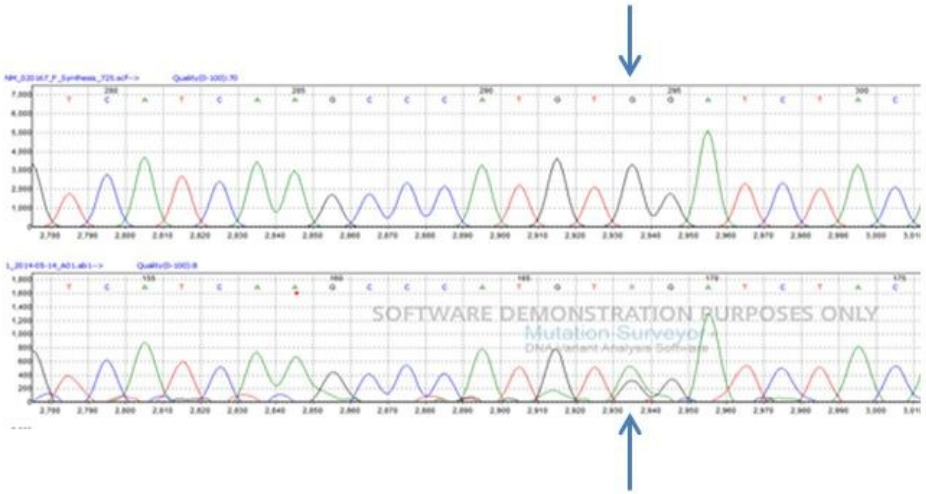
A) *PCSK1* variant





B) *NMUR2* variant





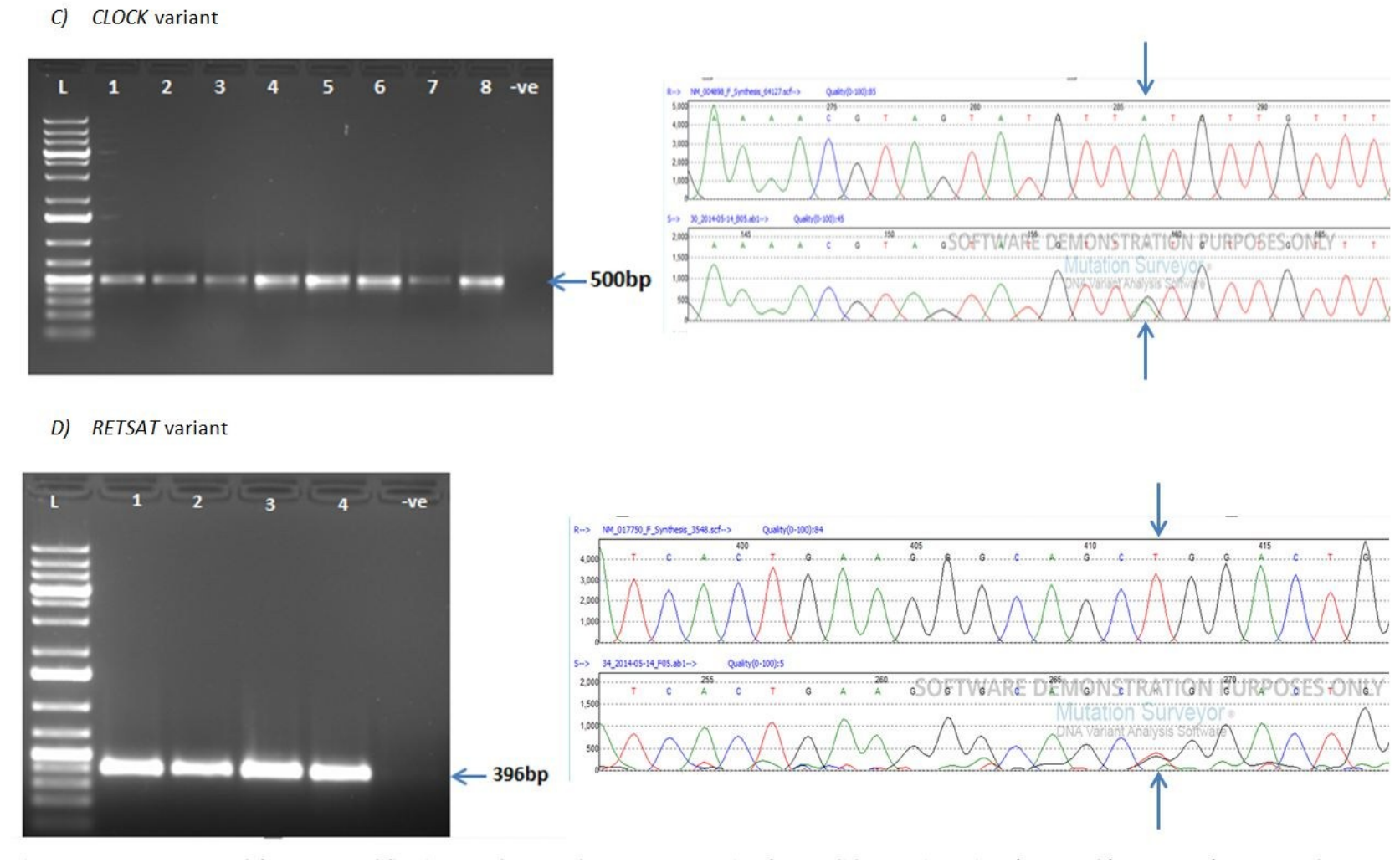


Figure 25. 1% agarose gel for PCR amplification products and Sanger sequencing traces for candidate variants in: A) *PCSK1*, B) *NMUR2*, C) *CLOCK* and D) *RETSAT*. Lane L shows 100bp DNA ladder (New England Biolabs) followed by test samples, and lane "–ve" for a negative control. Sanger sequencing trace is shown for heterozygous carriers for each of the variants with base change position indicated by the arrows.

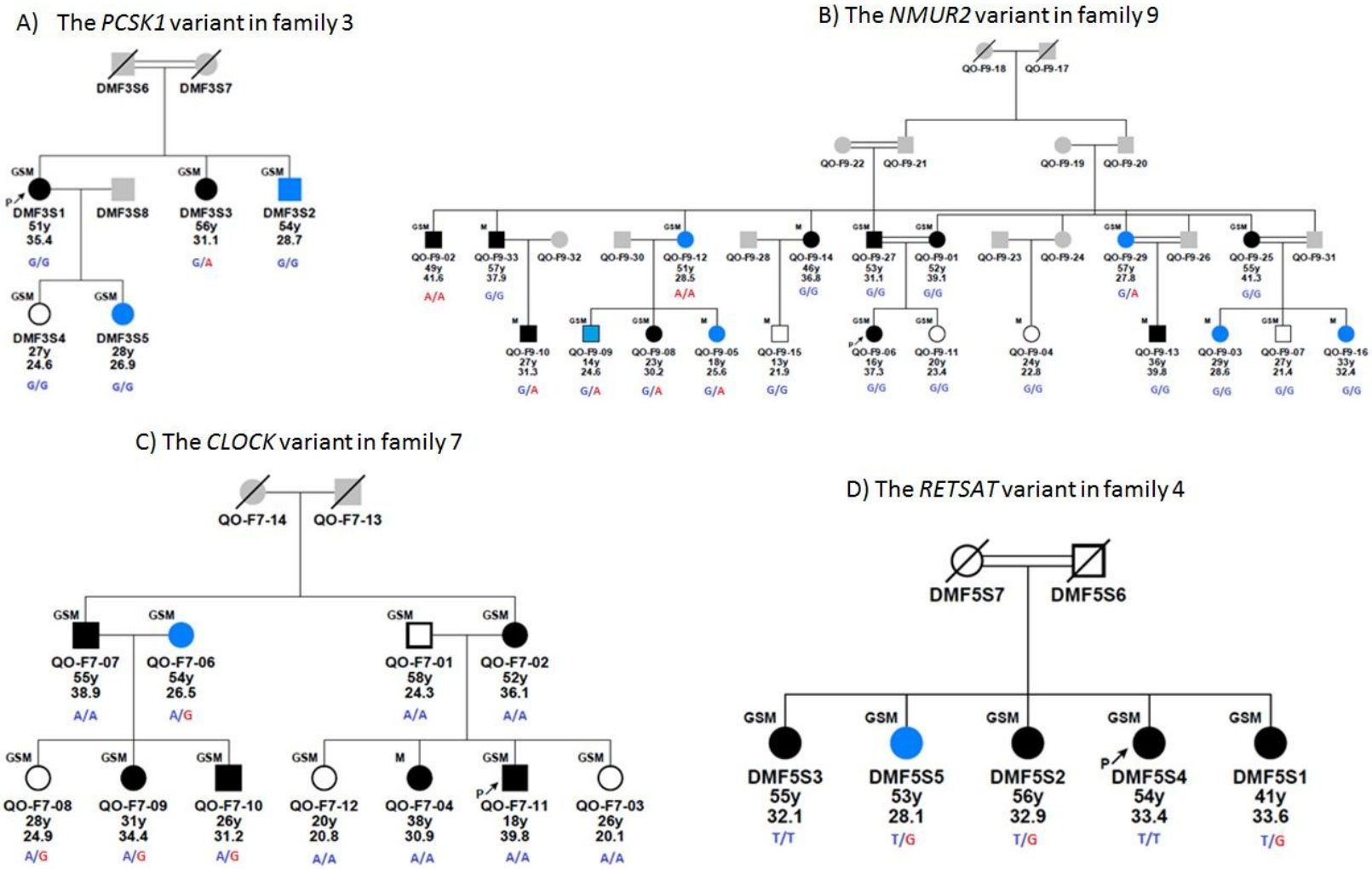


Figure 26. The genotypes of the candidate variants in the families in which they were detected. Obese subjects are shown in black, overweight in blue, normal in white and unknowns in grey. Probands are indicated with arrows. Sample ID, age and BMI are given below the individual symbol, respectively. The GSM label beside some individuals indicates the availability of genotyping, sequencing and methylation data for these subjects, while the M label indicates the availability of methylation data only. The reference allele is shown in blue while the alternative allele is shown in red.

3.3 Discussion

3.3.1 Candidate Obesity Regions

As SNPs are the most abundant source of genetic variations, it was not surprising to obtain about four million SNPs, including both frequent and rare SNPs, per genome through whole genome sequencing. This makes the identification of causative disease loci challenging especially for a common disease like obesity. Therefore, I applied a family-based strategy that combines linkage and ROH analyses with WGS to facilitate my search for rare obesity variants.

The linkage signals were generally low both in the NPL and the VC linkage analyses, and no significant linkage peak (LOD≥3.3) was obtained. However, the results obtained were somewhat expected for a common disease like obesity, where even families with highly penetrant rare variants show imperfect co-segregation (Hinney et al. 2006; Stutzmann et al. 2008). In addition, the modest sample size of my study resulted in lack of statistical power for linkage detection. Some inconsistency has been observed between NPL and VC linkage signals, as well as between the signals obtained for different quantitative traits using VC linkage analysis particularly for key obesity traits like BMI. This has also been seen in other genome-wide linkage scans for obesity and obesity-related traits (Saunders et al. 2007). A likely explanation for such inconsistency is the high locus heterogeneity underlying obesity and obesity-related traits. This corresponds with the fact that BMI values may not always represent adiposity consistently and may be influenced by other body composition traits like muscle mass and bone mass.

Some of the overlapping linkage regions identified across different traits seemed interesting. For example, four linkage signals were observed at the same position on 16p12.3 for four quantitative traits (muscle mass, fat free mass (FFM), bone mass and basal metabolic rate (BMR). The region identified for these four traits encompass G protein-coupled receptor class C group 5 member B (*GPRC5B*) gene, which encodes a G protein couples receptor that has unknown function, and a SNP located near this gene has been reported in a large meta-analysis of GWASs with BMI (Speliotes et al. 2010). The region identified with BMR also extends further and includes SH2B adapter protein 1 (*SH2B1*) gene, which encodes for an adaptor protein involved in leptin signalling, close to which a SNP has been identified in association with BMI (Thorleifsson et al. 2009; Willer et al. 2009). In addition, the signal identified on 13q14.13 has been observed in both NPL (LOD=1.3) and VC linkage with waist circumference (LOD=1.4). The overlapping region is about 22 megabases in size and it overlaps with 156 annotated genes.

It is not easy to prioritize linkage regions (and variants) based on the genes they overlap or based on the previously published linkage studies since a large number of linkage scans has been performed and they identified a large number of linkage regions for obesity and obesity-related traits, which are usually of megabases in size and may overlap hundreds of genes. Therefore, my strategy was to consider linkage regions with a LOD score of ≥1 for further analysis, and to also combine ROH information and WGS data for variant prioritization analyses (through Ingenuity Variant Analysis as well as gene functions and the available relevant literature) to facilitate the search for potential obesity variants. However, the success in finding rare obesity risk variants can be dependent on the nature of the variants in terms of frequency and functional recognisability.

My assumption was that as an inbreed population; the obesity risk variants might be present in homozygous state. Therefore, ROH analysis was applied to help prioritizing candidate linkage regions and to identify ROH regions that are enriched in case versus controls, however the sample size may have been too small for a robust enrichment analysis. Therefore, I investigated the linkage regions for harbouring rare obesity variants even if they did not overlap with ROH region, and also because the obesity variants are not necessarily present in homozygous state.

3.3.2 Whole Genome Sequencing

One of the strengths of my study is that WGS (with 40x coverage) were generated, which can provide a picture of the variants present within a genome. Whole exome sequencing (WES) provides information on the coding genomic regions, which represent only about 1% of the human genome, therefore it is widely used as an alternative to WGS to avoid the high cost, storage need as well as complexity of data interpretation in WGS. However, WES has some drawbacks including the bias in exome capture steps, missing variants within noncoding elements, and difficulty in using the data for other purposes like the detections of copy number variants. Therefore, WGS is the technique of choice for the future, especially when it becomes more affordable. The reduction in cost of WGS facilities and services is expected to accelerate the discovery of variants involved in complex diseases such as obesity. In addition, it will facilitate the routine screening for pathogenic mutations in clinical settings.

Analysing WGS data has several challenges including the low concordance between some variant calling algorithms that are currently available and the lack of efficient and publicly available tools for the annotation of the millions of variants identified through sequencing, including variants that are not

present in variant databases. Therefore, different research groups are developing their own in-house WGS pipelines using a combination of different softwares and tools.

I applied the WGS pipelines implemented at WCMC-Q for performing quality control (using FastQC, QPLOT and others), sequence reads alignment (using Bowtie2) and SNP variant calling (using GATK). The genotypes obtained using this pipeline showed high concordance with the genotypes obtained from SNP genotyping arrays. In addition, the genotypes of the four candidate variants that were validated with Sanger sequencing were concordant with the genotype calls given by the WGS pipeline, which reflects the high quality of the WGS data and analysis pipeline applied.

3.3.3 Candidate Rare Obesity Variants

In this study four potential rare obesity variants were identified through the filtration and prioritization steps in IVA and focusing on variants within monogenic obesity genes and candidate obesity regions identified through linkage and ROH. These rare variants were in *PCSK1*, *NMUR2*, *CLOCK* and *RETSAT*.

3.3.3.1 The Rare Variant in *PCSK1* Gene

By prioritizing variants within known monogenic obesity genes, I identified a candidate variant in *PCSK1*. *PCSK1* is a very interesting gene in relation to obesity due to the fact that a number of homozygous, compound heterozygous or heterozygous mutations in the gene can cause monogenic obesity (Farooqi et al. 2007; Jackson et al. 1997; Philippe et al. 2015; Zhu et al. 2002) and heterozygous mutations causing partial PC1 deficiency result in increased risk for obesity development (Creemers et al. 2012). In addition, frequent variants that modestly impair the function of PC1/3 were also found to increase the risk for common obesity (Benzinou et al. 2008; Creemers et al. 2012; Villalobos-Comparan et al. 2012). Such wide spectrum of *PCSK1* variants contributing to obesity underscores the critical role of this gene and its encoded enzyme in the regulation of energy metabolism.

Interestingly, the rare *PCSK1* variant (p.R80Q**,** rs1799904), which I detected in one obese subject of my samples, has been functionally studied in vitro by Pickett *et al.* (Pickett et al. 2013). They selected this rare variant for studying from the 1000 Genome project since it is comparatively more frequent (MAF 0.87%) in the 1000 genome project than other *PCSK1* variants and because it is predicted to be possibly damaging by Polyphen. The study showed that this variant, which is located within the secondary cleavage site of the prodomain of the enzyme, had significantly adverse effects on the maturation and catalytic function of the enzyme compared to wild type. Moreover, they illustrated that effect observed

of this rare variant was even greater than the effect observed for other previously described common obesity risk SNPs in *PCSK1* (Benzinou et al. 2008; Pickett et al. 2013). This finding supports the role of the *PCSK1* variant identified in this study to obesity susceptibility, and also highlights the potential role of rare variants in *PCSK1* as contributors to obesity risk in the general population.

3.3.3.2 The Rare Variant in *NMUR2* Gene

By prioritizing variants within overlapping regions between linkage and ROH, a promising candidate obesity variant in *NMUR2* (p.W211*) was identified. *NMUR2* encodes a G-protein coupled receptor widely expressed in the central nervous system and in the paraventricular nucleus (PVN) of the hypothalamus (Howard et al. 2000). NMUR2 functions as a receptor for neuromedin U, neuropeptide that is widely distributed in the gut and central nervous system and modulate appetite and energy homeostasis (Howard et al. 2000). Knock-out mice for *Nmu* are hyperphagic and obese whereas mice overexpressing Nmu are hypophagic and lean, suggesting that NMU is involved in energy regulation in humans. The acute central administration of NMU results in reduced food intake and increased physical activity in mice (Peier et al. 2009). However, such anorexigenic effect of NMU is not seen in *Nmu2* knock-out mice (Peier et al. 2009). Moreover, common variants in *NMU* were reported in association with overweight and obesity (Hainerova et al. 2006). The functional importance of NMUR2 encouraged Bhattacharyya *et al.* to screen the coding region of *NMUR2* in 96 subjects with severe early onset obesity to look for potential rare obesity causing variants in order to confirm the contribution of *NMUR2* to obesity in man (Bhattacharyya et al. 2004). However, no rare missense variants were detected in any of their study subjects.

In a recent study conducted by Benzon *et al.* (Benzon et al. 2014), viral-mediated RNAi was used to knock down *NMUR2* gene expression in the PVN in rats brains. When the rats were fed a standard chow (15% fat), *nmur2* knockdown produced no significant effect on food intake or body weight compared to wild type rats. However, when the same rats were fed a high-fat diet (45% fat), they consumed significantly more food (with preference for higher-fat food) and gained more body weight compared to wild type rats (Benzon et al. 2014). Therefore, I would expect that variants in *NMUR2* may have variable obesity outcomes depending on environmental effects like food type.

NMUR2 consists of 412 amino acids with seven transmembrane domains. The novel stop codon variant identified in this study is located within the third extracellular topological domain of the receptor. This means that this stop codon variant will result in a truncated receptor that lacks the C-terminus signalling

domain. Collectively, *NMUR2* studies supports the idea that the novel *NMUR2* variant detected in this study is of potential interest and that it deserves further follow-up studies to confirm its functional consequences on obesity, especially since it was present in homozygous state in two of the study subjects.

3.3.3.3 The Rare Variant in *CLOCK* Gene

In addition, a candidate variant was identified (p.M1V) in *CLOCK* gene. This gene encodes for the clock transcription factor which is a key component of the circadian clock system in mammals that allows a cell to perceive the time of the day (Gamble et al. 2014). Animal studies have demonstrated that homozygous Clock mutant mice have a greatly attenuated diurnal feeding rhythm in addition to obesity, hyperphagia and metabolic syndromes including hyperleptinemia, hyperlipidemia, hyperglycemia and hypoinsulinemia (Turek et al. 2005). Feeding these mice with high-fat diet accelerated the development of the obesity phenotype.

The variant identified in this study results in loss of the translation start site, therefore it might result in reduction of CLOCK protein levels. However, further investigations are needed to investigate the potential effect of this variant on circadian clock impairment and obesity before reaching a conclusion about this potential effect of this variant, especially since no obesity causing mutations have been previously identified within *CLOCK* in humans.

3.3.3.4 The Rare Variant in *RETSAT* Gene

Another potential candidate variant was detected in *RETSAT* gene. This gene encodes for the enzyme retinol saturase that catalyses the saturation of vitamin A, all-trans-retinol, to produce all-trans-13,14-dihydroretinol (Moise et al. 2010). Vitamin A metabolites play a role in the regulation of fat stores and in the differentiation of adipocytes. In addition, administration of vitamin A supplementation was shown to delay the onset of obesity in obesity rat models (Jeyakumar, Vajreswari, Giridharan 2006). Animal studies on *retsat* have shown that *retsat* knockout mice have similar weight gain and food intake rates compared to wild type littermates. However, they are characterized with increased adiposity and increased expression of adipogenic markers while maintained on high-fat diet (Moise et al. 2010). This shows an example of gene-environment interaction in modulating obesity. The variant I detected in *RETSAT* is a potential obesity candidate, and it might confer obesity risk under certain environmental conditions. However, these arguments are modest at this stage and more work is needed before drawing a valid conclusion.

3.3.4 Limitations

The complexity of obesity is one of the features that make the search for rare obesity variants difficult even when using family-based study designs. Such complexity arises from several factors including the high heterogeneity of obesity, the fact that obesity is not monogenic in several cases even if causal variants are within key obesity genes, the variable penetrance of obesity variants in age and gender-dependent manners which was even seen for rare monogenic obesity variants like some variants in *MC4R* (Hinney et al. 2006; Stutzmann et al. 2008), and the gene-environment interactions that play a role in determining the effect of some risk variants as in the case of variants near *FTO* for example. Such complexity may explain why there was no clear co-segregation or association evidence for the variant identified in this study. Therefore, further investigations are necessary in order to obtain valid conclusions about the role of the identified candidate variants in relation to obesity.

In addition, having WGS data with coverage of 40x can be a limiting factor when searching for rare disease variants as many rare SNPs are expected to be missed out, therefore, deeper coverage of about 100x is usually desired for such study purpose. Another limitation of the study is the difficulty in inferring the frequency of the identified variants in the general population due to the small sample size and the familial structure of the dataset, thus affecting my selection of rare SNPs. Therefore, I selected the rare obesity variants based on the frequencies observed in the 1000 Genome project, the public Complete Genomics genomes or the NHLBIESP exomes despite the fact they may not reflect the frequencies on such variants in the Qatari population.

As discussed in Section 3.2.4, none of the candidate variants were detected in the available replication dataset of the 108 Qatari genomes. The rarity of the identified candidate variants and their existence as family-specific suggests that large sample sets are in need in order to increase the chance for replicating such variants. The lack of other available datasets from Qatar or other Arab countries to be used for replication purposes is one of the study limitations at the current stage, which will be hopefully addressed in the future with the advancement in the genetic research projects in the Arab world.

3.3.5 Future Directions

The current study combined linkage (using Merlin), ROH (using Plink) and variant prediction from WGS data (using IVA) to identify candidate obesity variants. One of the future plans of this study is to apply another tool for that purpose, called the Pedigree Variant Annotation, Analysis, and Search Tool (pVAAST) (Hu et al. 2014). Unlike IVA, pVAAST is freely available at

http://www.hufflab.org/software/pvaast/. pVAAST is designed for the identification of disease genes from high throughput sequencing data in pedigrees. The tool performs linkage analysis and integrates functional prediction and case-control association features in a unified statistical framework. This approach was found to be applicable and robust for monogenic to polygenic phenotypes. It will be interesting to see whether the candidates I obtained will also be identified as disease variants using pVAAST.

Another direction for future research is to sequence the candidate genes identified in this study in large population-based cohorts in order to see whether variants in these genes are enriched in obese cases compared to controls. This can be done through collaborative efforts with Qatar Biobank and or other research institutions in the Arabian Gulf.

In addition, the possible effects of the candidate variants identified in this study can be assessed functionally, which was beyond the scope of this thesis. This would be particularly interesting for the *NMUR2* as it is the most promising candidate among the four candidates. Variants of interest can be studied in vitro by expressing them in adipose cell lines for example and measuring their effect on adiposity, mRNA levels and/or protein levels compared to wild type littermates. Other experiments may also involve the use of reporter genes to study the effect of the variant on protein localization. Different experimental designs can be applied to different candidate variants based on the variant and gene properties.

3.4 Conclusions

In conclusion, this chapter presents four potential obesity variants. For the *PCSK1*, it is likely that this variant contributes to obesity susceptibility since the effect of this particular variant on obesity has already been established functionally. The other candidate variants in *NMUR2*, *CLOCK* and *RETSAT* present very interesting obesity candidates based on gene function. However, further replication and/or functional evidence on these variants are necessary to confirm their functional role as contributors to obesity susceptibility.

CHAPTER FOUR

Identification of Common and Rare Obesity CNVs

4. Identification of Common and Rare Obesity CNVs

As discussed in Section 1.8.3 of this thesis, several common (Falchi et al. 2014; Jarick et al. 2011; Sha et al. 2009; Speliotes et al. 2010; Willer et al. 2009) and rare CNVs (Bochukova et al. 2010; Glessner et al. 2010; Walters et al. 2013; Wang et al. 2010a; Yu et al. 2011) have been associated with obesity in global populations. Our knowledge about obesity-related CNVs mainly came from studies on samples from European descents. Arabs are understudied in this research field despite the increasing incidence of obesity among Arabs.

In this chapter, I present and discuss the CNV analyses I performed on my sample set using the intensity data obtained from the HumanOmni2.5-8 SNP genotyping data for the purpose of identifying common and rare CNVs possibly contributing to obesity risk in the Qatari population. The CNV analyses were performed using the SNP genotyping data available for a total of 63 subjects (34 obese, 17 lean and 12 overweight). These subjects are indicated by the "GSM" label in their pedigrees which are given in Figure 2 in Section 2.4. CNV prediction was performed for the 63 sample using PennCNV (Wang et al. 2007), and the obtained CNVs were subjected to quality control checks and then analysed in order to identify common and rare obesity CNVs. Description of the methods is given in Chapter 2 in Section 2.8.3 while the results and discussion are given below.

4.1 CNV Prediction

The LRR and BAF information was exported form GenomeStudio (Illumina) and used for CNV detection in PennCNV (Wang et al. 2007). The initial CNV calling gave a total of 5,759 CNVs in 63 subjects (34 obese, 12 overweight and 17 lean). As quality control, CNV merging was applied to merge large (≥500kb) adjacent CNVs that tend to split automatically by PennCNV, which gave a total of 5,621 CNVs in the 63 samples. None of the samples had LRR_SD above the quality control threshold of 0.350 (samples LRR_SD ranged between 0.127 and 0.315). In addition, four outlier samples (3 obese and 1 overweight), that had a total number of predicted CNVs >2SD above the mean, were excluded from the study. A total of 4,481 CNVs in 59 subjects (31 obese, 11 overweigh and 17 lean) remained after quality control.

The most frequent type of CNVs within the 4,481 remaining CNVs was the heterozygous deletions (with 55.2%), the heterozygous duplications (with 33.1%), the homozygous deletions (with 11.6%) and the homozygous duplication (with 0.1%) respectively, as shown in Figure 27. Within each CNV state

category, there was no significant difference between the number of CNVs detected in the obese, overweight and lean subjects.

Two different approaches were applied for analysing these 4,481 CNVs for the purpose of identifying common and rare obesity CNVs (as summarized in the CNV analysis workflow given in Figure 15 in Section 2.8.2.1).

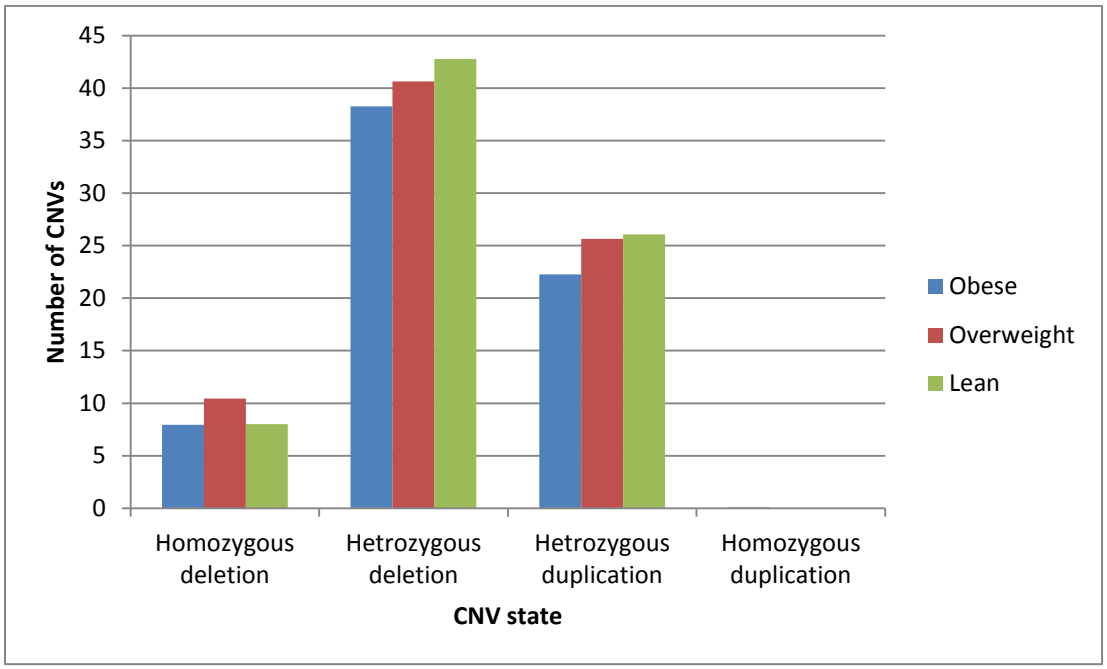


Figure 27. The frequency of the different CNV states detected in the obese, overweight and lean subjects. Number of CNVs was corrected for the number of subjects within the obese (n=34), overweight (n=11) and lean (n=17) categories.

4.2 Association of Common CNVs with BMI

To perform the common CNV association with BMI, I included SNP probes that were within inferred CNV regions of ≥1kb in size, encompassing at least 25 SNPs, and that had a copy number change in more than 10% of the samples. This gave a total of 2,236 SNP probes that correspond to 555 CNVs (detected in the 59 subjects) within 39 CNV regions. These 39 CNVs regions are summarized in Table 14. The CNV states at the 2,236 SNP probes were tested for association with BMI in a variance component framework.

Chr.	Start position (bp)	Start SNP	End position (bp)	End SNP	Size (kb)	# of SNPs	# of CNVs	Copy number state
1	798,959	rs11240777	915,543	kgp15845210	116.584	35	11	Het duplication
1	1,119,858	kgp2944849	1,425,700	kgp8416131	305.842	32	8	Het duplication
1	187,899,551	kgp15159942	188,339,497	kgp15729292	439.946	260	7	Het deletion
2	89,978,069	kgp22790915	90,275,455	kgp22747169	297.386	37	9	Het deletion/Het duplication
2	242,819,535	kgp14410578	243,068,869	kgp14601317	249.334	163	11	Het deletion/Het duplication
3	65,189,366	kgp7986924	65,214,685	rs1479940	25.319	29	15	Het deletion
3	75,419,736	rs2091866	75,682,933	kgp22757643	263.197	64	15	Het deletion/ Hom deletion
4	942,439	rs10516158	999,255	kgp22755893	56.816	20	7	Het duplication
4	25,557,047	rs7686265	25,578,269	kgp10702045	21.222	35	12	Het duplication
4	69,374,369	kgp20682437	69,601,154	kgp1629687	226.785	30	19	Het deletion/ Hom deletion
4	73,816,871	kgp6857414	74,059,396	kgp20702777	242.525	39	6	Het duplication
5	8,703,085	rs17289138	8,748,627	kgp8401171	45.542	34	6	Het duplication
5	9,902,340	kgp5416696	9,928,862	rs1035961	26.522	40	15	Het deletion
6	29,853,641	rs2394231	29,899,389	rs9259910	45.748	95	24	Het deletion/ Hom deletion
6	67,011,900	kgp1762263	67,048,885	kgp17083576	36.985	59	12	Het deletion
6	78,972,930	rs818248	79,035,739	kgp1698345	62.809	32	23	Het deletion/ Hom deletion
7	41,421	kgp7844626	79,243	kgp2138630	37.822	37	6	Het duplication
7	111,010,389	kgp13620239	111,020,216	kgp8817054	9.827	6	6	Het deletion
8	2,131,125	rs11989864	2,184,115	kgp6516407	52.99	39	6	Het deletion/ Hom deletion
8	39,232,581	kgp7704339	43,791,691	rs10958798	4559.11	75	38	Het deletion/ Hom deletion/Het duplication
8	144,894,496	rs4875060	145,025,515	kgp20103881	131.019	43	22	Het duplication
11	50,677,929	kgp1850048	51,373,124	kgp12658484	695.195	46	6	Het deletion
11	55,365,761	rs4459320	55,451,045	rs12789430	85.284	40	34	Het deletion/ Hom deletion
11	129,864,983	kgp10349003	129,959,902	kgp2634961	94.919	33	8	Het duplication
14	22,770,738	kgp331668	22,955,890	kgp37081	185.152	105	21	Het deletion
14	106,103,358	rs7492390	106,949,307	rs12590146	845.949	68	53	Het duplication
15	22,299,434	rs2605929	22,585,470	kgp8159039	286.036	28	6	Het duplication
15	34,720,794	kgp22810994	34,814,869	kgp4635701	94.075	40	9	Het deletion/ Hom deletion

Chr.	Start position (bp)	Start SNP	End position (bp)	End SNP	Size (kb)	# of SNPs	# of CNVs	Copy number state
16	568,748	kgp11405027	847,469	kgp16358310	278.721	41	7	Het duplication
16	1,916,103	rs453494	2,263,836	kgp10861734	347.733	38	6	Het duplication
16	32,363,472	kgp16212574	32,649,736	kgp16279043	286.264	40	13	Het deletion
19	797,032	kgp5849680	1,526,931	kgp5590298	729.899	124	37	Het duplication
19	2,422,177	rs735911	2,478,753	kgp4884420	56.576	31	7	Het duplication
19	20,601,006	kgp3163698	20,717,536	kgp578458	116.53	62	13	Het deletion
19	43,370,213	kgp5660821	53,552,296	rs12151194	10182.08	60	13	Het duplication
20	25,828,885	kgp5800882	25,896,760	kgp19273570	67.875	27	10	Het deletion
21	10,766,457	kgp242945	10,874,452	kgp13171982	107.995	47	8	Het duplication
21	38,347,199	kgp8090543	44,854,995	rs162365	6507.796	65	19	Het duplication
22	25,663,994	kgp7778013	25,985,070	kgp4381032	321.076	137	7	Het duplication

Table 14. The 39 CNV regions overlapping the 2,236 SNP probes that were included for the common CNV association analysis with BMI.

As shown in Table 15, seven probes showed significant association with BMI at FDR 5%. These seven probes were located at 19p13.3 (chr19:812,172-817,783; hg19). The minimum P-value ($1.0X10^{-4}$) observed was at the SNP probes: kgp21514629, kgp6465525, kgp6806336 and kgp7302446. This size of the duplication at the 19p13.3 region was ~5.6kb and it was associated with higher BMI. This duplication overlaps with two annotated genes: the lipid phosphatase-related protein type 3 (*LPPR3)* and polypyrimidine tract binding protein 1 (*PTBP1*).

I investigated the CNV on 19p13.3 in the database of genomic variations (DGV), which is a database that aims to provide a comprehensive summary of the structural variation in the human genome using CNV information from healthy controls. The database currently consists of 55 published CNV studies (MacDonald et al. 2014). The duplication on 19p13.3 which I identified in this study seems to be infrequent in the database since only three CNV variants (seen in 5 subjects) were overlapping with this duplication (Conrad et al. 2010; Park et al. 2010), as shown in Table 16. The studies that reported these variants did not investigate or relate these variants to any obesity phenotype; therefore, they don't provide any insight about the identified 19p13.3 CNV in relevance to obesity.

Marker	Chr.	Position (hg19)	P-value	Q-value	Beta
kgp21514629	19	816,065	$1.0X10^{-4}$	0.04	10.1
kgp6465525	19	817,708	$1.0X10^{-4}$	0.04	10.1
kgp6806336	19	817,783	$1.0X10^{-4}$	0.04	10.1
kgp7302446	19	815,086	$1.0X10^{-4}$	0.04	10.1
kgp21389405	19	812,309	$2.0X10^{-4}$	0.05	10.3
kgp4757336	19	814,680	$2.0X10^{-4}$	0.05	10.3
kgp590383	19	812,172	$2.0X10^{-4}$	0.05	10.3

Table 15. SNP probes associated with BMI at FDR 5%

Variant ID	Position (hg19)	Size (kb)	Frequency	Reference
nsv828354	chr19:796,933-815,173	18.2	1 in 31	(Park et al. 2010)
dgv566n67	chr19:805,448-814,402	9.0	2 in 31	(Park et al. 2010)
esv26343	chr19:811,335-812,583	12.5	1 in 40	(Conrad et al. 2010)

Table 16. CNV duplications in the database of genetic variants overlapping with the 19p13.3 duplication detected in my study.

4.3 Investigation of Rare Obesity CNVs

For the identification of rare obesity CNVs, I investigated large CNVs that are ≥ 500kb in size in order to find out whether any of the obese subjects in the sample set carry a CNV that may explain the phenotype. 24 large CNVs were detected in the dataset, as given in Appendix 1.7, and after excluding those that are located within centromeric regions, 16 CNVs remained. Visual examination of the log R ratio and B allele frequency plots, which were generated using a script in PennCNV, showed that seven of the large non-centromeric CNVs were reliable, as given in Table 17 and Figure 28. Reliable deletions are supposed to be reflected by the decrease in LRR and the loss of heterozygosity demonstrated by BAF while duplications are supposed to be reflected by the increase in LRR and the allelic imbalance demonstrated BAF. The change in LRR is usually more apparent in the case of homozygous deletions and duplications compared to heterozygous deletions and duplications.

These seven reliable CNVs were detected in seven individuals (5 obese and 2 overweight). The CNVs were located on 2p11.2, 6q21, 14q12 and 16p11.2. I checked these CNVs in other obesity CNV datasets to see whether they exist in any cases or controls, which will help prioritizing most potential obesity CNVs. These datasets including the childhood obesity cohort (CHO), the adulthood obesity cohort (ADO) from Imperial College London, the Saudi obesity and mental retardation (MR) cohort from Kings Faisal Hospital and the 108 Qatari genomes from WCMC-Q, which were all described in Section 2.7. As shown in Table 17, the large CNVs detected on 2p11.2 were present in controls from the other CNV studies. Whereas, the CNVs detected on 6q21, 14q12 and 16p11.2 were not seen in any controls across these different studies, which makes them the most promising obesity CNVs.

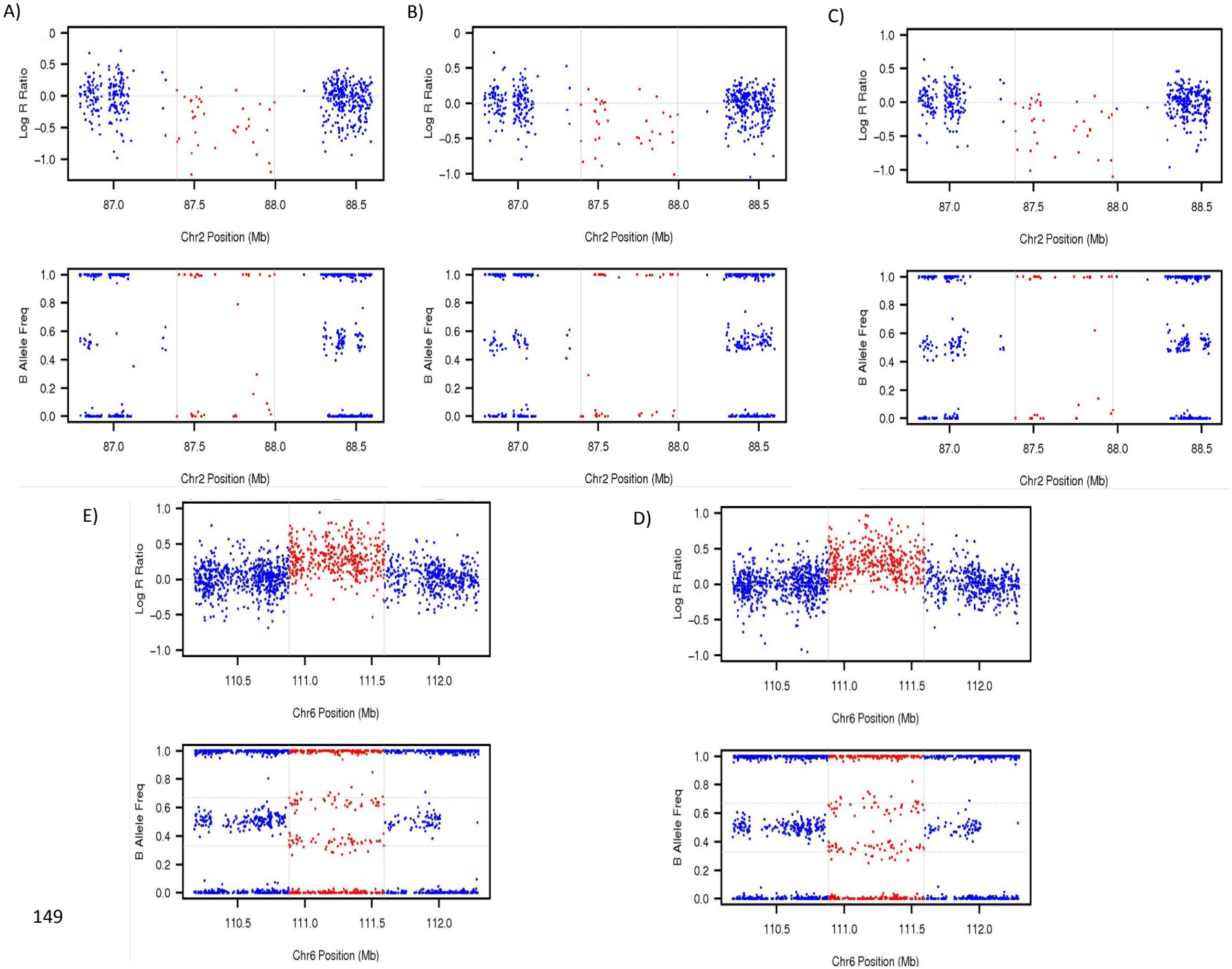
In addition, the large CNVs detected were examined in UCSC genome browser (http://genome.ucsc.edu/). The script visulaize_cnv.pl available in PennCNV was used to convert the PennCNV output to BED format. This was done to explore these CNV by visualizing them alongside with other genomic features like the overlapping RefSeq genes, SNPs and CNVs from CNV databases such as DGV (MacDonald et al. 2014) and the Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources (DECIPHER) (Firth et al. 2009), as given in Appendix 1.8.

I also investigated the large CNVs detected in my samples in the DECIPHER, which catalogues information about rare CNVs in patients with rare disorders based on information obtained from more than 200 academic departments of clinical genetics centres (Firth et al. 2009), to check whether any of

these large CNVs detected in my study were previously reported to cause obesity, as shown in Table 17. The only CNV reported in patients with severe obesity was the 16p11.2 deletion.

Chr	Start	End	Size (kb)	Number of SNPs	Cytoband	CN state	Subject ID	Age (y)	Family ID	BMI (kg/m^2)	Overlapping genes	COB/AOB data	Saudi obesity and MR data	108 Qatari genomes data
2	873,92,157	87,995,393	603.2	36	p11.2	Het deletion	DMF5S1	56	4	32.9	*LINC00152, MIR4435-1 and MIR4435-2*	1 COB control	4 controls	-
2	87,392,157	87,995,393	603.2	36		Het deletion	DMF5S2	53		28.1				
2	87,392,157	87,972,393	580.2	35		Het deletion	DMF5S5	41		33.6				
6	110,883,633	111,591,037	707.4	401	q21	Het duplication	DMF3S1	51	3	35.4	*AMD1, CDK19, GSTM2P1,GTF3C6, KIAA1919, RPF2* and *SLC16A10*	-	-	-
6	110,887,169	111,591,037	703.9	396		Het duplication	DMF3S2	54		28.7				
14	25,099,980	25,612,838	512.9	469	q12	Het duplication	QO_F5_05	20	5	39.9	*GZMB* and *STXBP6*	-	-	1 case
16	29,581,764	30,199,579	617.8	300	p11.2	Het deletion	QO_F4_02	32	1	51.8	*ALDOA, ASPHD1, C16orf54, C16orf92, CDIPT, CDIPTAS1, CORO1A, DOC2A, FAM57B, GDPD3, HIRIP3, INO80E, KCTD13, KIF22, MAPK3, MAZ, MVP, PAGR1, PPP4C, PRRT2, QPRT, SEZ6L2, SLC7A5P1, SPN, TAOK2, TBX6, TMEM219, YPEL3 and ZG16*	Observed in 4 COB cases and 4 AOB cases	1 case	-

Table 17. Large (≥500kb) rare non-centromeric CNVs detected in seven study subjects, and validated by visual examination of LRR and BAF plots. Coordinates are given according to the GRCh37/hg19 genome assembly. Het stands for heterozygous.


A)
B)
C)
D)
E)
Log R Ratio
B Allele Freq
Chr2 Position (Mb)
Chr6 Position (Mb)
87.0
87.5
88.0
88.5
110.5
111.0
111.5
112.0

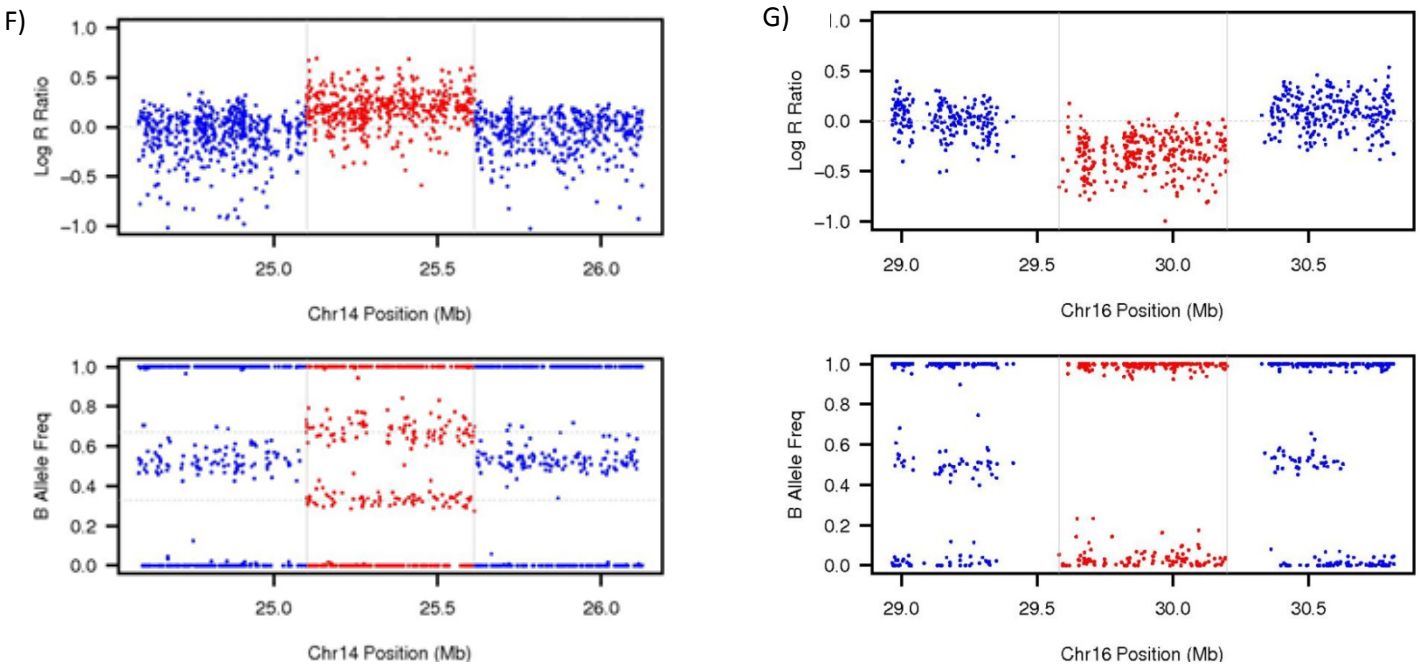


Figure 28. LRR and BAF plots of the seven large non-centromeric CNVs detected in seven study subjects. A, B and C show the heterozygous deletions observed on 2p11.2 in DMF5S1, DMF5S2 and DMF5S5, respectively. D and E show the heterozygous duplication observed on 6q21 in DMF3S1 and DMF3S2, respectively. F and G show the heterozygous duplication observed on 14q12 in QO_F5_05 and the heterozygous deletion observed on 16p11.2 in QO_F4_02, respectively. Heterozygous deletions are reflected by the decrease in LRR and the loss of heterozygosity demonstrated by BAF, which support the deletion of one allele. Heterozygous duplications are reflected by the increase in LRR and the gain of heterozygosity demonstrated by BAF, which support the gain of one allele.

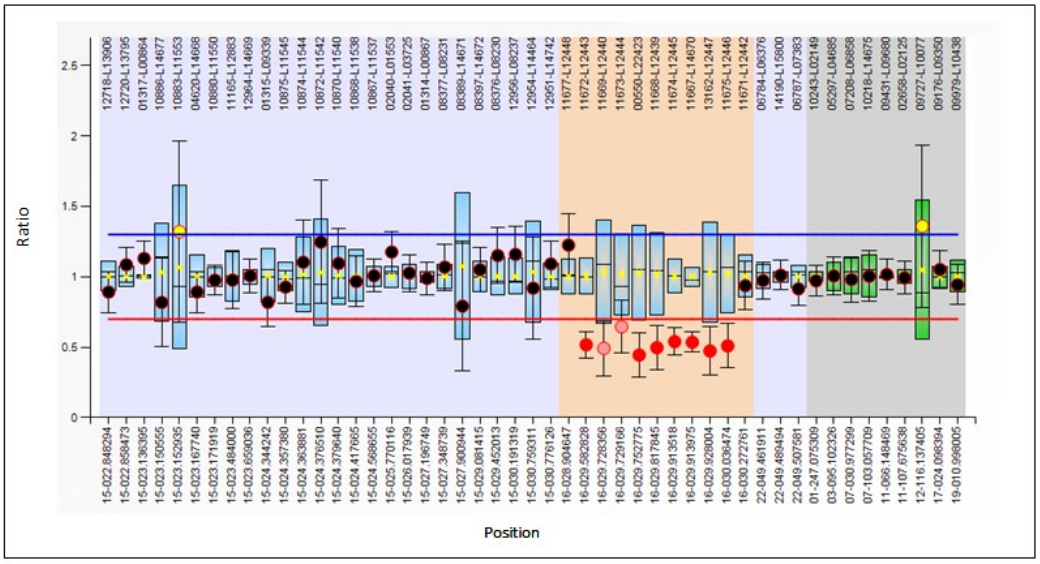
Chr	Query CNV state	Patient ID	CNV start-end	Phenotype	Size (kb)
2	Deletion	252352	84,470,510-89,026,775	Abnormality of dental morphology, broad thumb, delayed speech and language development, long toe, macrotia, muscular hypotonia, narrow palate, sandal gap, short philtrum, thin lower lip vermilion, thin upper lip vermilion, widely spaced teeth	4,556.3
		307922	87,374,062-88,297,268	Constipation, EEG abnormality, epicanthus, expressive language delay, high-arched palate	923.2
		256076	87,376,107-87,832,122	Autism, cognitive impairment	456.0
		274865	87,771,617-89,050,417	No available phenotypes	1,278.8
6	Duplication	266259	401,016-170,888,348	Arrhythmia, depressed nasal bridge	170,487.3
		4145	105,704,971-113,968,447	Brachycephaly, epicanthus, intellectual disability, intrauterine growth retardation, muscular hypotonia, ptosis	8,263.5
		252054	110,967,228-111,260,923	Delayed speech and language development, high palate, intellectual disability, microcephaly, micrognathia, muscular hypotonia, short stature	293.7
		287630	111,100,244-111,321,676	Behavioural/psychiatric abnormality, global developmental delay	221.4
		258535	111,250,449-111,651,155	No available phenotypes	400.7
		254935	111,583,447-111,620,386	No available phenotypes	36.9
14	Duplication	301147	20,253,739-41,764,477	No available phenotypes	21,510.7
		253272	20,295,211-29,615,149	No available phenotypes	9,319.9
		283651	20,472,548-31,139,579	Generalized hypotonia, seizures, severe global developmental delay, short stature	10,667.0
		262903	20,608,216-29,238,842	Intellectual disability	8,630.6
		249644	25,478,462-28,626,135	Deeply set eye, frontal upsweep of hair, hypertelorism, hypoplasia of the corpus callosum, pointed chin, prominent ears, prominent fingertip pads, short philtrum, spasticity, thick lower lip vermilion, thick upper lip vermilion	3147.7
16	Deletion	In about 260 patients	Vary in different patients	Intellectual disability, obesity, global developmental delay, delayed speech and language development, autism spectrum disorders, muscular hypotonia, seizures and others.	~600kb

Table 18. CNVs in the DECIPHER that overlap with the large rare CNVs detected in my study.

4.4 Confirmation of the 16p11.2 Deletion

I decided to independently investigate the 16p11.2 deletion since previous reports have associated this CNV to obesity (Bochukova et al. 2010; Walters et al. 2010). The CNV detected on 16p11.2 was a heterozygous deletion (chr16:29,581,764-30,199,579; hg19), as illustrated by the LRR and the BAF plots given in Figure 28 (G). The presence or absence of the deletion in the carrier (QO_F4_02) and the other family members (Family 1) was examined through Multiplex Ligation-Dependent Probe Amplification (MLPA), using the SALSA MLPA P343 Autism-1 probemix kit available by MRC-Holland (as was described in Section 2.8.2.2.3). The deletion was confirmed in the deletion carrier but was not detected in any other family members (Figure 29). The absence of the deletion in both parents of the deletion carrier suggests that it occurred *de novo*. Interestingly, this subject was the most extremely obese subject in the family and also in the whole sample set with a BMI of 51.8kg/m^2. In addition to the extreme obesity phenotype, this subject had speech delay and learning disabilities, consistent with a diagnosis of developmental delay. Interestingly, this subject had two siblings only and they were both obese (BMI= 42.4 and 37.8kg/m^2) but they were not affected with developmental delay and were not carrying the 16p11.2 deletion. In addition to the MLPA analysis, I investigated the deletion using the available WGS data. In the case of a heterozygous deletion, I would expect a drop in the sequencing read depth within the deletion region of about half the average genomic read depth of 40x. As shown in Figure 30, the WGS data also confirmed the presence of the 16p11.2 in the deletion carrier.

A)



B)

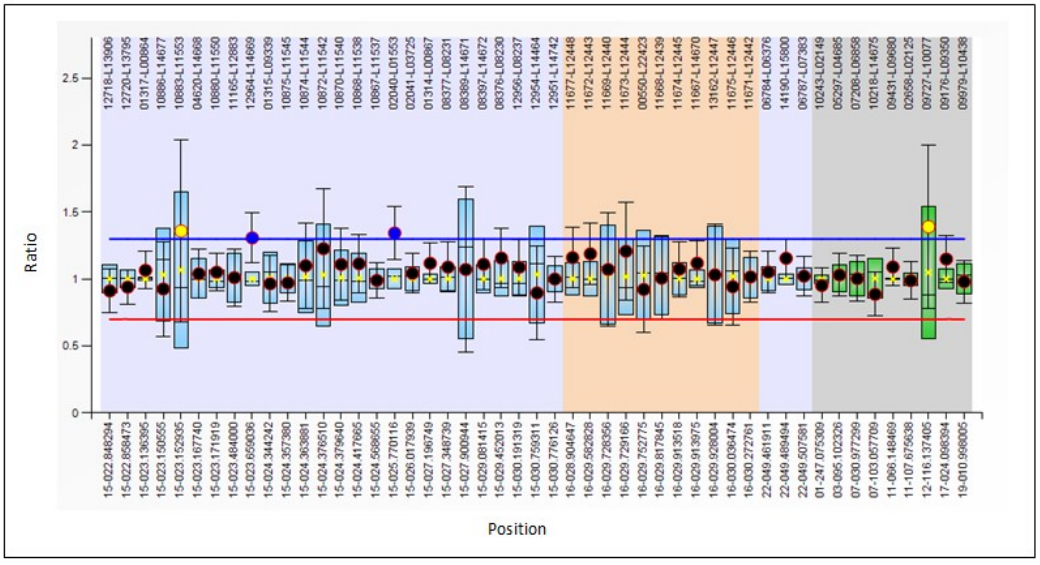


Figure 29. The normalised ratios of peak heights of the MLPA probes for: A) the 16p11.2 deletion carrier and B) a sibling that does not carry the deletion. The x-axis shows the chromosomal position of the probes, and probe IDs are given on the top of the plots. Boxes represent the 95% confidence interval of peak height ratios of reference samples (blue boxes are for test probes while green boxes are for control probes). The yellow "x" symbol shows the average ratios of three reference samples. Blue and red lines represent upper and lower arbitrary borders of the ratios. Circles represent ratios in test samples and error bars indicate their confidence range (black circles = normal ratio, yellow = ambiguous increase, pink = ambiguous decrease, and red = significant decrease).

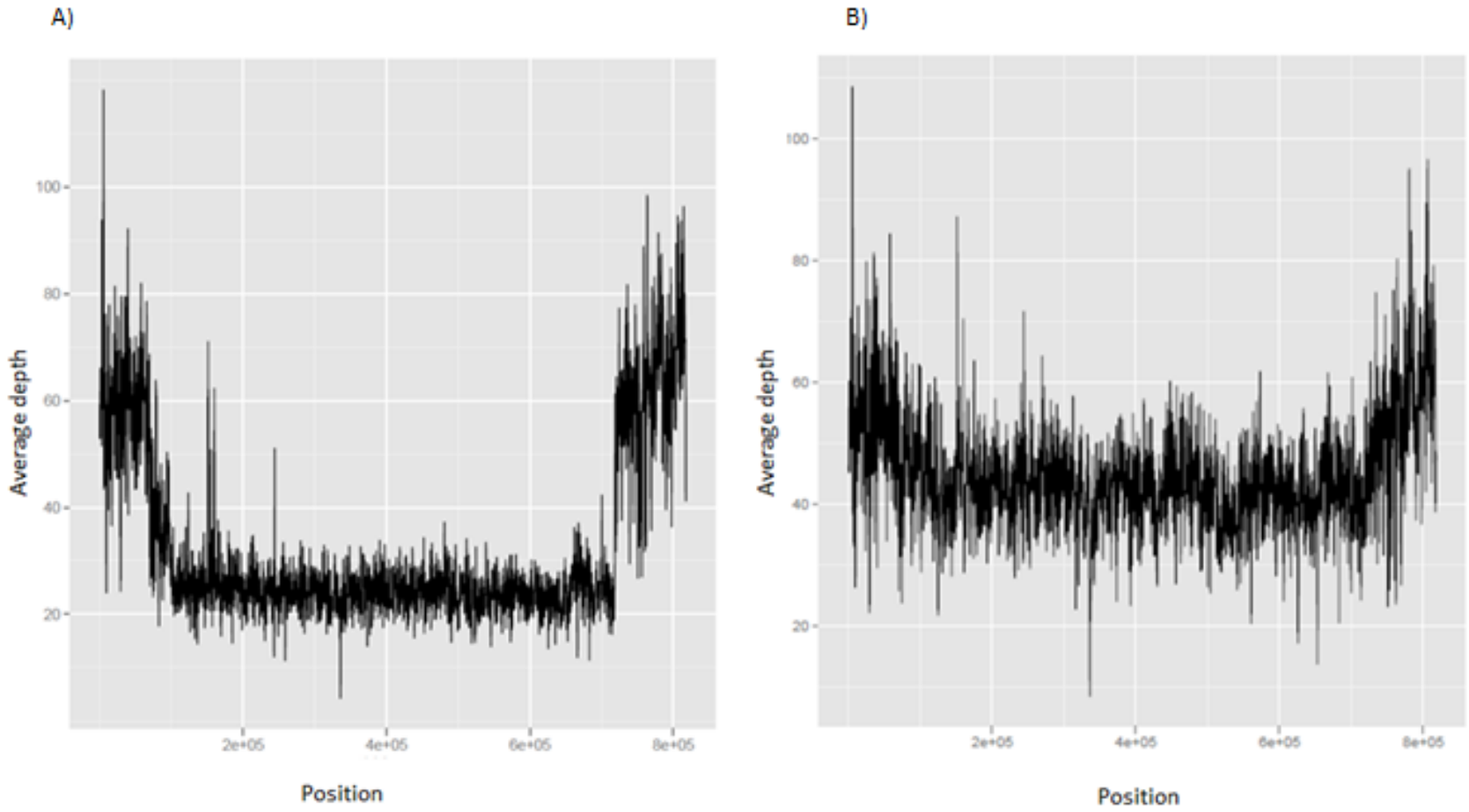


Figure 30. Read depth across the 16p11.2 deletion region (chr16:29,581,764-30,199,579 ± 100kb; hg19). A) Shows read depth in the 16p11.2 deletion carrier (QO_F4_02), where the rapid drop in read depth across the deletion region compared to the neighbouring regions supports the deletion of one allele. B) Shows read depth in a sibling that does not carry the deletion (QO_F4_03), where the observed normal read depth across the deletion region supports the presence of two alleles.

4.5 Discussion

4.5.1 Association of Common CNVs with BMI

Despite the alarming figures of obesity in the Arab world and the Gulf Arabian countries in particular, the genetics of obesity remain understudied in Arabs. To the best of my knowledge, this is the first genome-wide CNV association study with BMI in Arabs. At 5% FDR (Benjamini et al. 2001), I identified a common CNV located on 19p13.3 that associates with higher BMI. This CNV is a duplication of ~5.6kb in size and it overlaps two annotated genes, which are the lipid phosphatase-related protein type 3 (*LPPR3)* and the polypyrimidine tract binding protein 1 (*PTBP1*). LPPR3 belongs to the lipid phosphate phosphatase (LPP) protein family, which consists of integral membrane proteins that play a role in the modulation of bioactive lipid phosphates including phosphatidate, lysophosphatidate, and sphingosine-1-phosphate. *PTBP1* is part of the heterogeneous nuclear ribonucleoproteins (hnRNPs) protein family, which are RNA-binding proteins involved in pre-mRNA splicing, transport and metabolism. *PTBP1* might have some obesity and obesity-related metabolic actions through affecting the expression of adiponectin receptor 1 (AdipoR1), needed to mediate the pleiotropic effects of adiponectin (Lustig et al. 2014; Yamauchi et al. 2007). However, the identified duplication on the 19p13.3 only overlaps with the sequence of last exon of *PTBP1,* which lowers the potential of this CNV for altering the dosage of the gene.

4.5.2 Candidate Rare Obesity CNVs

In regard to rare CNVs, I investigated large (≥500kb) rare CNVs to identify potential obesity causing CNVs. I confirmed the contribution of the 16p11.2 deletion to severe obesity beyond European populations. The short arm of chromosome 16 is enriched with large segmental duplications which facilitates the occurrence of non-allelic homologous recombination, resulting in recurrent copy number changes at this locus. As was discussed in Section 1.8.3, Bochukova *et al.* (Bochukova et al. 2010) has previously reported a rare deletion of ~220kb at the 16p11.2 region (chr16:28,731,428-28,951,376; hg18) in association with extreme obesity in European children. In addition, my group at Imperial College London has identified a separate deletion of ~593kb in size at the 16p11.2 region (chr16:29,421,854-30,014,587; hg19) that was associated with severe obesity in European adults (Walters et al. 2010). Interestingly, reciprocal duplication at the same region was found to confer an increased risk for underweight (Jacquemont et al. 2011).

In this study, I detected a heterozygous *de novo* deletion of 617kb at the 16p11.2 region in the most obese subject in my samples with a BMI of 51.8 kg/m^2 who was also affected with developmental delay. The similarity between the deletion I identified in my study and the 16p11.2 deletion previously reported to cause a highly penetrant form of obesity in European adults (Walters et al. 2010) in terms of deletion state (heterozygous), position (29,581,764-30,199,579 vs 29,421,854-30,014,857; hg19), size (617kb vs 593kb), the genes they encompass (same 29 annotated genes which are given in Table 17), and the clinical phenotype (severe obesity with developmental delay) supports the causality of the deletion in the deletion carrier in my study. To the best of my knowledge, this is the first report of the 16p11.2 deletion in obese subjects from a non-European population.

Another interesting observation regarding the 16p11.2 deletion detected in this study is that the deletion carrier had two siblings who were obese but they were not carrying the deletion and were not affected with developmental delay. Such observation shows that even within the same family, obesity may result from different genetic causes or be non-genetic, and can also be associated with different phenotypic features, which highlights the genetic and phenotypic heterogeneity of such devastating health condition.

In addition to the 16p11.2 deletion, I identified other large CNV loci. The CNV on the 6q21 locus, which was detected in two siblings (1 obese and 1 overweight) from Family 3, seems to be of potential interest. This duplication was ~700kb in size and it overlaps with seven annotated genes: adenosylmethionine decarboxylase 1 (*AMD1*), cyclin-dependent kinase 19 (*CDK19*), glutathione-transferase mu 2 (muscle) pseudogene 1 (*GSTM2P1*), general transcription factor IIIC, polypeptide 6 (*GTF3C6*), ribosome production factor 2 homolog (S. cerevisiae) (*RPF2*), solute carrier family 16 (aromatic amino acid transporter), member 10 (*SLC16A10*) and *KIAA1919*. A study conducted by Tabassum *et al.* (Tabassum et al. 2012) has shown that a SNP (rs2796749) in *AMD1* gene, which encodes an enzyme involved in polyamine biosynthesis essential for cellular proliferation, is associated with childhood obesity, BMI, other measures of adiposity, and plasma leptin levels. The same study investigated the effect of rs2796749 on BMI in adults using the data available from the GIANT consortium meta-analysis (Speliotes et al. 2010). They found that rs2796749 showed consistency in association with BMI although it did not reach statistical significance.

Moreover, I detected two additional CNV loci located at 2p11.2 and 14q12. The CNV on 2p11.2 was a recurrent deletion observed in two obese and one overweight siblings form Family 4 with average size of ~596kb. The deletion overlaps with three non-protein coding RNA genes: long intergenic non-protein

coding RNA 152 (*LINC00152*), microRNA 4435-**1 (***MIR4435-1*), and microRNA 4435-**2** (*MIR4435-2*). The deletion was also detected in 1 control from the CHO cohort and 4 controls from the Saudi obesity and MR cohort (Table 17), which does not support the causality of the deletion.

The CNV on 14q12 was a ~513kb duplication detected in an obese subject from Family 5, and it overlaps granzyme B (granzyme 2, cytotoxic T-lymphocyte-associated serine esterase 1) (*GZMB*) gene and syntaxin binding protein 6 (amisyn) (*STXBP6*) gene. The duplication carrier had a high BMI of 39.9 kg/m^2 and the CNV was seen in another case from the 108 Qatari genomes dataset (Table 17). This gives potential importance to this duplication in relation to obesity, however further work is needed to investigate and confirm such potential effect, especially since the genes overlapping the duplication have not been previously implicated genetically or functionally in obesity.

4.5.3 Limitations

The small sample size of the study is considered one of the study limitations because PennCNV can generate better CNV calls (less false-positive calls) when larger sample size is available and also the larger the sample size the more robust the association analysis can be. Moreover, the small sample size makes it difficult to infer the frequency of the identified CNVs in the general population and thus affecting the selection of both common and rare CNVs.

Studies with large sample size have previously illustrated the rarity and the large effect size of the 16p11.2 deletion. This means that the 16p11.2 deletion has a large predictive value at the individual level for obesity and thus it can be used for clinical prediction of obesity (El-Sayed Moustafa and Froguel 2013). This is in fact important for a deletion like the 16p11.2 because it can be inherited from parents to children in an autosomal dominant fashion, and is associated with developmental delay as well.

Another limitation of the current study, which can be generalized to other CNV studies, is the limited ability for defining the functionally relevant gene(s) within a CNV in relevance to the disease or phenotype of interest. This is particularly an issue in the case of large CNVs that contain multiple potential disease genes. Therefore, it is vital to follow findings from CNV studies with functional work in order to understand and define the genes and mechanisms underlying the disease of interest, especially since CNVs may lead to increasing or decreasing effects on gene expression even when they are lying outside DNA coding regions (de Smith et al. 2008).

Golzio *et al.* (de Smith et al. 2008; Golzio et al. 2012) conducted a functional study on the 16p11.2 to study the macrocephalic phenotype associated with the 16p11.2 deletion reported by Walters *et al*. (Walters et al. 2013). They overexpressed and suppressed the expression of each of the 29 genes in zebrafish. They found that the suppression of potassium channel tetramerization domain containing 13 (*KCTD13)* expressions can solely induce the macrocephalic phenotype associated with the 16p11.2 deletion while its overexpression induced the macrocephalic phenotype associated with the 16p11.2 duplication. However, they could not draw a conclusion about the gene responsible for the obesity phenotypes, due to the limitation of zebrafish as animal model for studying obesity. Another study (Blaker-Lee et al. 2012) also performed on zebrafish, has investigated the role of the 16p11.2 deletion on brain development, and found that the majority of the genes within the deletion region affect brain development. For the kinesin family member 22 (*KIF22*) and aldolase A, fructose-bisphosphate (*ALDOA*) genes, a reduction of 50% in gene expression resulted in the development of abnormal brains and body shapes. While for most of the other genes, a reduction of at least 75% in gene expression was needed to observe such effects. Performing similar kind of studies but in mice models can help determining the obesity causing gene within the 16p11.2 deletion.

4.4.4 Future Directions

In this thesis, the CNV analysis was performed using the high-density SNP genotyping data. One of the future directions of the current study is to perform the CNV analysis using the available WGS data. This will allow the prediction of a wider spectrum of CNVs and at a higher resolution. In addition, it can facilitate the determination of the CNV breakpoint with a better accuracy.

In addition, since the CNVs detected on 6q21 and 14q12 seem to be of potential interest, follow-up investigations can be performed to confirm their possible impact on obesity. Overexpression and silencing experiments of the genes within these CNVs in animal models of obesity can be one of the approaches to achieve this aim.

4.6 Conclusions

In conclusion, this is the first CNV study on obesity in an Arab population. I conducted a genome-wide CNV association analysis with BMI and identified a common CNV at 19p13.3 that associates with higher BMI. In addition, I confirmed the contribution of deletions at 16p11.2 to severe obesity beyond European populations, and under different environmental pressures, which makes this study the first identifying any genetic cause of obesity in the Qatari population. Moreover, I found rare CNVs at 6q21

and 14q12 with potential contribution to obesity risk in this population, which merits further investigations. This work demonstrates the importance of common and rare CNVs in the susceptibility to common diseases like obesity. Also, it encourages researchers to study and/or extend their current genetic studies to accommodate samples from understudied populations such as the Gulf Arabs in order to better understand the genetic causes of obesity in such populations.

CHAPTER FIVE

Identification of Differentially Methylated Sites with BMI

5. Identification of Differentially Methylated Sites with BMI

As introduced in Section 1.9 of this thesis, variation in DNA methylation might provide a link between environmental exposures and phenotypes, and thus play a role in determining the susceptibility to complex diseases including obesity. The DNA methylation studies on obesity published so far (Almen et al. 2012; Almen et al. 2014; Feinberg et al. 2010; Godfrey et al. 2011; Groom et al. 2012; Kuehnen et al. 2012; Relton et al. 2012; Souren et al. 2011; Stepanow et al. 2011; Sun.D. et al. 2014; Wang et al. 2010b; Zhao, Goldberg, Vaccarino 2013) were summarized in Table 5 in Section 1.9.3.2. Given the early state of the epigenome-wide technologies for DNA methylation profiling, there is a small number of published EWASs with obesity and obesity-related traits. To date, no epigenetic study on obesity was performed on Arabs, therefore in this study I attempted to replicate previously reported DNA methylation associations from large EWASs with BMI (Dick et al. 2014; Sun.D. et al. 2014) in the Qatari population.

To achieve this aim, genome-wide methylation profiling was performed for 96 subjects of my samples using the Infinium HumanMethylation450 BeadChip, including 49 obese, 21 overweight and 26 lean subjects. These subjects are indicated by the “GSM” and the “M” labels in their pedigrees which are given in Figure 2 in Section 2.4. In addition, I combined the methylation data for additional 27 subjects from a datast of Qatari families with T2D, described in Section 2.7.4, giving a total of 123 subjects for the methylation analysis. The analysis was carried out for BMI as a continuous trait. Details the experimental and analytical aspects relevant to this chapter are given in Section 2.8.3. Here, I present the results and discussions related to the analyses performed.

5.1 Replication of Known Associations between DNA Methylation and BMI in the Qatari Samples

After extensive quality control steps that included sample and probe filtering, colour bias adjustment, quantile normalisation, BMIQ normalisation and estimation of the blood cell types coefficients**,** data on b-values for 468,472 CpG sites for 123 subjects were used to carry out an epigenome-wide association analysis with BMI. The association was evaluated using a linear mixed effect model in which DNA methylation levels as well as the covariates including age, gender, T2D status and the coefficients for the six blood cell components (monocytes, granulocytes, NK-cells, B-cells, $CD8^+$-T-cells, and $CD4^+$-T-cells) were used as fixed effects, while polygenic and environmental effects were modelled as random components, as discussed in Section 2.8.3.3.

My aim was to replicate the associations of six selected CpG sites with BMI that were reported in large EWASs published by the time this study was conducted (Dick et al. 2014; Sun.D. et al. 2014). These CpG sites are at *ABCG1* (cg06500161), *CPT1A* (cg00574958), *RNF145* (cg26403843), *HIF3A* (cg22891070), cluA/CLU1 (cg09664445) and *KLF13* (cg07814318). I did not carry out the association only at the selected probes in order to avoid missing any potential inflation of the association statistics that might be due to the limited size of my sample set. Instead, I first performed an EWAS and assessed inflation through QQ-plot and calculation of the lambda inflation coefficient of the EWAS. As shown in Figure 31, the QQ-plot showed negligible inflation of the association statistics (lambda=1.09) possibly due to hidden confounders. Therefore, no correction for inflation was applied.

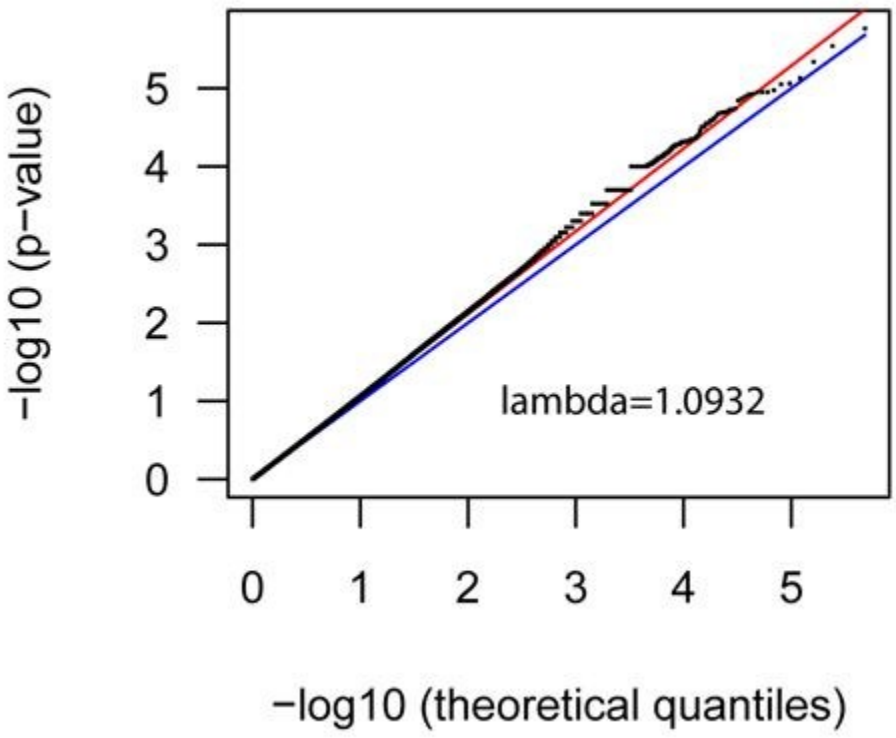


Figure 31. QQ-plot for the association between BMI and methylation sites. The blue line shows the expected P-values and the red is an interpolation of the P-values through the second and third quartile of the data. Here, the inflation coefficient is negligible (lambda= 1.0932).

I replicated the association between BMI and the CpG site cg06500161 in *ABCG1* (P-value=6.3X10^{-3}) and the CpG site cg00574958 in *CPT1A* (P-value=8.7X10^{-5}), as shown in Table 19. For every increase of BMI by a unit of 1, there is a corresponding increase by 0.0134 in the DNA methylation b-value at cg06500161 (in *ABCG1)* and a decrease by 0.0073 at cg00574958 (in *CPT1A*). The distributions of the DNA methylation b-values for these two CpG sites are shown in Figure 32, and scatterplots of their associations with BMI are given in Figure 33.

The most significant association in my dataset was detected with the CpG site cg17501210 at the Ribosomal protein S6 kinase, 90kDa, polypeptide 2 (*RPS6KA2*) gene (P-value=8.2X10^{-7}) which did not reach genome-wide significance after Bonferroni correction (Bland and Altman 1995). The top ten associated CpG sites are given in Table 20 while the top 100 associated sites are given in Appendix 1.9.

Probe	Chromosome	Position (hg19)	Gene symbol	Gene name	Reported P-value	P-value in my study	Beta in my study	Reference
cg00574958	**11**	**68,607,622**	***CPT1A***	**Carnitine palmitoyltransferase 1A**	**$1.1X10^{-9}$**	**$8.7X10^{-5}$**	-136.92	(Sun.D. et al. 2014)
cg06500161	**21**	**43,656,587**	***ABCG1***	**ATP-binding cassette, sub-family G (WHITE), member 1**	**$2.2X10^{-9}$**	**$6.2 X 10^{-3}$**	74.52	
cg26403843	5	158,634,085	*RNF145*	Ring finger protein 145	$4.0X10^{-3}$	0.02	24.07	
cg07814318	15	31,624,584	*KLF13*	Kruppel-like factor 13	$4.6X10^{-7}$	0.10	28.18	(Dick et al. 2014)
cg22891070	19	46,801,642	*HIF3A*	Hypoxia inducible factor 3, alpha subunit	$4{\cdot}0X10^{-8}$	0.14	10.40	
cg09664445	17	2,612,406	*CLUH*	Clustered mitochondria (cluA/CLU1) homolog	$2.3X10^{-7}$	0.87	-6.10	

Table 19. Known CpG sites associations with BMI that were attempted using my Qatari sample set. Replicated associations with BMI in my dataset are shown in bold. Replicated associations are those with P-value less than ($0.05/6=8.333 10^{-8}$).

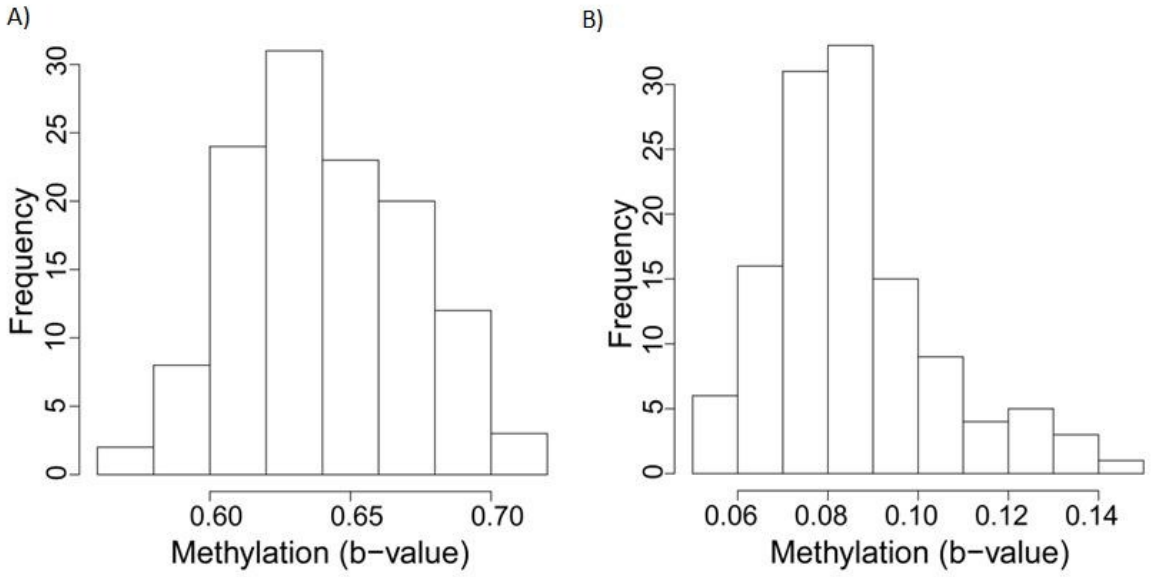


Figure 32. Histograms of the methylation b-values distribution for (A) cg06500161 at *ABCG1* and (B) cg00574958 at *CPT1A*.

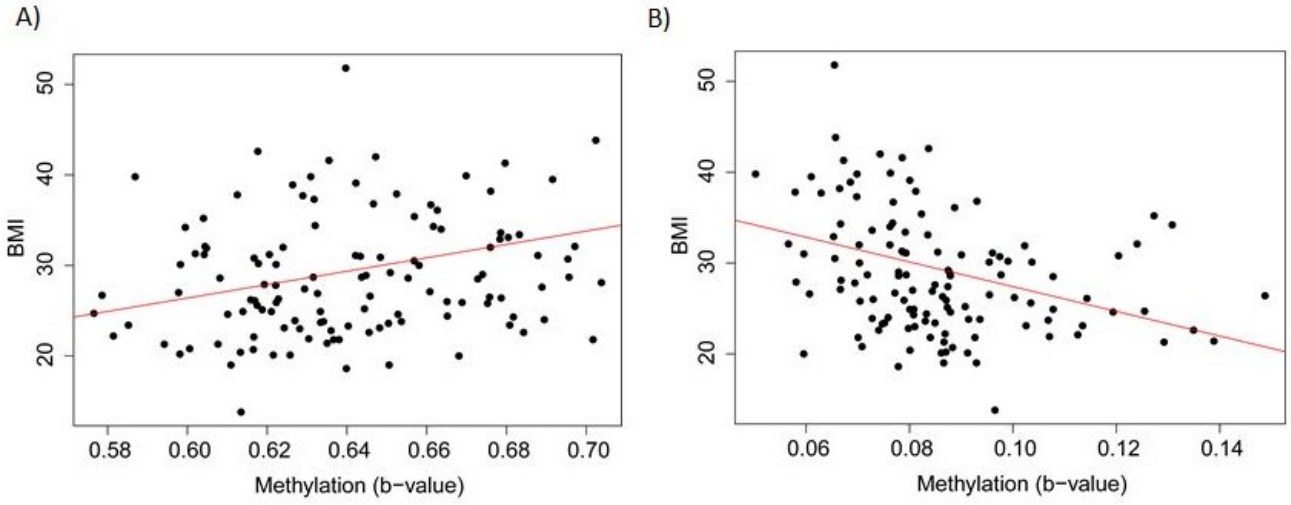


Figure 33. Scatterplots of methylation b-values against BMI for (A) cg06500161 at *ABCG1* and (B) cg00574958 at *CPT1A*. Red lines indicate the slopes of the regression models (Beta = 74.5 and -136.9, respectively).

Probe	Chromosome	Position (hg19)	Gene symbol	Gene name	P-value
cg17501210	6	166,970,252	*RPS6KA2*	Ribosomal protein S6 kinase, 90kDa, polypeptide 2	$8.2X10^{-7}$
cg18181703	17	76,354,621	*SOCS3*	Suppressor of cytokine signalling 3	$1.7X10^{-6}$
cg23943013	5	79,551,904	*SERINC5*	Serine incorporator 5	$2.9X10^{-8}$
cg05864010	11	85,376,948	*CREBZF*	CREB/ATF bZIP transcription factor	$4.6X10^{-6}$
cg06183831	1	9,883,466	*CLSTN1*	Calsyntenin 1	$7.4X10^{-6}$
cg14551938	8	49,173,274	NA	NA	$8.7X10^{-6}$
cg07146912	3	73,671,377	*PDZRN3*	PDZ domain containing ring finger 3	$8.9X10^{-6}$
cg13057167	14	103,750,565	NA	NA	$1.1X10^{-5}$
cg03717755	6	16,136,539	*MYLIP*	Myosin regulatory light chain interacting protein	$1.1X10^{-5}$
cg25499461	19	46,389,574	*IRF2BP1*	Interferon regulatory factor 2 binding protein 1	$1.1X10^{-5}$

Table 20. The top ten associated CpG sites with BMI in my Qatari data.

5.2 Association of DNA Methylation at *ABCG1* and *CPT1A* with BMI in the TwinsUK Cohort

As epigenetic patterns are strongly influenced by environmental factors, I compared in collaboration with my colleague Dr Alessia Visconti and my supervisor Dr Mario Falchi the effects of the associations I observed at *ABCG1* and *CPT1A* with BMI in my Qatari dataset to the effects they observed in the TwinsUK cohort (which is described in Section 2.7.5). Such comparison will address the question of whether the associations observed are homogeneous between these two populations and thus may give an indication of whether they are independent from variations due to differing genetic backgrounds and lifestyles.

The association between DNA methylation levels of the two probes at *ABCG1* and *CPT1A* with BMI in the TwinsUK cohort was assessed by a linear mixed effect model using also PopPANTE. The association model included DNA methylation levels and the covariates T2D status and the blood cell types coefficients for monocytes, granulocytes, NK-cells, B-cells, $CD8^+$-T-cells, and $CD4^+$-T-cells as fixed effects, while polygenic and environmental effects were modelled as random components. The analysis in the TwinsUK cohort confirmed significant association between BMI and the probes at *ABCG1* (P-value=$1.3X10^{-11}$) *and CPT1A* (P-value=$7.1X10^{-12}$).

5.3 Meta-analysis of the DNA Methylation Association Studies with BMI using the Qatari and TwinsUK Data

A meta-analysis was carried out combining the effects I observed for the associations between DNA methylation level with BMI at *ABCG1* (cg06500161) and *CPT1A* (cg00574958) probes using my Qatari samples data together with those observed using the TwinsUK data. The meta-analysis showed low to moderate heterogeneity between the two studies (I^2 =11.7% and 41.7%, for *ABCG1* and *CPT1A*, respectively; Cochran's heterogeneity P-value > 0.05). It also increased the associations at *ABCG1* (P-value=$2.5x10^{-13}$) and *CPT1A* (P-value=$1.9x10^{-15}$). 1

5.4 Discussion

Since the prevalence of obesity is high in Arab countries including Qatar, it is essential to initiate genetic and epigenetic research on obesity in this under-studied region of the world. To the best of my knowledge, this study is the first epigenetic study on obesity in an Arab population. As I did not have the power the carry out a full EWAS, I aimed to replicate in the Qatari population known associations with BMI from large EWASs that were mainly carried out in Caucasian populations.

5.4.1 Replicated Associations between DNA Methylation Levels and BMI

I attempted to replicate six CpG sites at *ABCG1, CPT1A, RNF145, HIF31A*, *CLUH* and *KLF13* and I successfully replicated the association of cg06500161 at *ABCG1* and cg00574958 at *CPT1A* (Table 19). The link between DNA methylation levels at *ABCG1* and *CPT1A* to obesity is not fully established yet. However, these genes can be functionally linked to obesity-related phenotypes.

ABCG1 encodes for ATP-binding cassette transporter that acts as a facilitator of cholesterol and phospholipid transportation in macrophages and other cell types including pancreatic beta-cells. In agreement with *ABCG1* function, a recent EWAS has reported higher levels of DNA methylation at cg06500161 in *ABCG1* with increased levels of chylomicrons and VLDL-cholesterol (Petersen et al. 2014), which supports the importance of the epigenetic regulation of *ABCG1* in relation to obesity. Animal studies on mice have shown that the disruption of *abcg1* leads to reduced granule cholesterol, altered granule morphology, impaired insulin secretion and glucose intolerance (Kruit et al. 2012). In addition, T2D patients have been characterized with reduced *ABCG1* mRNA and protein levels, which leads to altered cholesterol efflux, increased intracellular cholesterol accumulation and vascular complications (Mauldin et al. 2008).

Very recently, an EWAS on T2D carried out by Chambers *et al.* (Chambers et al. 2015) have reported association between DNA methylation levels at cg06500161 in peripheral blood with incident T2D in Indian Asians and Europeans. Interestingly, cg06500161 showed also significant association with BMI, waist-hip ratio, glucose concentrations, and other metabolic measures of insulin resistance (Chambers et al. 2015). Another study carried out by Kulkarni *et al.* (Kulkarni et al. 2015) has also recently reported association between cg06500161 methylation levels and T2D in Mexican American families. In concordance to Sun *et al.* (Sun.D. et al. 2014) and the recent Chambers's paper (Chambers et al. 2015), higher levels of DNA methylation at cg06500161 were associated with higher BMI in my study. Such increase in methylation levels at this site might be one of the mechanisms that mediates the reduction

of *ABCG1* expression and its consequences on lipid homeostasis and insulin resistance. Further work is needed to confirm such an effect and the underlying mechanisms, especially since *ABCG1* can provide a link between obesity, insulin resistance and T2D.

In addition, I replicated the association of cg00574958 at *CPT1A,* where lower levels of methylation were associated with higher BMI. *CPT1A* encodes for hepatic lipid oxidation enzyme called carnitine palmitoyltransferase 1A that catalyses the transfer of the acyl group of long-chain fatty acids into carnitine. This reaction is essential for the transportation of long-chain fatty acids into the mitochondria for beta-oxidation. Interestingly, a study conducted by Lemas *et al.* (Lemas et al. 2012) reported associations between genetic variants in *CPT1A* with obesity and fasting lipid traits. In addition, Sanchez *et al*. (Sanchez et al. 2012) found that lower levels of *CPT1A* mRNA levels in peripheral blood are protective against increased plasma triglycerides, cholesterol, BMI and also against insulin resistance. In addition to the reported association of cg00574958 with BMI (Dick et al. 2014), combined analysis of the methylation levels of CpG sites at *CPT1A* has recently revealed its association with T2D (Kulkarni et al. 2015).

The probes at *ABCG1* and *CPT1A* replicated in the Qatari samples (P-values of $6.3X10^{-3}$ and $8.7X10^{-5}$ for *ABCG1* and *CPT1A*, respectively) were also significantly associated with BMI in the TwinsUK samples (P-values of $1.3X10^{-11}$ and $7.1X10^{-12}$ for *ABCG1* and *CPT1A*, respectively) with similar direction and magnitude of association effect. These findings from my study as well as those from the epigenetic studies discussed above might underscore the role of these two CpG sites in relevance to obesity independently form differing genetic backgrounds and environmental pressures between different populations. It might also mean that the methylation patterns at these loci are consequences of obesity instead of cause (i.e. reverse causation) in which external environment might play a small role. These replicated CpG sites (and their relative genes) merit further investigations to better understand their role in relevance to obesity risk.

One of the most functionally interesting CpGs sites that was not attempted for replication at the beginning of this study but showed association with BMI was the CpG site cg18181703 at *SOCS3* ($P=1.7X10^{-6}$), which was ranked the second top associated site (Table 20). SOCS3 is part of the SOCS protein family that is usually induced by cytokines, and functions as inhibitors of several cytokine signalling pathways. Interestingly, SOCS3 acts as a negative regulator of leptin (Bjorbaek et al. 1998; Bjorbaek et al. 1999) and insulin signalling (Emanuelli et al. 2000; Rui et al. 2002; Shi et al. 2004), and SNPs located close to *SOCS3* were associated with BMI and other obesity related traits as well as glucose

homeostasis (Talbert et al. 2009; Tang et al. 2011). In addition, a study conducted by Mori *et al.* (Mori et al. 2004) has shown that *socs3* deficiency in the brains of mice models results in enhanced leptin-induced hypothalamic Stat3 tyrosine phosphorylation which consequently leads to loss in body weight and reduction in food intake compared to their wild type littermates. In addition, these mice showed resistance to high fat-induced weight gain and hyperleptinemia.

Chambers *et al.* (Chambers et al. 2015) have recently reported cg18181703 at *SOCS3* in association with T2D, BMI, waist-hip ratio, glucose concentrations, and other metabolic measures of insulin resistance. This means that my study replicates the association between cg18181703 at *SOCS3* and BMI, although this probe was not selected for replication at the start of my study since the Chamber's paper (Chambers et al. 2015) has just been published. In addition, methylation levels of CpG sites at *SOCS3,* in a combined analysis, has revealed association with T2D liability in Mexican American families (Kulkarni et al. 2015). Similar to *ABCG1* and *CPT1A*, *SOCS3* merits further investigations to better understand the epigenetic mechanism underlying its effects on obesity and its associated metabolic abnormalities.

5.4.2 Challenges of DNA Methylation Studies

Performing DNA methylation studies especially at the genome-wide level has many challenges. One of these is applying a proper quality control pipeline that is essential to minimize bias in methylation measures due to technical variations, and help focusing the downstream analysis to answer the biological question. For example, these technical issues includes the potential bias in methylation measures due to variation between the two colour channels in the Infinium HumanMethylation450K arrays, which I corrected for through a normalisation step in the quality control pipeline, and the potential bias due to variation in probe types and samples positions on the array, which we corrected for using quantile normalisation and BMIQ normalisation. Therefore, applying a proper quality control pipeline is crucial for obtaining clean data set before drawing biological conclusions (Section 2.8.3.2).

In addition to the technical issues of methylation, the instable nature of DNA methylation patterns over time and also between different cell and tissue types, and in response to environmental changes adds further complexity to the analysis of DNA methylation patterns (Sandholt, Hansen, Pedersen 2012). Such instability means that the association analysis can be confounded by many factors including age, tissue type and disease conditions. Therefore, age, gender, coefficients of blood cell types and T2D status were included in the association analysis. In addition to these factors, the degree of relatedness and thus the underlying genetic similarity between study subjects may bias the association analysis especially in

family-based studies. Therefore, we used a family-based approach for performing the association analysis. The availability of advanced technologies for DNA methylation profiling at the genome-wide level together with the required computational tools for addressing the related technical and analytical issues is expected to accelerate the number of epigenetic studies in the field of obesity and other complex diseases. However, a major problem with DNA methylation association studies is the lack of evidence for causality, which means that associations of DNA methylation levels at a particular CpG site with a disease does not necessarily mean that the methylation is responsible for causing the disease but it might reflect the effect of obesity on methylation patterns, and this could also be the basis for the low heterogeneity observed between the associations observed in the Qatari samples and the TwinsUK cohort for *ABCG1* and/or *CPT1A*.

5.4.3 Limitations

There are some limitations of this study that I'm aware off. One of these limitations is the small sample size that limited the study power from performing a full EWAS. Therefore, I focused on replicating known associations with BMI that were reported in large EWASs. Future studies with larger sample size are likely to detect associations that were not observed in the current study due to power limitations.

Another limitation of this study is that DNA methylation profiling was performed using the Infinium HumanMethylation450K arrays. These arrays can target a number of CpG sites across the genome, and therefore can provide limited methylation information to the targeted sites. On the other hand, the use of some other available whole genome sequencing-based methylation technologies can provide a more complete and comprehensive view of the methylation levels across the genome.

Moreover, the methylation measures obtained from array-based technologies can be biased in the case of the presence of SNPs within probe binding sites. Therefore, in most studies probes that contain known SNPs, based on the annotations given in the Illumina manifest, are excluded to address this problem. However, there is a chance for the existence of other unknown SNPs within the study samples, especially if they are non-Caucasian samples, since these annotations are mainly based on SNP array genotyping studies performed on Caucasian genomes. However, one of the strengths of this study is that I also have whole genome sequencing data with 40x coverage for my study subjects, which allowed the removal of such potentially confounding genetic effect.

Furthermore, DNA methylation was measured using whole-blood DNA since it was the only accessible type of sample as blood samples can be easily obtained compared to other kinds of samples. DNA

methylation measures on blood might not be representative of more disease-relevant tissues such as adipose tissues due to the existence of tissues-specific methylation patterns. In addition, as discussed in Section 2.8.3.2.6, blood is made of a cocktail of cellular subtypes in variable proportions and these subtypes carry variable methylation signatures, which may bias methylation estimates from whole blood. This problem was addressed here by estimating the cell type coefficients from the methylation data using the MethylSpectrum software (Houseman et al. 2012) and taking them into account in the association analysis with BMI, as discussed in Section 2.8.3.3.

Furthermore, it is usually common to independently validate the methylation profiles obtained using array-based methods for the candidate CpG sites of interest, like those at *ABCG1* and *CPT1A* in my study, using other technologies. However, I have not performed such experiments since the CpG sites at *ABCG1* and *CPT1A* were already included in the validation experiments performed by Petersen *et al.* (Petersen et al. 2014) using the EpiTYPER system. In addition, Illumina has reported high correlation between the DNA methylation profiles obtained using the HumanMethylation 450K array and whole genome bisulfite sequencing-based methods (Illumina 2012a).

5.4.4 Future Directions

As the association between BMI and the CpG site at *SOCS3* was significantly replicated in the Qatari samples, it will be interesting to also investigate such association in the TwinsUK cohort. Another interesting future directions on the methylation work presented in this chapter is to investigate the heritability of DNA methylation patterns especially at *ABCG1*, *CPT1A* and *SOCS3.* Such study is possible due to the family structure in the dataset that is considered a quite unique feature. This could help the determination of CpG sites with heritable patterns of methylation, and such CpG sites can be prioritized as potential obesity causative candidates from other differentially methylated sites with BMI that might be due to reverse causation.

Performing genome-wide expression analysis on the sample set, using array or sequencing-based technologies, can help understanding the consequences of the methylation patterns at *ABCG1* and *CPT1A, and SOCS3* on gene expression. It is possible to conduct such experiments since I have collected additional blood samples from my study participants in PAXgene Blood RNA tubes that stabilize intracellular RNA at the point of collection (PreAnalytiX, cat: 762165 (BD)) for downstream RNA extractions and profiling experiments.

5.5 Conclusions

This study is the first epigenetic study on obesity performed on samples from an Arab Gulf population. I replicated the previously associations of CpG sites at *ABCG1* and *CPT1A* in the Qatari samples and also in the TwinsUK samples, and also the association of a CpG site at *SOCS3* in the Qatari samples. Such finding supports the role of these CpG sites to obesity independently from variations due to differing genetic backgrounds and environmental pressures between different populations.

CHAPTER SIX

Conclusion

6. Conclusions

Obesity is a complex disease that results from interactions between genetic and environmental factors. Although there has been a significant progress in understanding the genetic contribution to the susceptibility to obesity, much remain to be revealed. Most of the success in the field of obesity genetics came from studying monogenic forms of obesity whilst, the identification and understanding of the genetic contributors to common obesity has proven to be more challenging. Rare genomic variants including SNPs and CNVs have gained a lot of attention as contributors to the missing heritability of obesity. In addition, the role of DNA methylation in complex diseases has recently become a hot research topic. Therefore, I decided to initiate through my PhD work the first genetic and epigenetic study on obesity in Qatar.

In chapter three of this thesis, I identified four rare candidate obesity mutations at *PCSK1*, *NMUR2*, *CLOCK* and *RETSAT*. Among these, the most interesting candidate was the *NMUR2* mutation and this was based on supporting evidence for its obesity-related function from the available literature. In chapter four, I identified a common CNV on 19p13.3 in associated with higher BMI, and I detected and validated a rare deletion at 16p11.2 in the most extremely obese adult in my samples, which confirmed the contribution of the 16p11.2 deletions to obesity beyond European populations. Moreover, I detected rare candidate obesity CNVs at 6q21 and 14q12. Further replication and/or functional investigation of the candidate mutations (from chapter three) and CNVs (from chapter four) are needed to confirm their potential effect on obesity. In chapter five, I replicated previously reported associations between BMI and DNA methylation levels at two CpG sites at *ABCG1* and *CPT1A* in both the Qatari samples and the TwinsUK cohort, and I also replicated the association at *SOCS3* in the Qatari samples. Replicating such associations in populations with differing genetic and environmental background supports the role of these genes in obesity and encourages further research on this topic.

There are some limitations and strengths of the work presented in this thesis that I'm aware off. One of these limitations is the small sample size of the discovery cohort, due to time and cost reasons, which limited the detection power of the study. In addition, the lack of large Arab cohorts or databases with sequencing, CNV and/or methylation data has limited my ability to replicate findings from this study in Arabs. However, there are some research projects in the Arab region that will help addressing such limitation in the future. One of these projects is the Saudi Human Genome Project by the Saudi Arabian national science agency, which aims to identify genetic basis of disease in the Saudi population by

sequencing the ~100,000 Saudi genomes representing normal and disease conditions. Similar research project is also planned in Qatar by the Qatar Biobank in collaboration with Hamad Medical Corporation and Sidra Medical and Research Centre.

The Qatar Biobank project has been established in Qatar and they have recruited about 2,000 Qatari subjects so far. Such project offers a great opportunity for researcher to access various types of samples based on their study designs, which will save the time, cost and effort needed for sample recruitment and study initiation.

The Biobank can also address another limitation that I encountered in this study, which is the limited phenotypic data that can be obtained from patient through secondary health care clinics and home visits. Having a wide range of phenotypic data and medical tests through the Biobank records and databases can facilitate the careful selection of most suitable study subjects for a particular study purpose. Also, it can help researchers providing links between closely related diseases such as obesity and T2D.

The development in the genetics research field is highly technology-driven. In this study, I utilized the state of the art technologies for genetic and epigenetic profiling such as the high-density genotyping arrays (HumanOmni2.5 array), WGS (paired-end sequencing using Hiseq2500 platform) and high-density methylation arrays (Infinium HumanMethylation450 BeadChip) that allowed the performance of genetic and epigenetic analyses at genome-wide level. Having multi-level data such as WGS and DNA methylation profiling on the same study subjects is considered as one the strengths of this dataset as it can be utilized in the future for integrating and/or confirming study findings toward system biology strategies. The advantage of having WGS and methylation data was clear in this study as the availability of the WGS data allowed the removal of the bias in DNA methylation measures that may arise from genetic variations.

One of the future strategies that can maximize the use of the current dataset is to also perform genome-wide expression profiling using expression arrays or RNA-seq technology from the study sample set using the blood samples that were collected for RNA experiments. This will be particularly useful to establish links between DNA methylation levels and their impact on gene expression in relation to obesity.

The development of analytical tools for studying the omics data is as important as the availability of the omics platforms. For example, analysing my WGS data required the implementation of the WGS pipeline

at WCMC-Q, which consisted of a number of tools for quality control, sequence alignment, and variant calling and annotation. High computing facilities were also needed for processing and storage of the large sequencing files. Similar to the WGS data, extensive quality control pipeline was applied to the methylation data to address the various sources of bias in methylation measures due to experimental factors, which also consisted of a number of tools. My study was one of the first studies at WCMC-Q that were utilized for the testing and implementation of these pipelines, which are currently utilized for the purpose of other studies at WCMC-Q.

The conductance of genetic and epigenetic studies at genome-wide level is usually a starting point that helps identifying disease-related candidates. Functional studies are then usually required to better understand the underlying mechanisms that might link such candidates to the disease of interest, and that might pave the way for drug discovery and personalized therapy in the future. Conducting functional studies was beyond the reach of this thesis due to time limitations but can be one as one of the future research directions.

The initiation of genetic and epigenetic research on obesity in Qatar was a big aim that has been successfully achieved by the conductance of this study as it represents the first genetic and epigenetic study on obesity in Arabs. In addition, it is the first study identifying any genetic cause of obesity (the 16p11.2 deletion) in a Qatari patient. It will be interesting to witness the development of research in Qatar and particularly in the field of obesity research with the completion of the Biobank project and the official opening of the Sidra Medical and Research Centre, as well as the drop in NGS technologies that will facilitate the conductance of sequencing projects at a large-scale.

This thesis encourages further research in this under-studied region of the world and also motivates existing projects to include samples from under-studied populations like Arabs. It is hoped that further follow-up of findings from this thesis will provide better understating of obesity, and might help providing better treatment and/or prevention options for obesity in the future.

References

Abecasis GR and Wigginton JE. 2005. Handling marker-marker linkage disequilibrium: Pedigree analysis with clustered markers. Am J Hum Genet 77(5):754-67.

Abecasis GR, Cherny SS, Cookson WO, Cardon LR. 2002. Merlin--rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet 30(1):97-101.

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. 2010. A method and server for predicting damaging missense mutations. Nat Methods 7(4):248-9.

Almen MS, Nilsson EK, Jacobsson JA, Kalnina I, Klovins J, Fredriksson R, Schioth HB. 2014. Genome-wide analysis reveals DNA methylation markers that vary with both age and obesity. Gene 548(1):61-7.

Almen MS, Jacobsson JA, Moschonis G, Benedict C, Chrousos GP, Fredriksson R, Schioth HB. 2012. Genome wide analysis reveals association of a FTO gene variant with epigenetic changes. Genomics 99(3):132-7.

Atwood LD, Heard-Costa NL, Cupples LA, Jaquish CE, Wilson PW, D'Agostino RB. 2002. Genomewide linkage analysis of body mass index across 28 years of the framingham heart study. Am J Hum Genet 71(5):1044-50.

Baldini A. 2005. Dissecting contiguous gene defects: TBX1. Curr Opin Genet Dev 15(3):279-84.

Barsh GS and Schwartz MW. 2002. Genetic approaches to studying energy balance: Perception and integration. Nat Rev Genet 3(8):589-600.

Beckmann JS, Sharp AJ, Antonarakis SE. 2008. CNVs and genetic medicine (excitement and consequences of a rediscovery). Cytogenet Genome Res 123(1-4):7-16.

Bell CG, Walley AJ, Froguel P. 2005. The genetics of human obesity. Nat Rev Genet 6(3):221-34.

Bell JT, Tsai PC, Yang TP, Pidsley R, Nisbet J, Glass D, Mangino M, Zhai G, Zhang F, Valdes A, et al. 2012. Epigenome-wide scans identify differentially methylated regions for age and age-related phenotypes in a healthy ageing population. PLoS Genet 8(4):e1002629.

Bener A and Alali KA. 2006. Consanguineous marriage in a newly developed country: The qatari population. J Biosoc Sci 38(2):239-46.

Benjamini Y, Drai D, Elmer G, Kafkafi N, Golani I. 2001. Controlling the false discovery rate in behavior genetics research. Behav Brain Res 125(1-2):279-84.

Benzinou M, Walley A, Lobbens S, Charles MA, Jouret B, Fumeron F, Balkau B, Meyre D, Froguel P. 2006. Bardet-biedl syndrome gene variants are associated with both childhood and adult common obesity in french caucasians. Diabetes 55(10):2876-82.

Benzinou M, Creemers JW, Choquet H, Lobbens S, Dina C, Durand E, Guerardel A, Boutin P, Jouret B, Heude B, et al. 2008. Common nonsynonymous variants in PCSK1 confer risk of obesity. Nat Genet 40(8):943-5.

Benzon CR, Johnson SB, McCue DL, Li D, Green TA, Hommel JD. 2014. Neuromedin U receptor 2 knockdown in the paraventricular nucleus modifies behavioral responses to obesogenic high-fat food and leads to increased body weight. Neuroscience 258:270-9.

Berndt SI, Gustafsson S, Magi R, Ganna A, Wheeler E, Feitosa MF, Justice AE, Monda KL, Croteau-Chonka DC, Day FR, et al. 2013. Genome-wide meta-analysis identifies 11 new loci for anthropometric traits and provides insights into genetic architecture. Nat Genet .

Bhattacharyya S, Luan J, Farooqi IS, Keogh J, Montague C, Brennand J, Jorde L, Wareham NJ, O'Rahilly S. 2004. Studies of the neuromedin U-2 receptor gene in human obesity: Evidence for the existence of two ancestral forms of the receptor. J Endocrinol 183(1):115-20.

Bibikova M, Le J, Barnes B, Saedinia-Melnyk S, Zhou L, Shen R, Gunderson KL. 2009. Genome-wide DNA methylation profiling using infinium(R) assay. Epigenomics 1(1):177-200.

Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, Le JM, Delano D, Zhang L, Schroth GP, Gunderson KL, et al. 2011. High density DNA methylation array with single CpG site resolution. Genomics 98(4):288-95.

Bjorbaek C, El-Haschimi K, Frantz JD, Flier JS. 1999. The role of SOCS-3 in leptin signaling and leptin resistance. J Biol Chem 274(42):30059-65.

Bjorbaek C, Elmquist JK, Frantz JD, Shoelson SE, Flier JS. 1998. Identification of SOCS-3 as a potential mediator of central leptin resistance. Mol Cell 1(4):619-25.

Blakemore AI and Froguel P. 2010. Investigation of mendelian forms of obesity holds out the prospect of personalized medicine. Ann N Y Acad Sci 1214:180-9.

Blakemore AI, Meyre D, Delplanque J, Vatin V, Lecoeur C, Marre M, Tichet J, Balkau B, Froguel P, Walley AJ. 2009. A rare variant in the visfatin gene (NAMPT/PBEF1) is associated with protection from obesity. Obesity (Silver Spring) 17(8):1549-53.

Blaker-Lee A, Gupta S, McCammon JM, De Rienzo G, Sive H. 2012. Zebrafish homologs of genes within 16p11.2, a genomic region associated with brain disorders, are active during brain development, and include two deletion dosage sensor genes. Dis Model Mech 5(6):834-51.

Bland JM and Altman DG. 1995. Multiple significance tests: The bonferroni method. BMJ 310(6973):170.

Bochukova EG, Huang N, Keogh J, Henning E, Purmann C, Blaszczyk K, Saeed S, Hamilton-Shield J, Clayton-Smith J, O'Rahilly S, et al. 2010. Large, rare chromosomal deletions associated with severe early-onset obesity. Nature 463(7281):666-70.

Bodmer W and Bonilla C. 2008. Common and rare variants in multifactorial susceptibility to common diseases. Nat Genet 40(6):695-701.

Bonnefond A, Raimondo A, Stutzmann F, Ghoussaini M, Ramachandrappa S, Bersten DC, Durand E, Vatin V, Balkau B, Lantieri O, et al. 2013. Loss-of-function mutations in SIM1 contribute to obesity and prader-willi-like features. J Clin Invest 123(7):3037-41.

Bradfield JP, Taal HR, Timpson NJ, Scherag A, Lecoeur C, Warrington NM, 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(5):526-31.

Bruder CE, Piotrowski A, Gijsbers AA, Andersson R, Erickson S, Diaz de Stahl T, Menzel U, Sandgren J, von Tell D, Poplawski A, et al. 2008. Phenotypically concordant and discordant monozygotic twins display different DNA copy-number-variation profiles. Am J Hum Genet 82(3):763-71.

Cai G, Cole SA, Bastarrachea RA, Maccluer JW, Blangero J, Comuzzie AG. 2004. Quantitative trait locus determining dietary macronutrient intakes is located on human chromosome 2p22. Am J Clin Nutr 80(5):1410-4.

Campfield LA, Smith FJ, Guisez Y, Devos R, Burn P. 1995. Recombinant mouse OB protein: Evidence for a peripheral signal linking adiposity and central neural networks. Science 269(5223):546-9.

Carter NP. 2007. Methods and strategies for analyzing copy number variation using DNA microarrays. Nat Genet 39(7 Suppl):S16-21.

Challis BG, Pritchard LE, Creemers JW, Delplanque J, Keogh JM, Luan J, Wareham NJ, Yeo GS, Bhattacharyya S, Froguel P, et al. 2002. A missense mutation disrupting a dibasic prohormone processing site in pro-opiomelanocortin (POMC) increases susceptibility to early-onset obesity through a novel molecular mechanism. Hum Mol Genet 11(17):1997-2004.

Chambers JC, Elliott P, Zabaneh D, Zhang W, Li Y, Froguel P, Balding D, Scott J, Kooner JS. 2008. Common genetic variation near MC4R is associated with waist circumference and insulin resistance. Nat Genet 40(6):716-8.

Chambers JC, Loh M, Lehne B, Drong A, Kriebel J, Motta V, Wahl S, Elliott HR, Rota F, Scott WR, et al. 2015. Epigenome-wide association of DNA methylation markers in peripheral blood from indian asians and europeans with incident type 2 diabetes: A nested case-control study. Lancet Diabetes Endocrinol .

Chiang DY, Getz G, Jaffe DB, O'Kelly MJ, Zhao X, Carter SL, Russ C, Nusbaum C, Meyerson M, Lander ES. 2009. High-resolution mapping of copy-number alterations with massively parallel sequencing. Nat Methods 6(1):99-103.

Chua SC,Jr, Chung WK, Wu-Peng XS, Zhang Y, Liu SM, Tartaglia L, Leibel RL. 1996. Phenotypes of mouse diabetes and rat fatty due to mutations in the OB (leptin) receptor. Science 271(5251):994-6.

Church C, Moir L, McMurray F, Girard C, Banks GT, Teboul L, Wells S, Bruning JC, Nolan PM, Ashcroft FM, et al. 2010. Overexpression of fto leads to increased food intake and results in obesity. Nat Genet 42(12):1086-92.

Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) 6(2):80-92.

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

Clarke L, Zheng-Bradley X, Smith R, Kulesha E, Xiao C, Toneva I, Vaughan B, Preuss D, Leinonen R, Shumway M, et al. 2012. The 1000 genomes project: Data management and community access. Nat Methods 9(5):459-62.

Clement K, Vaisse C, Lahlou N, Cabrol S, Pelloux V, Cassuto D, Gourmelen M, Dina C, Chambaz J, Lacorte JM, et al. 1998. A mutation in the human leptin receptor gene causes obesity and pituitary dysfunction. Nature 392(6674):398-401.

Coin LJ, Asher JE, Walters RG, Moustafa JS, de Smith AJ, Sladek R, Balding DJ, Froguel P, Blakemore AI. 2010. cnvHap: An integrative population and haplotype-based multiplatform model of SNPs and CNVs. Nat Methods 7(7):541-6.

Colella S, Shen L, Baggerly KA, Issa JP, Krahe R. 2003. Sensitive and quantitative universal pyrosequencing methylation analysis of CpG sites. BioTechniques 35(1):146-50.

Conrad DF and Hurles ME. 2007. The population genetics of structural variation. Nat Genet 39(7 Suppl):S30-6.

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

Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, Aerts J, Andrews TD, Barnes C, Campbell P, et al. 2010. Origins and functional impact of copy number variation in the human genome. Nature 464(7289):704-12.

Considine RV, Sinha MK, Heiman ML, Kriauciunas A, Stephens TW, Nyce MR, Ohannesian JP, Marco CC, McKee LJ, Bauer TL. 1996. Serum immunoreactive-leptin concentrations in normal-weight and obese humans. N Engl J Med 334(5):292-5.

Cooper GM, Nickerson DA, Eichler EE. 2007. Mutational and selective effects on copy-number variants in the human genome. Nat Genet 39(7 Suppl):S22-9.

Cotsapas C, Speliotes EK, Hatoum IJ, Greenawalt DM, Dobrin R, Lum PY, Suver C, Chudin E, Kemp D, Reitman M, et al. 2009. Common body mass index-associated variants confer risk of extreme obesity. Hum Mol Genet 18(18):3502-7.

Creemers JW, Choquet H, Stijnen P, Vatin V, Pigeyre M, Beckers S, Meulemans S, Than ME, Yengo L, Tauber M, et al. 2012. Heterozygous mutations causing partial prohormone convertase 1 deficiency contribute to human obesity. Diabetes 61(2):383-90.

Dabelea D, Hanson RL, Lindsay RS, Pettitt DJ, Imperatore G, Gabir MM, Roumain J, Bennett PH, Knowler WC. 2000. Intrauterine exposure to diabetes conveys risks for type 2 diabetes and obesity: A study of discordant sibships. Diabetes 49(12):2208-11.

de Smith AJ, Walters RG, Froguel P, Blakemore AI. 2008. Human genes involved in copy number variation: Mechanisms of origin, functional effects and implications for disease. Cytogenet Genome Res 123(1-4):17-26.

de Smith AJ, Purmann C, Walters RG, Ellis RJ, Holder SE, Van Haelst MM, Brady AF, Fairbrother UL, Dattani M, Keogh JM, et al. 2009. A deletion of the HBII-85 class of small nucleolar RNAs (snoRNAs) is associated with hyperphagia, obesity and hypogonadism. Hum Mol Genet 18(17):3257-65.

Deaton AM and Bird A. 2011. CpG islands and the regulation of transcription. Genes Dev 25(10):1010-22.

Dedeurwaerder S, Defrance M, Bizet M, Calonne E, Bontempi G, Fuks F. 2014. A comprehensive overview of infinium HumanMethylation450 data processing. Brief Bioinform 15(6):929-41.

Deeb SS, Fajas L, Nemoto M, Pihlajamaki J, Mykkanen L, Kuusisto J, Laakso M, Fujimoto W, Auwerx J. 1998. A Pro12Ala substitution in PPARgamma2 associated with decreased receptor activity, lower body mass index and improved insulin sensitivity. Nat Genet 20(3):284-7.

Deng HW, Deng H, Liu YJ, Liu YZ, Xu FH, Shen H, Conway T, Li JL, Huang QY, Davies KM, et al. 2002. A genomewide linkage scan for quantitative-trait loci for obesity phenotypes. Am J Hum Genet 70(5):1138-51.

Diamond J. 2003. The double puzzle of diabetes. Nature 423(6940):599-602.

Dick KJ, Nelson CP, Tsaprouni L, Sandling JK, Aissi D, Wahl S, Meduri E, Morange PE, Gagnon F, Grallert H, et al. 2014. DNA methylation and body-mass index: A genome-wide analysis. Lancet 383(9933):1990-8.

Dina C, Meyre D, Gallina S, Durand E, Korner A, Jacobson P, Carlsson LM, Kiess W, Vatin V, Lecoeur C, et al. 2007. Variation in FTO contributes to childhood obesity and severe adult obesity. Nat Genet 39(6):724-6.

Dong C, Wang S, Li WD, Li D, Zhao H, Price RA. 2003. Interacting genetic loci on chromosomes 20 and 10 influence extreme human obesity. Am J Hum Genet 72(1):115-24.

Dorajoo R, Blakemore AI, Sim X, Ong RT, Ng DP, Seielstad M, Wong TY, Saw SM, Froguel P, Liu J, et al. 2012. Replication of 13 obesity loci among singaporean chinese, malay and asian-indian populations. Int J Obes (Lond) 36(1):159-63.

Drong AW, Lindgren CM, McCarthy MI. 2012. The genetic and epigenetic basis of type 2 diabetes and obesity. Clin Pharmacol Ther 92(6):707-15.

Du P, Kibbe WA, Lin SM. 2008. Lumi: A pipeline for processing illumina microarray. Bioinformatics 24(13):1547-8.

Du P, Zhang X, Huang CC, Jafari N, Kibbe WA, Hou L, Lin SM. 2010. Comparison of beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC Bioinformatics 11:587,2105-11-587.

Duan J, Zhang JG, Deng HW, Wang YP. 2013. Comparative studies of copy number variation detection methods for next-generation sequencing technologies. PLoS One 8(3):e59128.

Dubern B, Clement K, Pelloux V, Froguel P, Girardet JP, Guy-Grand B, Tounian P. 2001. Mutational analysis of melanocortin-4 receptor, agouti-related protein, and alpha-melanocyte-stimulating hormone genes in severely obese children. J Pediatr 139(2):204-9.

Eleftherohorinou H, Andersson-Assarsson JC, Walters RG, El-Sayed Moustafa JS, Coin L, Jacobson P, Carlsson LM, Blakemore AI, Froguel P, Walley AJ, et al. 2011. famCNV: Copy number variant association for quantitative traits in families. Bioinformatics 27(13):1873-5.

El-Gharbawy AH, Adler-Wailes DC, Mirch MC, Theim KR, Ranzenhofer L, Tanofsky-Kraff M, Yanovski JA. 2006. Serum brain-derived neurotrophic factor concentrations in lean and overweight children and adolescents. J Clin Endocrinol Metab 91(9):3548-52.

Elks CE, den Hoed M, Zhao JH, Sharp SJ, Wareham NJ, Loos RJ, Ong KK. 2012. Variability in the heritability of body mass index: A systematic review and meta-regression. Front Endocrinol (Lausanne) 3:29.

El-Sayed Moustafa JS and Froguel P. 2013. From obesity genetics to the future of personalized obesity therapy. Nat Rev Endocrinol 9(7):402-13.

Emanuelli B, Peraldi P, Filloux C, Sawka-Verhelle D, Hilton D, Van Obberghen E. 2000. SOCS-3 is an insulin-induced negative regulator of insulin signaling. J Biol Chem 275(21):15985-91.

Falchi M, El-Sayed Moustafa JS, Takousis P, Pesce F, Bonnefond A, Andersson-Assarsson JC, Sudmant PH, Dorajoo R, Al-Shafai MN, Bottolo L, et al. 2014. Low copy number of the salivary amylase gene predisposes to obesity. Nat Genet 46(5):492-7.

Farooqi IS. 2005. Genetic and hereditary aspects of childhood obesity. Best Pract Res Clin Endocrinol Metab 19(3):359-74.

Farooqi IS, Keogh JM, Yeo GS, Lank EJ, Cheetham T, O'Rahilly S. 2003. Clinical spectrum of obesity and mutations in the melanocortin 4 receptor gene. N Engl J Med 348(12):1085-95.

Farooqi IS, Yeo GS, Keogh JM, Aminian S, Jebb SA, Butler G, Cheetham T, O'Rahilly S. 2000. Dominant and recessive inheritance of morbid obesity associated with melanocortin 4 receptor deficiency. J Clin Invest 106(2):271-9.

Farooqi IS, Volders K, Stanhope R, Heuschkel R, White A, Lank E, Keogh J, O'Rahilly S, Creemers JW. 2007. Hyperphagia and early-onset obesity due to a novel homozygous missense mutation in prohormone convertase 1/3. J Clin Endocrinol Metab 92(9):3369-73.

Farooqi IS, Jebb SA, Langmack G, Lawrence E, Cheetham CH, Prentice AM, Hughes IA, McCamish MA, O'Rahilly S. 1999. Effects of recombinant leptin therapy in a child with congenital leptin deficiency. N Engl J Med 341(12):879-84.

Feinberg AP, Irizarry RA, Fradin D, Aryee MJ, Murakami P, Aspelund T, Eiriksdottir G, Harris TB, Launer L, Gudnason V, et al. 2010. Personalized epigenomic signatures that are stable over time and covary with body mass index. Sci Transl Med 2(49):49ra67.

Firth HV, Richards SM, Bevan AP, Clayton S, Corpas M, Rajan D, Van Vooren S, Moreau Y, Pettett RM, Carter NP. 2009. DECIPHER: Database of chromosomal imbalance and phenotype in humans using ensembl resources. Am J Hum Genet 84(4):524-33.

Fischer J, Koch L, Emmerling C, Vierkotten J, Peters T, Bruning JC, Ruther U. 2009. Inactivation of the fto gene protects from obesity. Nature 458(7240):894-8.

Fox CS, Heard-Costa NL, Wilson PW, Levy D, D'Agostino RB S, Atwood LD. 2004. Genome-wide linkage to chromosome 6 for waist circumference in the framingham heart study. Diabetes 53(5):1399-402.

Frayling TM, Timpson NJ, Weedon MN, Zeggini E, Freathy RM, Lindgren CM, Perry JR, Elliott KS, Lango H, Rayner NW, et al. 2007. A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. Science 316(5826):889-94.

Freedman DS, Khan LK, Dietz WH, Srinivasan SR, Berenson GS. 2001. Relationship of childhood obesity to coronary heart disease risk factors in adulthood: The bogalusa heart study. Pediatrics 108(3):712-8.

Friedman JM and Halaas JL. 1998. Leptin and the regulation of body weight in mammals. Nature 395(6704):763-70.

Froguel P and Blakemore AI. 2008. The power of the extreme in elucidating obesity. N Engl J Med 359(9):891-3.

Fujisawa T, Ikegami H, Kawaguchi Y, Ogihara T. 1998. Meta-analysis of the association of Trp64Arg polymorphism of beta 3-adrenergic receptor gene with body mass index. J Clin Endocrinol Metab 83(7):2441-4.

Gamble KL, Berry R, Frank SJ, Young ME. 2014. Circadian clock control of endocrine factors. Nat Rev Endocrinol 10(8):466-75.

Gerken T, Girard CA, Tung YC, Webby CJ, Saudek V, Hewitson KS, Yeo GS, McDonough MA, Cunliffe S, McNeill LA, et al. 2007. The obesity-associated FTO gene encodes a 2-oxoglutarate-dependent nucleic acid demethylase. Science 318(5855):1469-72.

Gibson WT, Farooqi IS, Moreau M, DePaoli AM, Lawrence E, O'Rahilly S, Trussell RA. 2004. Congenital leptin deficiency due to homozygosity for the Delta133G mutation: Report of another case and evaluation of response to four years of leptin therapy. J Clin Endocrinol Metab 89(10):4821-6.

Glessner JT, Bradfield JP, Wang K, Takahashi N, Zhang H, Sleiman PM, Mentch FD, Kim CE, Hou C, Thomas KA, et al. 2010. A genome-wide study reveals copy number variants exclusive to childhood obesity cases. Am J Hum Genet 87(5):661-6.

Godfrey KM, Sheppard A, Gluckman PD, Lillycrop KA, Burdge GC, McLean C, Rodford J, Slater-Jefferies JL, Garratt E, Crozier SR, et al. 2011. Epigenetic gene promoter methylation at birth is associated with child's later adiposity. Diabetes 60(5):1528-34.

Golzio C, Willer J, Talkowski ME, Oh EC, Taniguchi Y, Jacquemont S, Reymond A, Sun M, Sawa A, Gusella JF, et al. 2012. KCTD13 is a major driver of mirrored neuroanatomical phenotypes of the 16p11.2 copy number variant. Nature 485(7398):363-7.

Gorlova OY, Amos CI, Wang NW, Shete S, Turner ST, Boerwinkle E. 2003. Genetic linkage and imprinting effects on body mass index in children and young adults. Eur J Hum Genet 11(6):425-32.

Gray J, Yeo GS, Cox JJ, Morton J, Adlam AL, Keogh JM, Yanovski JA, El Gharbawy A, Han JC, Tung YC, et al. 2006. Hyperphagia, severe obesity, impaired cognitive function, and hyperactivity associated with functional loss of one copy of the brain-derived neurotrophic factor (BDNF) gene. Diabetes 55(12):3366-71.

Greer SN, Metcalf JL, Wang Y, Ohh M. 2012. The updated biology of hypoxia-inducible factor. EMBO J 31(11):2448-60.

Groom A, Potter C, Swan DC, Fatemifar G, Evans DM, Ring SM, Turcot V, Pearce MS, Embleton ND, Smith GD, et al. 2012. Postnatal growth and DNA methylation are associated with differential gene expression of the TACSTD2 gene and childhood fat mass. Diabetes 61(2):391-400.

Gu H, Bock C, Mikkelsen TS, Jager N, Smith ZD, Tomazou E, Gnirke A, Lander ES, Meissner A. 2010. Genome-scale DNA methylation mapping of clinical samples at single-nucleotide resolution. Nat Methods 7(2):133-6.

Guo DF and Rahmouni K. 2011. Molecular basis of the obesity associated with bardet-biedl syndrome. Trends Endocrinol Metab 22(7):286-93.

Guo Y, Lanktree MB, Taylor KC, Hakonarson H, Lange LA, Keating BJ, IBC 50K SNP array BMI Consortium. 2013. Gene-centric meta-analyses of 108 912 individuals confirm known body mass index loci and reveal three novel signals. Hum Mol Genet 22(1):184-201.

Hackett JA, Sengupta R, Zylicz JJ, Murakami K, Lee C, Down TA, Surani MA. 2013. Germline DNA demethylation dynamics and imprint erasure through 5-hydroxymethylcytosine. Science 339(6118):448-52.

Hager J, Dina C, Francke S, Dubois S, Houari M, Vatin V, Vaillant E, Lorentz N, Basdevant A, Clement K, et al. 1998. A genome-wide scan for human obesity genes reveals a major susceptibility locus on chromosome 10. Nat Genet 20(3):304-8.

Hainerova I, Torekov SS, Ek J, Finkova M, Borch-Johnsen K, Jorgensen T, Madsen OD, Lebl J, Hansen T, Pedersen O. 2006. Association between neuromedin U gene variants and overweight and obesity. J Clin Endocrinol Metab 91(12):5057-63.

Halaas JL, Gajiwala KS, Maffei M, Cohen SL, Chait BT, Rabinowitz D, Lallone RL, Burley SK, Friedman JM. 1995. Weight-reducing effects of the plasma protein encoded by the obese gene. Science 269(5223):543-6.

Hanson RL, Ehm MG, Pettitt DJ, Prochazka M, Thompson DB, Timberlake D, Foroud T, Kobes S, Baier L, Burns DK, et al. 1998. An autosomal genomic scan for loci linked to type II diabetes mellitus and body-mass index in pima indians. Am J Hum Genet 63(4):1130-8.

Haslam DW and James WP. 2005. Obesity. Lancet 366(9492):1197-209.

Heard-Costa NL, Zillikens MC, Monda KL, Johansson A, Harris TB, Fu M, Haritunians T, Feitosa MF, Aspelund T, Eiriksdottir G, et al. 2009. NRXN3 is a novel locus for waist circumference: A genome-wide association study from the CHARGE consortium. PLoS Genet 5(6):e1000539.

Heinzen EL, Depondt C, Cavalleri GL, Ruzzo EK, Walley NM, Need AC, Ge D, He M, Cirulli ET, Zhao Q, et al. 2012. Exome sequencing followed by large-scale genotyping fails to identify single rare variants of large effect in idiopathic generalized epilepsy. Am J Hum Genet 91(2):293-302.

Herman JG, Graff JR, Myohanen S, Nelkin BD, Baylin SB. 1996. Methylation-specific PCR: A novel PCR assay for methylation status of CpG islands. Proc Natl Acad Sci U S A 93(18):9821-6.

Hinds DA, Kloek AP, Jen M, Chen X, Frazer KA. 2006. Common deletions and SNPs are in linkage disequilibrium in the human genome. Nat Genet 38(1):82-5.

Hinney A, Nguyen TT, Scherag A, Friedel S, Bronner G, Muller TD, Grallert H, Illig T, Wichmann HE, Rief W, et al. 2007. Genome wide association (GWA) study for early onset extreme obesity supports the role of fat mass and obesity associated gene (FTO) variants. PLoS One 2(12):e1361.

Hinney A, Bettecken T, Tarnow P, Brumm H, Reichwald K, Lichtner P, Scherag A, Nguyen TT, Schlumberger P, Rief W, et al. 2006. Prevalence, spectrum, and functional characterization of melanocortin-4 receptor gene mutations in a representative population-based sample and obese adults from germany. J Clin Endocrinol Metab 91(5):1761-9.

Holder JL,Jr, Butte NF, Zinn AR. 2000. Profound obesity associated with a balanced translocation that disrupts the SIM1 gene. Hum Mol Genet 9(1):101-8.

Holliday R and Pugh JE. 1975. DNA modification mechanisms and gene activity during development. Science 187(4173):226-32.

HOTCHKISS RD. 1948. The quantitative separation of purines, pyrimidines, and nucleosides by paper chromatography. J Biol Chem 175(1):315-32.

Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, Wiencke JK, Kelsey KT. 2012. DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinformatics 13:86,2105-13-86.

Howard AD, Wang R, Pong SS, Mellin TN, Strack A, Guan XM, Zeng Z, Williams DL,Jr, Feighner SD, Nunes CN, et al. 2000. Identification of receptors for neuromedin U and its role in feeding. Nature 406(6791):70-4.

Hu H, Roach JC, Coon H, Guthery SL, Voelkerding KV, Margraf RL, Durtschi JD, Tavtigian SV, Shankaracharya, Wu W, et al. 2014. A unified test of linkage analysis and rare-variant association for analysis of pedigree sequence data. Nat Biotechnol 32(7):663-9.

Hunt SC, Abkevich V, Hensel CH, Gutin A, Neff CD, Russell DL, Tran T, Hong X, Jammulapati S, Riley R, et al. 2001. Linkage of body mass index to chromosome 20 in utah pedigrees. Hum Genet 109(3):279-85.

Hunter-Zinck H, Musharoff S, Salit J, Al-Ali KA, Chouchane L, Gohar A, Matthews R, Butler MW, Fuller J, Hackett NR, et al. 2010. Population genetic structure of the people of qatar. Am J Hum Genet 87(1):17-25.

Huszar D, Lynch CA, Fairchild-Huntress V, Dunmore JH, Fang Q, Berkemeier LR, Gu W, Kesterson RA, Boston BA, Cone RD, et al. 1997. Targeted disruption of the melanocortin-4 receptor results in obesity in mice. Cell 88(1):131-41.

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

Ichimura A, Hirasawa A, Poulain-Godefroy O, Bonnefond A, Hara T, Yengo L, Kimura I, Leloire A, Liu N, Iida K, et al. 2012. Dysfunction of lipid sensor GPR120 leads to obesity in both mouse and human. Nature 483(7389):350-4.

Illumina. BeadArray Microarray Technology [Internet]; c2015 [cited 2014 11/12]. Available from: http://technology.illumina.com/technology/beadarray-technology.html .

Illumina. Infinium Genotyping Data Analysis (Technical Note: Genotyping) [Internet]; c2014 [cited 2014 January 20]. Available from: http://res.illumina.com/documents/products/technotes/technote_infinium_genotyping_data_analysis.pdf .

Illumina. HumanOmni2.5-8 BeadChips (Data Sheet) [Internet]; c2013 [cited 2013 February 17]. Available from: http://www.illumina.com/content/dam/illumina-marketing/documents/products/datasheets/datasheet_human_omni2.5.pdf .

Illumina. Infinium HumanMethylation450 BeadChip (Data Sheet: Epigenetics) [Internet]; c2012a [cited 2014 September 20]. Available from: http://www.illumina.com/Documents/products/datasheets/datasheet_humanmethylation450.pdf .

Illumina. HumanMethylation450 BeadChip Achieves Breadth of Coverage Using Two Infinium® Chemistries (Technical Note) [Internet]; c2012b [cited 2012 January 20]. Available from: http://www.illumina.com/documents/products/technotes/technote_hm450_data_analysis_optimization.pdf .

Illumina. HumanLinkage Panels for Infinium and GoldenGate SNP Genotyping (Data Sheet) [Internet]; c2007 [cited 2012 . Available from: http://supportres.illumina.com/documents/products/datasheets/datasheet_goldengate_infinium_linkage_panels.pdf .

Ingalls AM, Dickie MM, Snell GD. 1950. Obese, a new mutation in the house mouse. J Hered 41(12):317-8.

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

International HapMap Consortium. 2003. The international HapMap project. Nature 426(6968):789-96.

International HapMap Consortium, Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, et al. 2007. A second generation human haplotype map of over 3.1 million SNPs. Nature 449(7164):851-61.

Jackson RS, Creemers JW, Ohagi S, Raffin-Sanson ML, Sanders L, Montague CT, Hutton JC, O'Rahilly S. 1997. Obesity and impaired prohormone processing associated with mutations in the human prohormone convertase 1 gene. Nat Genet 16(3):303-6.

Jacquemont S, Reymond A, Zufferey F, Harewood L, Walters RG, Kutalik Z, Martinet D, Shen Y, Valsesia A, Beckmann ND, et al. 2011. Mirror extreme BMI phenotypes associated with gene dosage at the chromosome 16p11.2 locus. Nature 478(7367):97-102.

Jaenisch R and Bird A. 2003. Epigenetic regulation of gene expression: How the genome integrates intrinsic and environmental signals. Nat Genet 33 Suppl:245-54.

Jamshidi Y, Snieder H, Ge D, Spector TD, O'Dell SD. 2007. The SH2B gene is associated with serum leptin and body fat in normal female twins. Obesity (Silver Spring) 15(1):5-9.

Jarick I, Vogel CI, Scherag S, Schafer H, Hebebrand J, Hinney A, Scherag A. 2011. Novel common copy number variation for early onset extreme obesity on chromosome 11q11 identified by a genome-wide analysis. Hum Mol Genet 20(4):840-52.

Jeyakumar SM, Vajreswari A, Giridharan NV. 2006. Chronic dietary vitamin A supplementation regulates obesity in an obese mutant WNIN/ob rat model. Obesity (Silver Spring) 14(1):52-9.

Kallioniemi A, Kallioniemi OP, Sudar D, Rutovitz D, Gray JW, Waldman F, Pinkel D. 1992. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. Science 258(5083):818-21.

Karimi M, Johansson S, Stach D, Corcoran M, Grander D, Schalling M, Bakalkin G, Lyko F, Larsson C, Ekstrom TJ. 2006. LUMA (LUminometric methylation assay)--a high throughput method to the analysis of genomic DNA methylation. Exp Cell Res 312(11):1989-95.

Khor CC, Chapman SJ, Vannberg FO, Dunne A, Murphy C, Ling EY, Frodsham AJ, Walley AJ, Kyrieleis O, Khan A, et al. 2007. A mal functional variant is associated with protection against invasive pneumococcal disease, bacteremia, malaria and tuberculosis. Nat Genet 39(4):523-8.

Kilpelainen TO, Zillikens MC, Stancakova A, Finucane FM, Ried JS, Langenberg C, Zhang W, Beckmann JS, Luan J, Vandenput L, et al. 2011. Genetic variation near IRS1 associates with reduced adiposity and an impaired metabolic profile. Nat Genet 43(8):753-60.

Kirin M, McQuillan R, Franklin CS, Campbell H, McKeigue PM, Wilson JF. 2010. Genomic runs of homozygosity record population history and consanguinity. PLoS One 5(11):e13996.

Kissebah AH, Sonnenberg GE, Myklebust J, Goldstein M, Broman K, James RG, Marks JA, Krakower GR, Jacob HJ, Weber J, et al. 2000. Quantitative trait loci on chromosomes 3 and 17 influence phenotypes of the metabolic syndrome. Proc Natl Acad Sci U S A 97(26):14478-83.

Klose RJ and Bird AP. 2006. Genomic DNA methylation: The mark and its mediators. Trends Biochem Sci 31(2):89-97.

Knowler WC, Pettitt DJ, Saad MF, Bennett PH. 1990. Diabetes mellitus in the pima indians: Incidence, risk factors and pathogenesis. Diabetes Metab Rev 6(1):1-27.

Kohno D, Gao HZ, Muroya S, Kikuyama S, Yada T. 2003. Ghrelin directly interacts with neuropeptide-Y-containing neurons in the rat arcuate nucleus: Ca2+ signaling via protein kinase A and N-type channel-dependent mechanisms and cross-talk with leptin and orexin. Diabetes 52(4):948-56.

Kondo T, Hayakawa T, Shibata T, Sato Y, Toda Y. 1988. Serum levels of pancreatic enzymes in lean and obese subjects. Int J Pancreatol 3(4):241-8.

Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, Kim PM, Palejev D, Carriero NJ, Du L, et al. 2007. Paired-end mapping reveals extensive structural variation in the human genome. Science 318(5849):420-6.

Korbie DJ and Mattick JS. 2008. Touchdown PCR for increased specificity and sensitivity in PCR amplification. Nat Protoc 3(9):1452-6.

Krude H and Gruters A. 2000. Implications of proopiomelanocortin (POMC) mutations in humans: The POMC deficiency syndrome. Trends Endocrinol Metab 11(1):15-22.

Krude H, Biebermann H, Luck W, Horn R, Brabant G, Gruters A. 1998. Severe early-onset obesity, adrenal insufficiency and red hair pigmentation caused by POMC mutations in humans. Nat Genet 19(2):155-7.

Kruglyak L. 1997. The use of a genetic map of biallelic markers in linkage studies. Nat Genet 17(1):21-4.

Kruit JK, Wijesekara N, Westwell-Roper C, Vanmierlo T, de Haan W, Bhattacharjee A, Tang R, Wellington CL, LutJohann D, Johnson JD, et al. 2012. Loss of both ABCA1 and ABCG1 results in increased disturbances in islet sterol homeostasis, inflammation, and impaired beta-cell function. Diabetes 61(3):659-64.

Ku CS, Loy EY, Pawitan Y, Chia KS. 2010. The pursuit of genome-wide association studies: Where are we now? J Hum Genet 55(4):195-206.

Kublaoui BM, Holder JL,Jr, Gemelli T, Zinn AR. 2006. Sim1 haploinsufficiency impairs melanocortin-mediated anorexia and activation of paraventricular nucleus neurons. Mol Endocrinol 20(10):2483-92.

Kuehnen P, Mischke M, Wiegand S, Sers C, Horsthemke B, Lau S, Keil T, Lee YA, Grueters A, Krude H. 2012. An alu element-associated hypermethylation variant of the POMC gene is associated with childhood obesity. PLoS Genet 8(3):e1002543.

Kulkarni H, Kos MZ, Neary J, Dyer TD, Harald HH, Cole SA, Comuzzie AG, Almasy L, Mahaney MC, Curran JE, et al. Novel epigenetic determinants of type 2 diabetes in Mexican American families [Internet]; c2015 [cited 2015 June 24]. Available from: http://hmg.oxfordjournals.org/content/early/2015/06/21/hmg.ddv232.abstract .

Kumar P, Al-Shafai M, Al Muftah WA, Chalhoub N, Elsaid MF, Aleem AA, Suhre K. 2014. Evaluation of SNP calling using single and multiple-sample calling algorithms by validation against array base genotyping and mendelian inheritance. BMC Res Notes 7:747,0500-7-747.

Kumar RA, KaraMohamed S, Sudi J, Conrad DF, Brune C, Badner JA, Gilliam TC, Nowak NJ, Cook EH,Jr, Dobyns WB, et al. 2008. Recurrent 16p11.2 microdeletions in autism. Hum Mol Genet 17(4):628-38.

Kurokawa N, Young EH, Oka Y, Satoh H, Wareham NJ, Sandhu MS, Loos RJ. 2008. The ADRB3 Trp64Arg variant and BMI: A meta-analysis of 44 833 individuals. Int J Obes (Lond) 32(8):1240-9.

Ladd-Acosta C, Aryee MJ, Ordway JM, Feinberg AP. 2010. Comprehensive high-throughput arrays for relative methylation (CHARM). Curr Protoc Hum Genet Chapter 20:Unit 20.1.1-19.

Lan X, Adams C, Landers M, Dudas M, Krissinger D, Marnellos G, Bonneville R, Xu M, Wang J, Huang TH, et al. 2011. High resolution detection and analysis of CpG dinucleotides methylation using MBD-seq technology. PLoS One 6(7):e22226.

Lander E and Kruglyak L. 1995. Genetic dissection of complex traits: Guidelines for interpreting and reporting linkage results. Nat Genet 11(3):241-7.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. 2001. Initial sequencing and analysis of the human genome. Nature 409(6822):860-921.

Langmead B and Salzberg SL. 2012. Fast gapped-read alignment with bowtie 2. Nat Methods 9(4):357-9.

Lee JH, Reed DR, Li WD, Xu W, Joo EJ, Kilker RL, Nanthakumar E, North M, Sakul H, Bell C, et al. 1999. Genome scan for human obesity and linkage to markers in 20q13. Am J Hum Genet 64(1):196-209.

Lemas DJ, Wiener HW, O'Brien DM, Hopkins S, Stanhope KL, Havel PJ, Allison DB, Fernandez JR, Tiwari HK, Boyer BB. 2012. Genetic polymorphisms in carnitine palmitoyltransferase 1A gene are associated with variation in body composition and fasting lipid traits in yup'ik eskimos. J Lipid Res 53(1):175-84.

Li B, Zhan X, Wing MK, Anderson P, Kang HM, Abecasis GR. 2013. QPLOT: A quality assessment tool for next generation sequencing data. Biomed Res Int 2013:865181.

Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo QM, et al. 2009. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature 462(7271):315-22.

Liu J, Morgan M, Hutchison K, Calhoun VD. 2010. A study of the influence of sex on genome wide methylation. PLoS One 5(4):e10028.

Liu Y, Aryee MJ, Padyukov L, Fallin MD, Hesselberg E, Runarsson A, Reinius L, Acevedo N, Taub M, Ronninger M, et al. 2013. Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. Nat Biotechnol 31(2):142-7.

Liu YJ, Rocha-Sanchez SM, Liu PY, Long JR, Lu Y, Elze L, Recker RR, Deng HW. 2004a. Tests of linkage and/or association of the LEPR gene polymorphisms with obesity phenotypes in caucasian nuclear families. Physiol Genomics 17(2):101-6.

Liu YJ, Xu FH, Shen H, Liu YZ, Deng HY, Zhao LJ, Huang QY, Dvornyk V, Conway T, Davies KM, et al. 2004b. A follow-up linkage study for quantitative trait loci contributing to obesity-related phenotypes. J Clin Endocrinol Metab 89(2):875-82.

Lloyd DJ, Bohan S, Gekakis N. 2006. Obesity, hyperphagia and increased metabolic efficiency in Pc1 mutant mice. Hum Mol Genet 15(11):1884-93.

Locke AE, Kahali B, Berndt SI, Justice AE, Pers TH, Day FR, Powell C, Vedantam S, Buchkovich ML, Yang J, et al. 2015. Genetic studies of body mass index yield new insights for obesity biology. Nature 518(7538):197-206.

Locke DP, Sharp AJ, McCarroll SA, McGrath SD, Newman TL, Cheng Z, Schwartz S, Albertson DG, Pinkel D, Altshuler DM, et al. 2006. Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome. Am J Hum Genet 79(2):275-90.

Loos RJ, Barroso I, O'rahilly S, Wareham NJ. 2007. Comment on "A common genetic variant is associated with adult and childhood obesity". Science 315(5809):187; author reply 187.

Loos RJ, Lindgren CM, Li S, Wheeler E, Zhao JH, Prokopenko I, Inouye M, Freathy RM, Attwood AP, Beckmann JS, et al. 2008. Common variants near MC4R are associated with fat mass, weight and risk of obesity. Nat Genet 40(6):768-75.

Lubrano-Berthelier C, Dubern B, Lacorte JM, Picard F, Shapiro A, Zhang S, Bertrais S, Hercberg S, Basdevant A, Clement K, et al. 2006. Melanocortin 4 receptor mutations in a large cohort of severely obese adults: Prevalence, functional classification, genotype-phenotype relationship, and lack of association with binge eating. J Clin Endocrinol Metab 91(5):1811-8.

Lucito R, Healy J, Alexander J, Reiner A, Esposito D, Chi M, Rodgers L, Brady A, Sebat J, Troge J, et al. 2003. Representational oligonucleotide microarray analysis: A high-resolution method to detect genome copy number variation. Genome Res 13(10):2291-305.

Luke A, Wu X, Zhu X, Kan D, Su Y, Cooper R. 2003. Linkage for BMI at 3q27 region confirmed in an african-american population. Diabetes 52(5):1284-7.

Lupski JR, de Oca-Luna RM, Slaugenhaupt S, Pentao L, Guzzetta V, Trask BJ, Saucedo-Cardenas O, Barker DF, Killian JM, Garcia CA, et al. 1991. DNA duplication associated with charcot-marie-tooth disease type 1A. Cell 66(2):219-32.

Lustig Y, Barhod E, Ashwal-Fluss R, Gordin R, Shomron N, Baruch-Umansky K, Hemi R, Karasik A, Kanety H. 2014. RNA-binding protein PTB and microRNA-221 coregulate AdipoR1 translation and adiponectin signaling. Diabetes 63(2):433-45.

Ma L, Tataranni PA, Bogardus C, Baier LJ. 2004. Melanocortin 4 receptor gene variation is associated with severe obesity in pima indians. Diabetes 53(10):2696-9.

MacDonald JR, Ziman R, Yuen RK, Feuk L, Scherer SW. 2014. The database of genomic variants: A curated collection of structural variation in the human genome. Nucleic Acids Res 42(Database issue):D986-92.

Maes HH, Neale MC, Eaves LJ. 1997. Genetic and environmental factors in relative body weight and human adiposity. Behav Genet 27(4):325-51.

Magi R and Morris AP. 2010. GWAMA: Software for genome-wide association meta-analysis. BMC Bioinformatics 11:288,2105-11-288.

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, et al. 2009. Finding the missing heritability of complex diseases. Nature 461(7265):747-53.

Marabita F, Almgren M, Lindholm ME, Ruhrmann S, Fagerstrom-Billai F, Jagodic M, Sundberg CJ, Ekstrom TJ, Teschendorff AE, Tegner J, et al. 2013. An evaluation of analysis pipelines for DNA methylation profiling using the illumina HumanMethylation450 BeadChip platform. Epigenetics 8(3):333-46.

Mauldin JP, Nagelin MH, Wojcik AJ, Srinivasan S, Skaflen MD, Ayers CR, McNamara CA, Hedrick CC. 2008. Reduced expression of ATP-binding cassette transporter G1 increases cholesterol accumulation in macrophages of patients with type 2 diabetes mellitus. Circulation 117(21):2785-92.

Mazen I, El-Gammal M, Abdel-Hamid M, Farooqi IS, Amr K. 2011. Homozygosity for a novel missense mutation in the leptin receptor gene (P316T) in two egyptian cousins with severe early onset obesity. Mol Genet Metab 102(4):461-4.

McCarroll SA and Altshuler DM. 2007. Copy-number variation and association studies of human disease. Nat Genet 39(7 Suppl):S37-42.

McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, Hirschhorn JN. 2008. Genome-wide association studies for complex traits: Consensus, uncertainty and challenges. Nat Rev Genet 9(5):356-69.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. 2010. The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20(9):1297-303.

McRae AF, Powell JE, Henders AK, Bowdler L, Hemani G, Shah S, Painter JN, Martin NG, Visscher PM, Montgomery GW. 2014. Contribution of genetic variation to transgenerational inheritance of DNA methylation. Genome Biol 15(5):R73.

Mejia-Benitez MA, Bonnefond A, Yengo L, Huyvaert M, Dechaume A, Peralta-Romero J, Klunder-Klunder M, Garcia Mena J, El-Sayed Moustafa JS, Falchi M, et al. 2015. Beneficial effect of a high number of copies of salivary amylase AMY1 gene on obesity risk in mexican children. Diabetologia 58(2):290-4.

Meselson M, Yuan R, Heywood J. 1972. Restriction and modification of DNA. Annu Rev Biochem 41:447-66.

Meyre D, Lecoeur C, Delplanque J, Francke S, Vatin V, Durand E, Weill J, Dina C, Froguel P. 2004. A genome-wide scan for childhood obesity-associated traits in french families shows significant linkage on chromosome 6q22.31-q23.2. Diabetes 53(3):803-11.

Meyre D, Delplanque J, Chevre JC, Lecoeur C, Lobbens S, Gallina S, Durand E, Vatin V, Degraeve F, Proenca C, et al. 2009. Genome-wide association study for early-onset and morbid adult obesity identifies three new risk loci in european populations. Nat Genet 41(2):157-9.

Michels KB, Binder AM, Dedeurwaerder S, Epstein CB, Greally JM, Gut I, Houseman EA, Izzi B, Kelsey KT, Meissner A, et al. 2013. Recommendations for the design and analysis of epigenome-wide association studies. Nat Methods 10(10):949-55.

Miller CA, Hampton O, Coarfa C, Milosavljevic A. 2011. ReadDepth: A parallel R package for detecting copy number alterations from short sequencing reads. PLoS One 6(1):e16327.

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

Moayyeri A, Hammond CJ, Valdes AM, Spector TD. 2013. Cohort profile: TwinsUK and healthy ageing twin study. Int J Epidemiol 42(1):76-85.

Mohn F, Weber M, Schubeler D, Roloff TC. 2009. Methylated DNA immunoprecipitation (MeDIP). Methods Mol Biol 507:55-64.

Moise AR, Lobo GP, Erokwu B, Wilson DL, Peck D, Alvarez S, Dominguez M, Alvarez R, Flask CA, de Lera AR, et al. 2010. Increased adiposity in the retinol saturase-knockout mouse. FASEB J 24(4):1261-70.

Monda KL, Chen GK, Taylor KC, Palmer C, Edwards TL, Lange LA, Ng MC, Adeyemo AA, Allison MA, Bielak LF, et al. 2013. A meta-analysis identifies new loci associated with body mass index in individuals of african ancestry. Nat Genet 45(6):690-6.

Montagne L, Raimondo A, Delobel B, Duban-Bedu B, Noblet FS, Dechaume A, Bersten DC, Meyre D, Whitelaw ML, Froguel P, et al. 2014. Identification of two novel loss-of-function SIM1 mutations in two overweight children with developmental delay. Obesity (Silver Spring) 22(12):2621-4.

Montague CT, Farooqi IS, Whitehead JP, Soos MA, Rau H, Wareham NJ, Sewter CP, Digby JE, Mohammed SN, Hurst JA, et al. 1997. Congenital leptin deficiency is associated with severe early-onset obesity in humans. Nature 387(6636):903-8.

Morgan HD, Sutherland HG, Martin DI, Whitelaw E. 1999. Epigenetic inheritance at the agouti locus in the mouse. Nat Genet 23(3):314-8.

Mori H, Hanada R, Hanada T, Aki D, Mashima R, Nishinakamura H, Torisu T, Chien KR, Yasukawa H, Yoshimura A. 2004. Socs3 deficiency in the brain elevates leptin sensitivity and confers resistance to diet-induced obesity. Nat Med 10(7):739-43.

Moslehi R, Goldstein AM, Beerman M, Goldin L, Bergen AW, Framingham Heart Study. 2003. A genome-wide linkage scan for body mass index on framingham heart study families. BMC Genet 4 Suppl 1:S97.

Murray SS, Oliphant A, Shen R, McBride C, Steeke RJ, Shannon SG, Rubano T, Kermani BG, Fan JB, Chee MS, et al. 2004. A highly informative SNP linkage panel for human genetic studies. Nat Methods 1(2):113-7.

Need AC, McEvoy JP, Gennarelli M, Heinzen EL, Ge D, Maia JM, Shianna KV, He M, Cirulli ET, Gumbs CE, et al. 2012. Exome sequencing followed by large-scale genotyping suggests a limited role for moderately rare risk factors of strong effect in schizophrenia. Am J Hum Genet 91(2):303-12.

Neel JV. 1999. Diabetes mellitus: A "thrifty" genotype rendered detrimental by "progress"? 1962. Bull World Health Organ 77(8):694,703; discussion 692-3.

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

Ng PC and Henikoff S. 2003. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res 31(13):3812-4.

Ng SF, Lin RC, Laybutt DR, Barres R, Owens JA, Morris MJ. 2010. Chronic high-fat diet in fathers programs beta-cell dysfunction in female rat offspring. Nature 467(7318):963-6.

Nguyen DQ, Webber C, Ponting CP. 2006. Bias of selection on human copy-number variants. PLoS Genet 2(2):e20.

Norman RA, Thompson DB, Foroud T, Garvey WT, Bennett PH, Bogardus C, Ravussin E. 1997. Genomewide search for genes influencing percent body fat in pima indians: Suggestive linkage at chromosome 11q21-q22. pima diabetes gene group. Am J Hum Genet 60(1):166-73.

Norman RA, Tataranni PA, Pratley R, Thompson DB, Hanson RL, Prochazka M, Baier L, Ehm MG, Sakul H, Foroud T, et al. 1998. Autosomal genomic scan for loci linked to obesity and energy metabolism in pima indians. Am J Hum Genet 62(3):659-68.

Norris JM, Langefeld CD, Scherzinger AL, Rich SS, Bookman E, Beck SR, Saad MF, Haffner SM, Bergman RN, Bowden DW, et al. 2005. Quantitative trait loci for abdominal fat and BMI in hispanic-americans and african-americans: The IRAS family study. Int J Obes (Lond) 29(1):67-77.

Ogden CL, Carroll MD, Curtin LR, McDowell MA, Tabak CJ, Flegal KM. 2006. Prevalence of overweight and obesity in the united states, 1999-2004. JAMA 295(13):1549-55.

O'Rahilly S, Farooqi IS, Yeo GS, Challis BG. 2003. Minireview: Human obesity-lessons from monogenic disorders. Endocrinology 144(9):3757-64.

Ozata M, Ozdemir IC, Licinio J. 1999. Human leptin deficiency caused by a missense mutation: Multiple endocrine defects, decreased sympathetic tone, and immune system dysfunction indicate new targets for leptin action, greater central than peripheral resistance to the effects of leptin, and spontaneous correction of leptin-mediated defects. J Clin Endocrinol Metab 84(10):3686-95.

Palmer LJ, Buxbaum SG, Larkin E, Patel SR, Elston RC, Tishler PV, Redline S. 2003. A whole-genome scan for obstructive sleep apnea and obesity. Am J Hum Genet 72(2):340-50.

Panoutsopoulou K and Zeggini E. 2009. Finding common susceptibility variants for complex disease: Past, present and future. Brief Funct Genomic Proteomic 8(5):345-52.

Park H, Kim JI, Ju YS, Gokcumen O, Mills RE, Kim S, Lee S, Suh D, Hong D, Kang HP, et al. 2010. Discovery of common asian copy number variants using integrated high-resolution array CGH and massively parallel DNA sequencing. Nat Genet 42(5):400-5.

Paternoster L, Evans DM, Nohr EA, Holst C, Gaborieau V, Brennan P, Gjesing AP, Grarup N, Witte DR, Jorgensen T, et al. 2011. Genome-wide population-based association study of extremely overweight young adults--the GOYA study. PLoS One 6(9):e24303.

Pearce LR, Atanassova N, Banton MC, Bottomley B, van der Klaauw AA, Revelli JP, Hendricks A, Keogh JM, Henning E, Doree D, et al. 2013. KSR2 mutations are associated with obesity, insulin resistance, and impaired cellular fuel oxidation. Cell 155(4):765-77.

Peier A, Kosinski J, Cox-York K, Qian Y, Desai K, Feng Y, Trivedi P, Hastings N, Marsh DJ. 2009. The antiobesity effects of centrally administered neuromedin U and neuromedin S are mediated predominantly by the neuromedin U receptor 2 (NMUR2). Endocrinology 150(7):3101-9.

Peiffer DA, Le JM, Steemers FJ, Chang W, Jenniges T, Garcia F, Haden K, Li J, Shaw CA, Belmont J, et al. 2006. High-resolution genomic profiling of chromosomal aberrations using infinium whole-genome genotyping. Genome Res 16(9):1136-48.

Perrone L, Marzuillo P, Grandone A, del Giudice EM. 2010. Chromosome 16p11.2 deletions: Another piece in the genetic puzzle of childhood obesity. Ital J Pediatr 36:43.

Perry GH, Ben-Dor A, Tsalenko A, Sampas N, Rodriguez-Revenga L, Tran CW, Scheffer A, Steinfeld I, Tsang P, Yamada NA, et al. 2008. The fine-scale and complex architecture of human copy-number variation. Am J Hum Genet 82(3):685-95.

Petersen AK, Zeilinger S, Kastenmuller G, Romisch-Margl W, Brugger M, Peters A, Meisinger C, Strauch K, Hengstenberg C, Pagel P, et al. 2014. Epigenetics meets metabolomics: An epigenome-wide association study with blood serum metabolic traits. Hum Mol Genet 23(2):534-45.

Pettitt DJ, Baird HR, Aleck KA, Bennett PH, Knowler WC. 1983. Excessive obesity in offspring of pima indian women with diabetes during pregnancy. N Engl J Med 308(5):242-5.

Philippe J, Stijnen P, Meyre D, De Graeve F, Thuillier D, Delplanque J, Gyapay G, Sand O, Creemers JW, Froguel P, et al. 2015. A nonsense loss-of-function mutation in PCSK1 contributes to dominantly inherited human obesity. Int J Obes (Lond) 39(2):295-302.

Pickett LA, Yourshaw M, Albornoz V, Chen Z, Solorzano-Vargas RS, Nelson SF, Martin MG, Lindberg I. 2013. Functional consequences of a novel variant of PCSK1. PLoS One 8(1):e55065.

Piotrowski A, Bruder CE, Andersson R, Diaz de Stahl T, Menzel U, Sandgren J, Poplawski A, von Tell D, Crasto C, Bogdan A, et al. 2008. Somatic mosaicism for copy number variation in differentiated human tissues. Hum Mutat 29(9):1118-24.

Price RA, Li WD, Bernstein A, Crystal A, Golding EM, Weisberg SJ, Zuckerman WA. 2001. A locus affecting obesity in human chromosome region 10p12. Diabetologia 44(3):363-6.

Pritchard JK. 2001. Are rare variants responsible for susceptibility to complex diseases? Am J Hum Genet 69(1):124-37.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, et al. 2007. PLINK: A tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81(3):559-75.

Radford EJ, Ito M, Shi H, Corish JA, Yamazawa K, Isganaitis E, Seisenberger S, Hore TA, Reik W, Erkek S, et al. 2014. In utero effects. in utero undernourishment perturbs the adult sperm methylome and intergenerational metabolism. Science 345(6198):1255903.

Ramachandrappa S, Raimondo A, Cali AM, Keogh JM, Henning E, Saeed S, Thompson A, Garg S, Bochukova EG, Brage S, et al. 2013. Rare variants in single-minded 1 (SIM1) are associated with severe obesity. J Clin Invest 123(7):3042-50.

Ramsahoye BH, Biniszkiewicz D, Lyko F, Clark V, Bird AP, Jaenisch R. 2000. Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a. Proc Natl Acad Sci U S A 97(10):5237-42.

Rankinen T, Zuberi A, Chagnon YC, Weisnagel SJ, Argyropoulos G, Walts B, Perusse L, Bouchard C. 2006. The human obesity gene map: The 2005 update. Obesity (Silver Spring) 14(4):529-644.

Razin A and Cedar H. 1977. Distribution of 5-methylcytosine in chromatin. Proc Natl Acad Sci U S A 74(7):2725-8.

Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, et al. 2006. Global variation in copy number in the human genome. Nature 444(7118):444-54.

Reinius LE, Acevedo N, Joerink M, Pershagen G, Dahlen SE, Greco D, Soderhall C, Scheynius A, Kere J. 2012. Differential DNA methylation in purified human blood cells: Implications for cell lineage and studies on disease susceptibility. PLoS One 7(7):e41361.

Relton CL, Groom A, St Pourcain B, Sayers AE, Swan DC, Embleton ND, Pearce MS, Ring SM, Northstone K, Tobias JH, et al. 2012. DNA methylation patterns in cord blood DNA and body size in childhood. PLoS One 7(3):e31821.

Ren D, Li M, Duan C, Rui L. 2005a. Identification of SH2-B as a key regulator of leptin sensitivity, energy balance, and body weight in mice. Cell Metab 2(2):95-104.

Ren D, Zhou Y, Morris D, Li M, Li Z, Rui L. 2007. Neuronal SH2B1 is essential for controlling energy and glucose homeostasis. J Clin Invest 117(2):397-406.

Ren H, Francis W, Boys A, Chueh AC, Wong N, La P, Wong LH, Ryan J, Slater HR, Choo KH. 2005b. BAC-based PCR fragment microarray: High-resolution detection of chromosomal deletion and duplication breakpoints. Hum Mutat 25(5):476-82.

Rhythm. Rhythm initiates clinical trial of RM-493 for obesity caused by genetic variant [Internet]; c2013 [cited 2014 September 20]. Available from: http://www.rhythmtx.com/NEWS/releases/2013%200924%20Rhythm%20RM-493%20MC4R-Deficiency%20Trial.html .

Ribases M, Gratacos M, Fernandez-Aranda F, Bellodi L, Boni C, Anderluh M, Cristina Cavallini M, Cellini E, Di Bella D, Erzegovesi S, et al. 2005. Association of BDNF with restricting anorexia nervosa and minimum body mass index: A family-based association study of eight european populations. Eur J Hum Genet 13(4):428-34.

Riggs AD. 1975. X inactivation, differentiation, and DNA methylation. Cytogenet Cell Genet 14(1):9-25.

Ristow M, Muller-Wieland D, Pfeiffer A, Krone W, Kahn CR. 1998. Obesity associated with a mutation in a genetic regulator of adipocyte differentiation. N Engl J Med 339(14):953-9.

Ross JS and Cronin M. 2011. Whole cancer genome sequencing by next-generation methods. Am J Clin Pathol 136(4):527-39.

Rotimi CN, Comuzzie AG, Lowe WL, Luke A, Blangero J, Cooper RS. 1999. The quantitative trait locus on chromosome 2 for serum leptin levels is confirmed in african-americans. Diabetes 48(3):643-4.

Rozen S and Skaletsky H. 2000. Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol 132:365-86.

Rui L, Yuan M, Frantz D, Shoelson S, White MF. 2002. SOCS-1 and SOCS-3 block insulin signaling by ubiquitin-mediated degradation of IRS1 and IRS2. J Biol Chem 277(44):42394-8.

Saeed S, Butt TA, Anwer M, Arslan M, Froguel P. 2012. High prevalence of leptin and melanocortin-4 receptor gene mutations in children with severe obesity from pakistani consanguineous families. Mol Genet Metab 106(1):121-6.

Sanchez J, Priego T, Pico C, Ahrens W, De Henauw S, Fraterman A, Marild S, Molnar D, Moreno LA, Peplies J, et al. 2012. Blood cells as a source of transcriptional biomarkers of childhood obesity and its related metabolic alterations: Results of the IDEFICS study. J Clin Endocrinol Metab 97(4):E648-52.

Sandholt CH, Hansen T, Pedersen O. 2012. Beyond the fourth wave of genome-wide obesity association studies. Nutr Diabetes 2:e37.

Saunders CL, Chiodini BD, Sham P, Lewis CM, Abkevich V, Adeyemo AA, de Andrade M, Arya R, Berenson GS, Blangero J, et al. 2007. Meta-analysis of genome-wide linkage studies in BMI and obesity. Obesity (Silver Spring) 15(9):2263-75.

Scherag A, Dina C, Hinney A, Vatin V, Scherag S, Vogel CI, Muller TD, Grallert H, Wichmann HE, Balkau B, et al. 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(4):e1000916.

Schneeman BO, Inman MD, Stern JS. 1983. Pancreatic enzyme activity in obese and lean zucker rats: A developmental study. J Nutr 113(4):921-5.

Schork NJ, Murray SS, Frazer KA, Topol EJ. 2009. Common vs. rare allele hypotheses for complex diseases. Curr Opin Genet Dev 19(3):212-9.

Scully T. 2012. Diabetes in numbers. Nature Outlook 485(7398):S2–S3.

Scuteri A, Sanna S, Chen WM, Uda M, Albai G, Strait J, Najjar S, Nagaraja R, Orru M, Usala G, et al. 2007. Genome-wide association scan shows genetic variants in the FTO gene are associated with obesity-related traits. PLoS Genet 3(7):e115.

Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Maner S, Massa H, Walker M, Chi M, et al. 2004. Large-scale copy number polymorphism in the human genome. Science 305(5683):525-8.

Seki Y, Williams L, Vuguin PM, Charron MJ. 2012. Minireview: Epigenetic programming of diabetes and obesity: Animal models. Endocrinology 153(3):1031-8.

Sha BY, Yang TL, Zhao LJ, Chen XD, Guo Y, Chen Y, Pan F, Zhang ZX, Dong SS, Xu XH, et al. 2009. Genome-wide association study suggested copy number variation may be associated with body mass index in the chinese population. J Hum Genet 54(4):199-202.

Sharp AJ, Locke DP, McGrath SD, Cheng Z, Bailey JA, Vallente RU, Pertz LM, Clark RA, Schwartz S, Segraves R, et al. 2005. Segmental duplications and copy-number variation in the human genome. Am J Hum Genet 77(1):78-88.

Shastry BS. 2009. Copy number variation and susceptibility to human disorders (review). Mol Med Rep 2(2):143-7.

Shi H, Tzameli I, Bjorbaek C, Flier JS. 2004. Suppressor of cytokine signaling 3 is a physiological regulator of adipocyte insulin signaling. J Biol Chem 279(33):34733-40.

Shugart YY, Chen L, Day IN, Lewis SJ, Timpson NJ, Yuan W, Abdollahi MR, Ring SM, Ebrahim S, Golding J, et al. 2009. Two british women studies replicated the association between the Val66Met polymorphism in the brain-derived neurotrophic factor (BDNF) and BMI. Eur J Hum Genet 17(8):1050-5.

Shungin D, Winkler TW, Croteau-Chonka DC, Ferreira T, Locke AE, Magi R, Strawbridge RJ, Pers TH, Fischer K, Justice AE, et al. 2015. New genetic loci link adipose and insulin biology to body fat distribution. Nature 518(7538):187-96.

Silventoinen K, Hasselbalch AL, Lallukka T, Bogl L, Pietilainen KH, Heitmann BL, Schousboe K, Rissanen A, Kyvik KO, Sorensen TI, et al. 2009. Modification effects of physical activity and protein intake on heritability of body size and composition. Am J Clin Nutr 90(4):1096-103.

Smemo S, Tena JJ, Kim KH, Gamazon ER, Sakabe NJ, Gomez-Marin C, Aneas I, Credidio FL, Sobreira DR, Wasserman NF, et al. 2014. Obesity-associated variants within FTO form long-range functional connections with IRX3. Nature 507(7492):371-5.

Smith ZD and Meissner A. 2013. DNA methylation: Roles in mammalian development. Nat Rev Genet 14(3):204-20.

Sobreira NL, Cirulli ET, Avramopoulos D, Wohler E, Oswald GL, Stevens EL, Ge D, Shianna KV, Smith JP, Maia JM, et al. 2010. Whole-genome sequencing of a single proband together with linkage analysis identifies a mendelian disease gene. PLoS Genet 6(6):e1000991.

Souren NY, Tierling S, Fryns JP, Derom C, Walter J, Zeegers MP. 2011. DNA methylation variability at growth-related imprints does not contribute to overweight in monozygotic twins discordant for BMI. Obesity (Silver Spring) 19(7):1519-22.

Spalding KL, Arner E, Westermark PO, Bernard S, Buchholz BA, Bergmann O, Blomqvist L, Hoffstedt J, Naslund E, Britton T, et al. 2008. Dynamics of fat cell turnover in humans. Nature 453(7196):783-7.

Speakman JR. 2010. FTO effect on energy demand versus food intake. Nature 464(7289):E1; discussion E2.

Speliotes EK, Willer CJ, Berndt SI, Monda KL, Thorleifsson G, Jackson AU, Allen HL, Lindgren CM, 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(11):937-48.

Steemers FJ, Chang W, Lee G, Barker DL, Shen R, Gunderson KL. 2006. Whole-genome genotyping with the single-base extension assay. Nat Methods 3(1):31-3.

Stefansson H, Rujescu D, Cichon S, Pietilainen OP, Ingason A, Steinberg S, Fossdal R, Sigurdsson E, Sigmundsson T, Buizer-Voskamp JE, et al. 2008. Large recurrent microdeletions associated with schizophrenia. Nature 455(7210):232-6.

Stein R, Sciaky-Gallili N, Razin A, Cedar H. 1983. Pattern of methylation of two genes coding for housekeeping functions. Proc Natl Acad Sci U S A 80(9):2422-6.

Steinberg MH and Adewoye AH. 2006. Modifier genes and sickle cell anemia. Curr Opin Hematol 13(3):131-6.

Stepanow S, Reichwald K, Huse K, Gausmann U, Nebel A, Rosenstiel P, Wabitsch M, Fischer-Posovszky P, Platzer M. 2011. Allele-specific, age-dependent and BMI-associated DNA methylation of human MCHR1. PLoS One 6(5):e17711.

Stoger R. 2008. Epigenetics and obesity. Pharmacogenomics 9(12):1851-60.

Storey JD. 2002. A direct approach to false discovery rates. Journal of the Royal Statistical Society.Series B (Statistical Methodology) 64(3):479-98.

Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, Redon R, Bird CP, de Grassi A, Lee C, et al. 2007. Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science 315(5813):848-53.

Strobel A, Issad T, Camoin L, Ozata M, Strosberg AD. 1998. A leptin missense mutation associated with hypogonadism and morbid obesity. Nat Genet 18(3):213-5.

Stutzmann F, Tan K, Vatin V, Dina C, Jouret B, Tichet J, Balkau B, Potoczna N, Horber F, O'Rahilly S, et al. 2008. Prevalence of melanocortin-4 receptor deficiency in europeans and their age-dependent penetrance in multigenerational pedigrees. Diabetes 57(9):2511-8.

Sun D, Fu X, Li S, Li Q, Chen T, Lian Y,et al. 2014. Race-specific association between DNA methylation and body mass index: the Bogalusa Heart Study [Internet]; c2014 [cited 2014 . Available from: http://www.ashg.org/2014meeting/abstracts/fulltext/f140122628.htm .

Szpiech ZA, Xu J, Pemberton TJ, Peng W, Zollner S, Rosenberg NA, Li JZ. 2013. Long runs of homozygosity are enriched for deleterious variation. Am J Hum Genet 93(1):90-102.

Tabassum R, Jaiswal A, Chauhan G, Dwivedi OP, Ghosh S, Marwaha RK, Tandon N, Bharadwaj D. 2012. Genetic variant of AMD1 is associated with obesity in urban indian children. PLoS One 7(4):e33162.

Tadmouri GO, Nair P, Obeid T, Al Ali MT, Al Khaja N, Hamamy HA. 2009. Consanguinity and reproductive health among arabs. Reprod Health 6:17,4755-6-17.

Talbert ME, Langefeld CD, Ziegler J, Mychaleckyj JC, Haffner SM, Norris JM, Bowden DW. 2009. Polymorphisms near SOCS3 are associated with obesity and glucose homeostasis traits in hispanic americans from the insulin resistance atherosclerosis family study. Hum Genet 125(2):153-62.

Tan JT, Dorajoo R, Seielstad M, Sim XL, Ong RT, Chia KS, Wong TY, Saw SM, Chew SK, Aung T, et al. 2008. FTO variants are associated with obesity in the chinese and malay populations in singapore. Diabetes 57(10):2851-7.

Tang W, Zou JJ, Chen XF, Zheng JY, Zeng HZ, Liu ZM, Shi YQ. 2011. Association of two polymorphisms within and near SOCS3 gene with obesity in three nationalities in xinjiang province of china. Acta Pharmacol Sin 32(11):1381-6.

Teschendorff AE, Marabita F, Lechner M, Bartlett T, Tegner J, Gomez-Cabrero D, Beck S. 2013. A beta-mixture quantile normalization method for correcting probe design bias in illumina infinium 450 k DNA methylation data. Bioinformatics 29(2):189-96.

Therneau T. Coxme: Mixed Effects Cox Models. R package version 2.2-4. [Internet]; c2015 [cited 2015 April 17]. Available from: http://CRAN.R-project.org/package=coxme .

Therneau T, Atkinson E, Sinnwell T, Schaid D, McDonnell S. Kinship2: Pedigree functions. R package version 1.6.0. [Internet]; c2014 [cited 2015 April 17]. Available from: http://CRAN.R-project.org/package=kinship2 .

Thiele H and Nurnberg P. 2005. HaploPainter: A tool for drawing pedigrees with complex haplotypes. Bioinformatics 21(8):1730-2.

Thompson RF, Fazzari MJ, Greally JM. 2010. Experimental approaches to the study of epigenomic dysregulation in ageing. Exp Gerontol 45(4):255-68.

Thomson JP, Skene PJ, Selfridge J, Clouaire T, Guy J, Webb S, Kerr AR, Deaton A, Andrews R, James KD, et al. 2010. CpG islands influence chromatin structure via the CpG-binding protein Cfp1. Nature 464(7291):1082-6.

Thorleifsson G, Walters GB, Gudbjartsson DF, Steinthorsdottir V, Sulem P, Helgadottir A, Styrkarsdottir U, Gretarsdottir S, Thorlacius S, Jonsdottir I, et al. 2009. Genome-wide association yields new sequence variants at seven loci that associate with measures of obesity. Nat Genet 41(1):18-24.

Tolson KP, Gemelli T, Gautron L, Elmquist JK, Zinn AR, Kublaoui BM. 2010. Postnatal Sim1 deficiency causes hyperphagic obesity and reduced Mc4r and oxytocin expression. J Neurosci 30(10):3803-12.

Tonjes A, Scholz M, Loeffler M, Stumvoll M. 2006. Association of Pro12Ala polymorphism in peroxisome proliferator-activated receptor gamma with pre-diabetic phenotypes: Meta-analysis of 57 studies on nondiabetic individuals. Diabetes Care 29(11):2489-97.

Tsaprouni LG, Yang TP, Bell J, Dick KJ, Kanoni S, Nisbet J, Vinuela A, Grundberg E, Nelson CP, Meduri E, et al. 2014. Cigarette smoking reduces DNA methylation levels at multiple genomic loci but the effect is partially reversible upon cessation. Epigenetics 9(10):1382-96.

Turek FW, Joshu C, Kohsaka A, Lin E, Ivanova G, McDearmon E, Laposky A, Losee-Olson S, Easton A, Jensen DR, et al. 2005. Obesity and metabolic syndrome in circadian clock mutant mice. Science 308(5724):1043-5.

Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, Pertz LM, Haugen E, Hayden H, Albertson D, Pinkel D, et al. 2005. Fine-scale structural variation of the human genome. Nat Genet 37(7):727-32.

Urban AE, Korbel JO, Selzer R, Richmond T, Hacker A, Popescu GV, Cubells JF, Green R, Emanuel BS, Gerstein MB, et al. 2006. High-resolution mapping of DNA copy alterations in human chromosome 22 using high-density tiling oligonucleotide arrays. Proc Natl Acad Sci U S A 103(12):4534-9.

Urnov FD and Wolffe AP. 2001. Above and within the genome: Epigenetics past and present. J Mammary Gland Biol Neoplasia 6(2):153-67.

Vaisse C, Clement K, Guy-Grand B, Froguel P. 1998. A frameshift mutation in human MC4R is associated with a dominant form of obesity. Nat Genet 20(2):113-4.

Villalobos-Comparan M, Villamil-Ramirez H, Villarreal-Molina T, Larrieta-Carrasco E, Leon-Mimila P, Romero-Hidalgo S, Jacobo-Albavera L, Liceaga-Fuentes AE, Campos-Perez FJ, Lopez-Contreras BE, et al. 2012. PCSK1 rs6232 is associated with childhood and adult class III obesity in the mexican population. PLoS One 7(6):e39037.

Vionnet N, Hani EH, Dupont S, Gallina S, Francke S, Dotte S, De Matos F, Durand E, Lepretre F, Lecoeur C, et al. 2000. Genomewide search for type 2 diabetes-susceptibility genes in french whites: Evidence for a novel susceptibility locus for early-onset diabetes on chromosome 3q27-qter and independent replication of a type 2-diabetes locus on chromosome 1q21-q24. Am J Hum Genet 67(6):1470-80.

Walder K, Hanson RL, Kobes S, Knowler WC, Ravussin E. 2000. An autosomal genomic scan for loci linked to plasma leptin concentration in pima indians. Int J Obes Relat Metab Disord 24(5):559-65.

Walley AJ, Asher JE, Froguel P. 2009. The genetic contribution to non-syndromic human obesity. Nat Rev Genet 10(7):431-42.

Walters RG, Coin LJ, Ruokonen A, de Smith AJ, El-Sayed Moustafa JS, Jacquemont S, Elliott P, Esko T, Hartikainen AL, Laitinen J, et al. 2013. Rare genomic structural variants in complex disease: Lessons from the replication of associations with obesity. PLoS One 8(3):e58048.

Walters RG, Jacquemont S, Valsesia A, de Smith AJ, Martinet D, Andersson J, Falchi M, Chen F, Andrieux J, Lobbens S, et al. 2010. A new highly penetrant form of obesity due to deletions on chromosome 16p11.2. Nature 463(7281):671-5.

Wang H, Dong S, Xu H, Qian J, Yang J. 2012. Genetic variants in FTO associated with metabolic syndrome: A meta- and gene-based analysis. Mol Biol Rep 39(5):5691-8.

Wang K, Li WD, Glessner JT, Grant SF, Hakonarson H, Price RA. 2010a. Large copy-number variations are enriched in cases with moderate to extreme obesity. Diabetes 59(10):2690-4.

Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SF, Hakonarson H, Bucan M. 2007. PennCNV: An integrated hidden markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res 17(11):1665-74.

Wang X, Zhu H, Snieder H, Su S, Munn D, Harshfield G, Maria BL, Dong Y, Treiber F, Gutin B, et al. 2010b. Obesity related methylation changes in DNA of peripheral blood leukocytes. BMC Med 8:87,7015-8-87.

Wang YC, McPherson K, Marsh T, Gortmaker SL, Brown M. 2011. Health and economic burden of the projected obesity trends in the USA and the UK. Lancet 378(9793):815-25.

Weisenberger DJ, Campan M, Long TI, Kim M, Woods C, Fiala E, Ehrlich M, Laird PW. 2005. Analysis of repetitive element DNA methylation by MethyLight. Nucleic Acids Res 33(21):6823-36.

Wheeler E, Huang N, Bochukova EG, Keogh JM, Lindsay S, Garg S, Henning E, Blackburn H, Loos RJ, Wareham NJ, et al. 2013. Genome-wide SNP and CNV analysis identifies common and low-frequency variants associated with severe early-onset obesity. Nat Genet .

Willer CJ, Speliotes EK, Loos RJ, Li S, Lindgren CM, Heid IM, Berndt SI, Elliott AL, Jackson AU, Lamina C, et al. 2009. Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. Nat Genet 41(1):25-34.

Willyard C. 2014. Heritability: The family roots of obesity. Nature 508(7496):S58-60.

Wilson AF, Elston RC, Tran LD, Siervogel RM. 1991. Use of the robust sib-pair method to screen for single-locus, multiple-locus, and pleiotropic effects: Application to traits related to hypertension. Am J Hum Genet 48(5):862-72.

Wolffe AP and Matzke MA. 1999. Epigenetics: Regulation through repression. Science 286(5439):481-6.

Wong KK, deLeeuw RJ, Dosanjh NS, Kimm LR, Cheng Z, Horsman DE, MacAulay C, Ng RT, Brown CJ, Eichler EE, et al. 2007. A comprehensive analysis of common copy-number variations in the human genome. Am J Hum Genet 80(1):91-104.

World Health Organization. Obesity and overweight [Internet]; c2015a [cited 2015 Febreuary/3rd]. Available from: http://www.who.int/mediacentre/factsheets/fs311/en/ .

Growth reference 5-19 years [Internet]; c2015b [cited 2015 February 5]. Available from: http://www.who.int/growthref/who2007_bmi_for_age/en/ .

Global health observatory data repository [Internet]; c2015c [cited 2015 February 5]. Available from: http://apps.who.int/gho/data/node.main.A900A?lang=en .

Wu X, Cooper RS, Borecki I, Hanis C, Bray M, Lewis CE, Zhu X, Kan D, Luke A, Curb D. 2002. A combined analysis of genomewide linkage scans for body mass index from the national heart, lung, and blood institute family blood pressure program. Am J Hum Genet 70(5):1247-56.

Xie C and Tammi MT. 2009. CNV-seq, a new method to detect copy number variation using high-throughput sequencing. BMC Bioinformatics 10:80,2105-10-80.

Xu X, Su S, Barnes VA, De Miguel C, Pollock J, Ownby D, Shi H, Zhu H, Snieder H, Wang X. 2013. A genome-wide methylation study on obesity: Differential variability and differential methylation. Epigenetics 8(5):522-33.

Yamauchi T, Nio Y, Maki T, Kobayashi M, Takazawa T, Iwabu M, Okada-Iwabu M, Kawamoto S, Kubota N, Kubota T, et al. 2007. Targeted disruption of AdipoR1 and AdipoR2 causes abrogation of adiponectin binding and metabolic actions. Nat Med 13(3):332-9.

Yang J, Ferreira T, Morris AP, Medland SE, Genetic Investigation of ANthropometric Traits (GIANT) Consortium, DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium, Madden PA, Heath AC, Martin NG, Montgomery GW, et al. 2012. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nat Genet 44(4):369,75, S1-3.

Yaswen L, Diehl N, Brennan MB, Hochgeschwender U. 1999. Obesity in the mouse model of pro-opiomelanocortin deficiency responds to peripheral melanocortin. Nat Med 5(9):1066-70.

Yeo GS, Farooqi IS, Aminian S, Halsall DJ, Stanhope RG, O'Rahilly S. 1998. A frameshift mutation in MC4R associated with dominantly inherited human obesity. Nat Genet 20(2):111-2.

Yoo KY, Shin HR, Chang SH, Choi BY, Hong YC, Kim DH, Kang D, Cho NH, Shin C, Jin YW. 2005. Genomic epidemiology cohorts in korea: Present and the future. Asian Pac J Cancer Prev 6(3):238-43.

Yu Y, Zhu H, Miller DT, Gusella JF, Platt OS, Wu BL, Shen Y, Children's Hospital Boston Genotype Phenotype Study Group. 2011. Age- and gender-dependent obesity in individuals with 16p11.2 deletion. J Genet Genomics 38(9):403-9.

Zaghlool SB, Al-Shafai M, Al Muftah WA, Kumar P, Falchi M, Suhre K. 2015. Association of DNA methylation with age, gender, and smoking in an arab population. Clin Epigenetics 7(1):6,014-0040-6. eCollection 2015.

Zegers D, Beckers S, Hendrickx R, Van Camp JK, de Craemer V, Verrijken A, Van Hoorenbeeck K, Verhulst SL, Rooman RP, Desager KN, et al. 2014. Mutation screen of the SIM1 gene in pediatric patients with early-onset obesity. Int J Obes (Lond) 38(7):1000-4.

Zhang Y, Proenca R, Maffei M, Barone M, Leopold L, Friedman JM. 1994. Positional cloning of the mouse obese gene and its human homologue. Nature 372(6505):425-32.

Zhao J, Goldberg J, Vaccarino V. 2013. Promoter methylation of serotonin transporter gene is associated with obesity measures: A monozygotic twin study. Int J Obes (Lond) 37(1):140-5.

Zhu X, Zhou A, Dey A, Norrbom C, Carroll R, Zhang C, Laurent V, Lindberg I, Ugleholdt R, Holst JJ, et al. 2002. Disruption of PC1/3 expression in mice causes dwarfism and multiple neuroendocrine peptide processing defects. Proc Natl Acad Sci U S A 99(16):10293-8.

Appendices

Appendix 1

Supplementary Materials

Appendix 1.1Study ethical approval



Weill Cornell Medical College in Qatar

Qatar Foundation

WCMC-Q Institutional Review Board
Office of Research

Weill Cornell Medical College in Qatar
Qatar Foundation - Education City
PO Box 24144, Doha, Qatar

APPROVAL NOTICE

26-Sep-2012

PI: Khaled Machaca,
Associate Dean for Research
IRB #: 2012-0025
Study Title: *Genetics of Obesity in Qatari Families*
Approval Period: 26-Sep-2012 - 25-Sep-2013
Submission Type: Initial Review
Approval Type: Expedited

The WCMC-Q Institutional Review Board (IRB) has reviewed and approved the above-referenced protocol, including the following documents:

Please comply with the following codicils imposed upon this approval by the WCMC-Q IRB:

1. The WCMC-Q IRB has determined that the inclusion of children is acceptable under the minimal risk category of the Qatari and US regulations and with the documented permission of at least one parent and documented assent from children 7 years of age or older.

2. The WCMC-Q IRB has determined that the inclusion of pregnant women poses minimal risk and is acceptable under Qatari and US regulations.

As the Principal Investigator of this research project, you are ultimately responsible for:

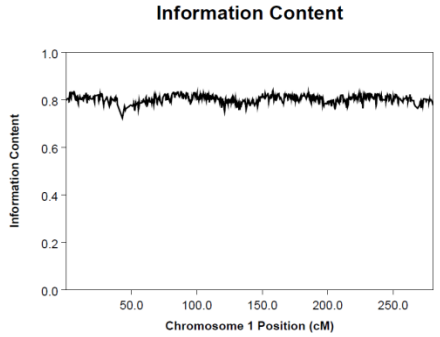
1. The conduct of the study in accordance with the IRB approved protocol;
2. The ethical performance of the project;
3. The protection of the rights and welfare of research subjects;
4. Strict adherence to any stipulations imposed by the IRB; and
5. Compliance with all applicable WCMC-Q and WCMC policies and procedures, as well as with all applicable regulations.

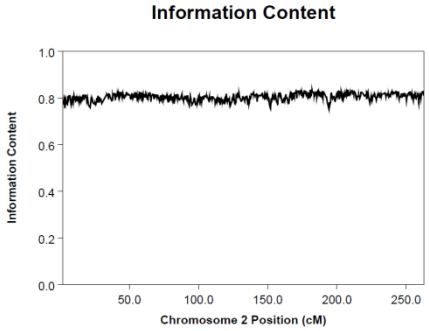
Please contact the WCMC-Q IRB office if you have any questions or need assistance in complying with the terms of this approval.

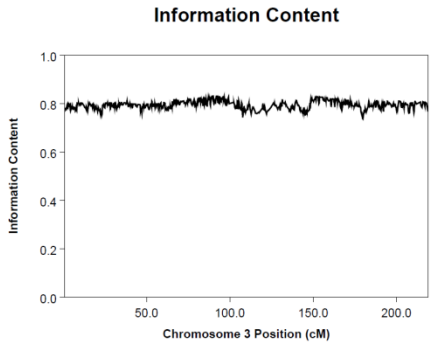
Sincerely,

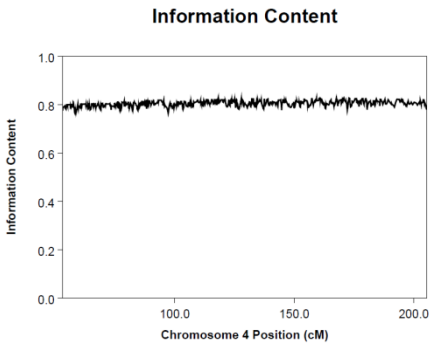
Kip Kantelo, CIP
Director, Research Compliance

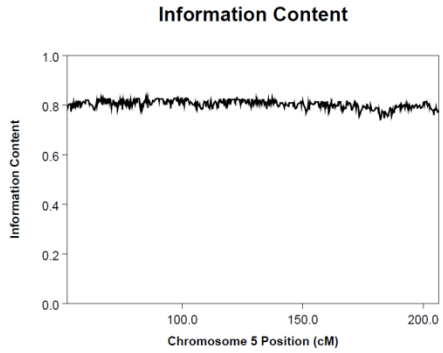
Appendix 1.2 Information content plots given by Method A marker set

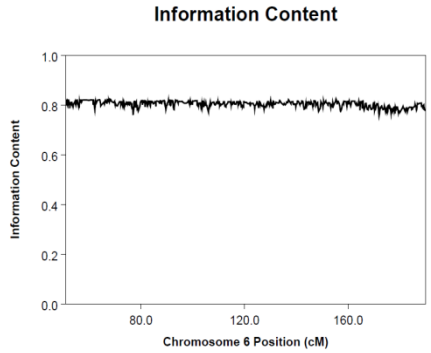


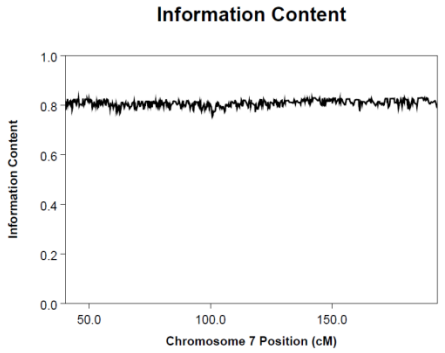


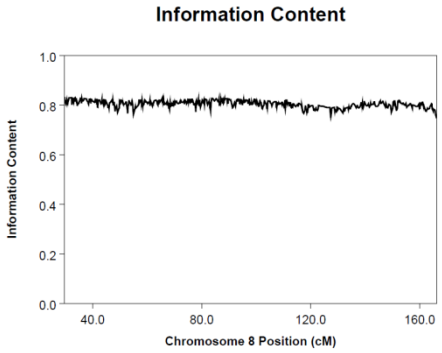


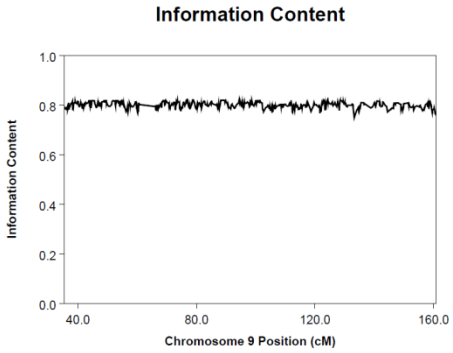


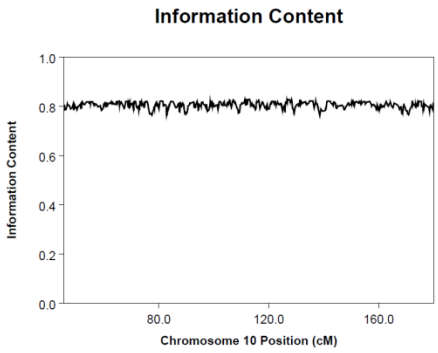


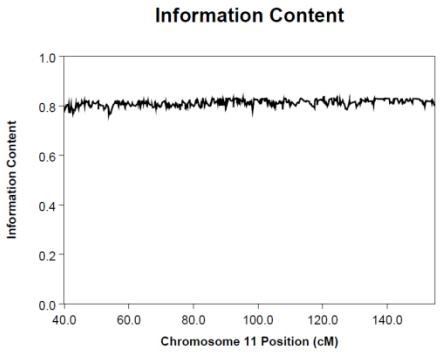


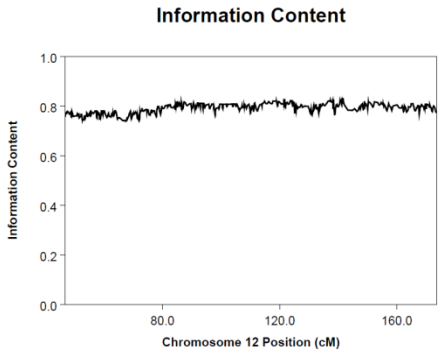

Information Content
Information Content
1.0
0.8
0.6
0.4
0.2
0.0
50.0
100.0
150.0
Chromosome 7 Position (cM)

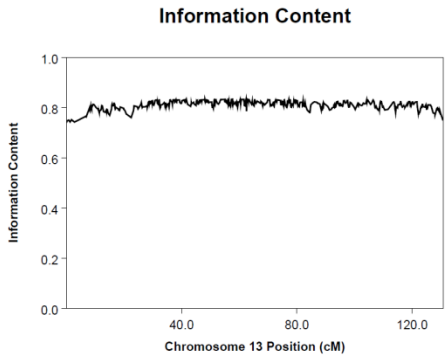

Information Content
Information Content
1.0
0.8
0.6
0.4
0.2
0.0
40.0
80.0
120.0
160.0
Chromosome 8 Position (cM)

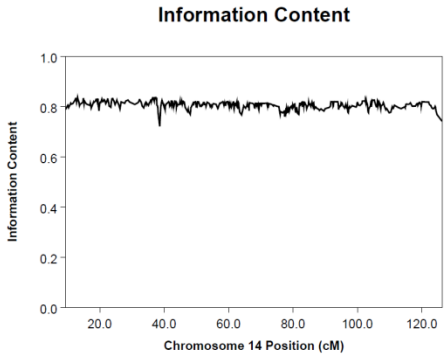

Information Content
Information Content
1.0
0.8
0.6
0.4
0.2
0.0
40.0
80.0
120.0
160.0
Chromosome 9 Position (cM)

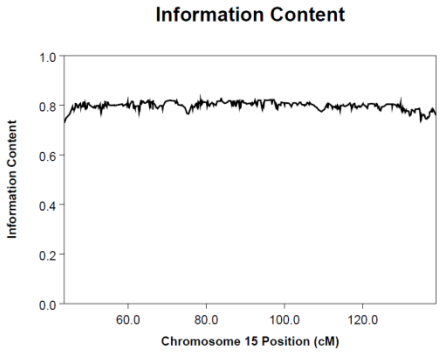

Information Content
Information Content
1.0
0.8
0.6
0.4
0.2
0.0
80.0
120.0
160.0
Chromosome 10 Position (cM)

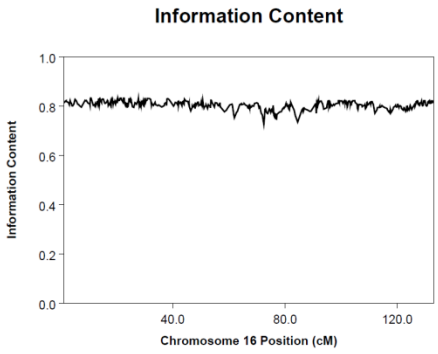

Information Content
Information Content
1.0
0.8
0.6
0.4
0.2
0.0
40.0
60.0
80.0
100.0
120.0
140.0
Chromosome 11 Position (cM)

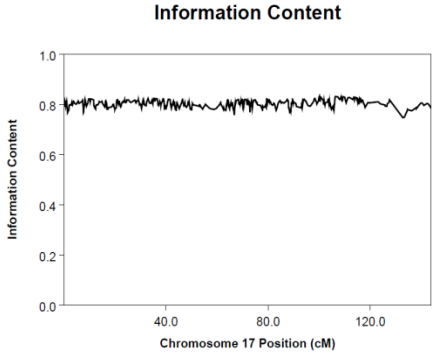

Information Content
Information Content
1.0
0.8
0.6
0.4
0.2
0.0
80.0
120.0
160.0
Chromosome 12 Position (cM)

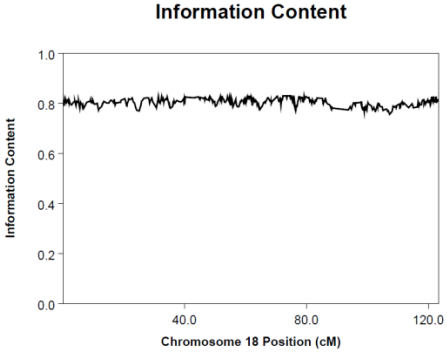

Information Content
Information Content
1.0
0.8
0.6
0.4
0.2
0.0
40.0
80.0
120.0
Chromosome 13 Position (cM)

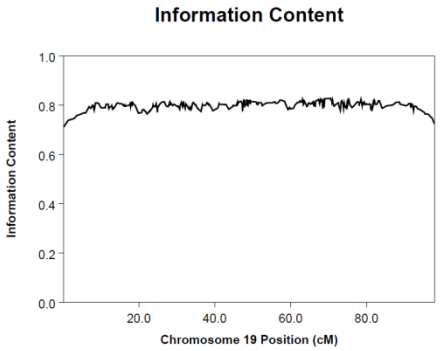

Information Content
Information Content
1.0
0.8
0.6
0.4
0.2
0.0
20.0
40.0
60.0
80.0
100.0
120.0
Chromosome 14 Position (cM)

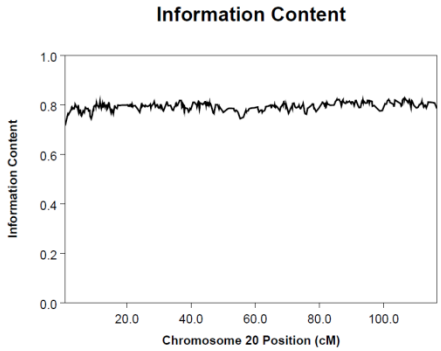

Information Content
Information Content
1.0
0.8
0.6
0.4
0.2
0.0
60.0
80.0
100.0
120.0
Chromosome 15 Position (cM)

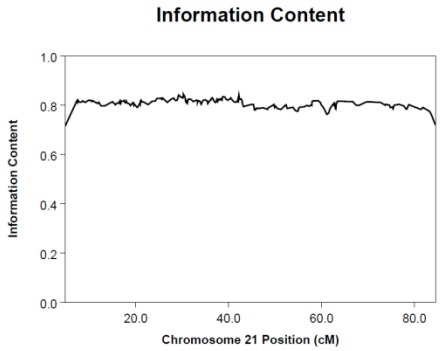

Information Content
Information Content
1.0
0.8
0.6
0.4
0.2
0.0
40.0
80.0
120.0
Chromosome 16 Position (cM)

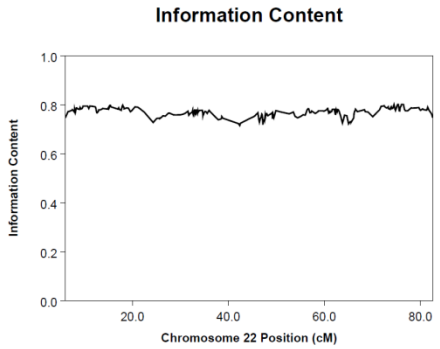

Information Content
Information Content
1.0
0.8
0.6
0.4
0.2
0.0
40.0
80.0
120.0
Chromosome 17 Position (cM)


Information Content
Information Content
1.0
0.8
0.6
0.4
0.2
0.0
40.0
80.0
120.0
Chromosome 18 Position (cM)

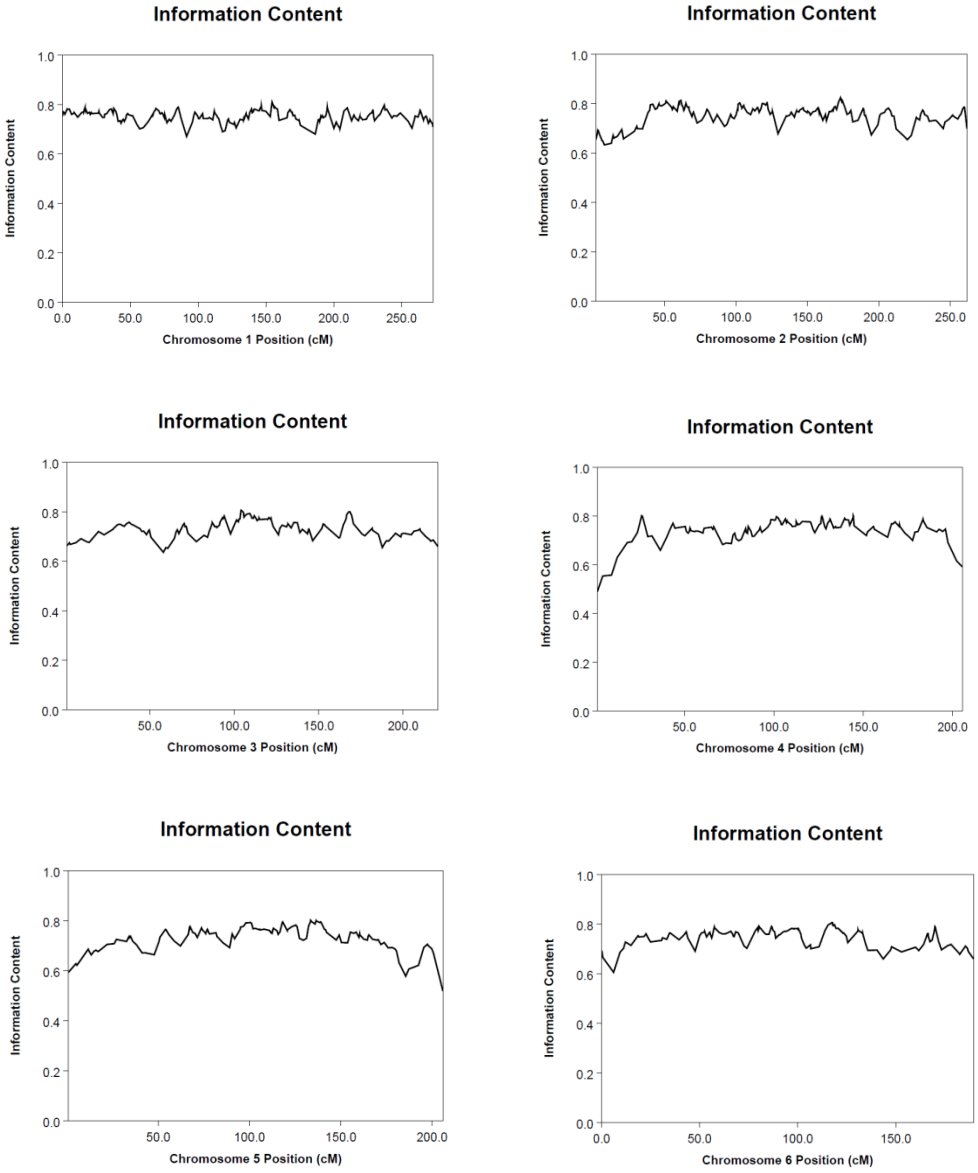

Information Content
Information Content
1.0
0.8
0.6
0.4
0.2
0.0
20.0
40.0
60.0
80.0
Chromosome 19 Position (cM)

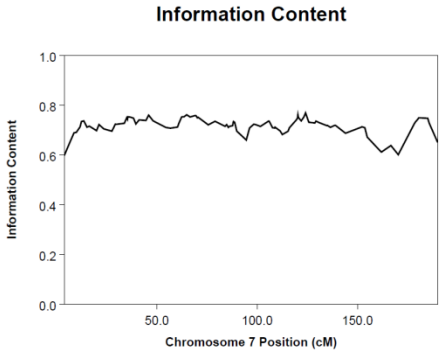

Information Content
Information Content
1.0
0.8
0.6
0.4
0.2
0.0
20.0
40.0
60.0
80.0
100.0
Chromosome 20 Position (cM)

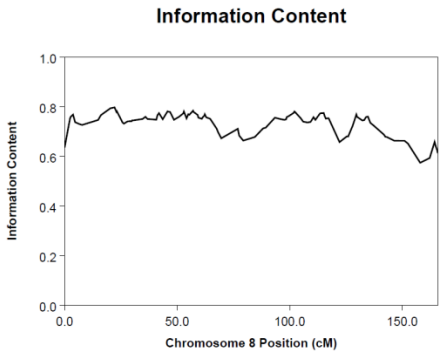

Information Content
Information Content
1.0
0.8
0.6
0.4
0.2
0.0
20.0
40.0
60.0
80.0
Chromosome 21 Position (cM)

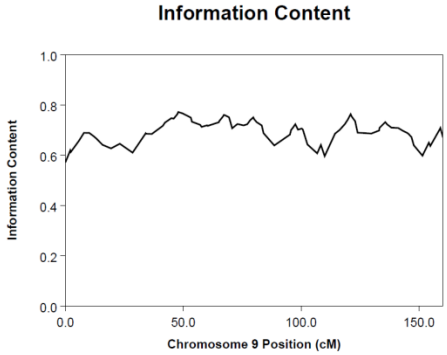

Information Content
Information Content
1.0
0.8
0.6
0.4
0.2
0.0
20.0
40.0
60.0
80.0
Chromosome 22 Position (cM)

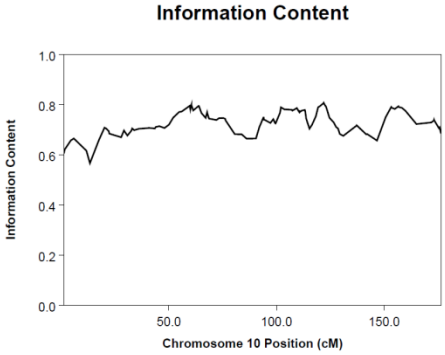
Appendix 1.3 Information content plots given by Method B marker set

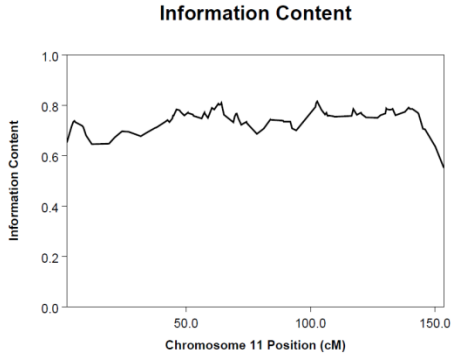


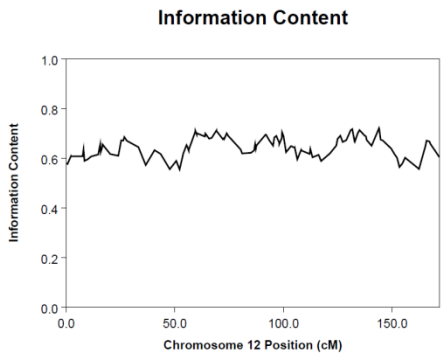

Information Content
Information Content
1.0
0.8
0.6
0.4
0.2
0.0
50.0
100.0
150.0
Chromosome 7 Position (cM)

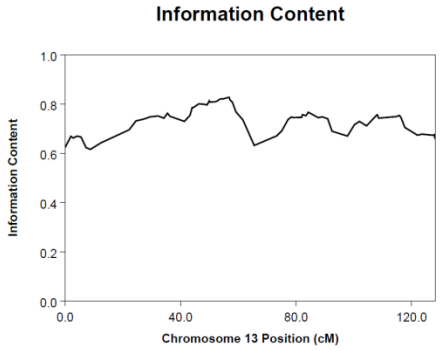

Information Content
Information Content
1.0
0.8
0.6
0.4
0.2
0.0
0.0
50.0
100.0
150.0
Chromosome 8 Position (cM)

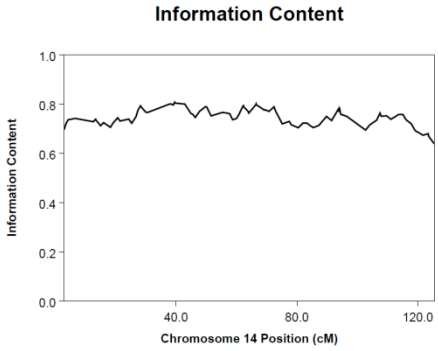

Information Content
Information Content
1.0
0.8
0.6
0.4
0.2
0.0
0.0
50.0
100.0
150.0
Chromosome 9 Position (cM)

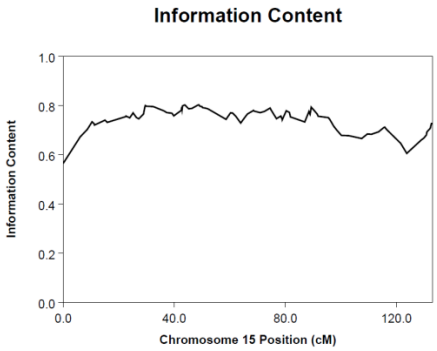

Information Content
Information Content
1.0
0.8
0.6
0.4
0.2
0.0
50.0
100.0
150.0
Chromosome 10 Position (cM)

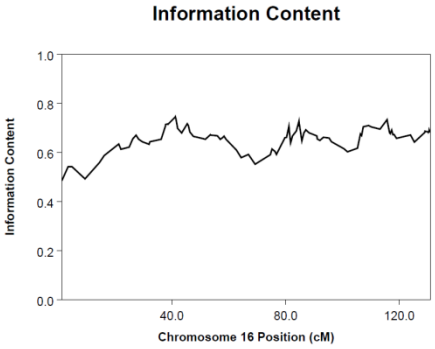

Information Content
Information Content
1.0
0.8
0.6
0.4
0.2
0.0
50.0
100.0
150.0
Chromosome 11 Position (cM)

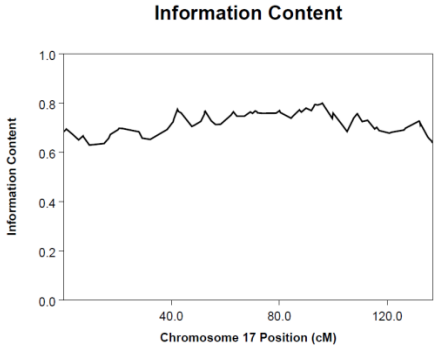

Information Content
Information Content
1.0
0.8
0.6
0.4
0.2
0.0
0.0
50.0
100.0
150.0
Chromosome 12 Position (cM)

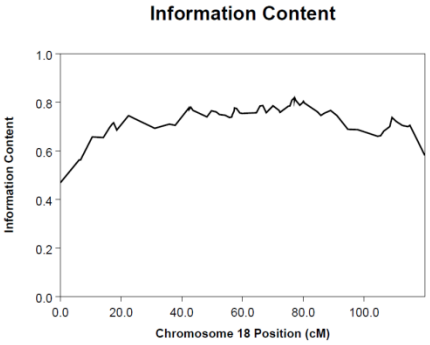

Information Content
Information Content
1.0
0.8
0.6
0.4
0.2
0.0
0.0
40.0
80.0
120.0
Chromosome 13 Position (cM)

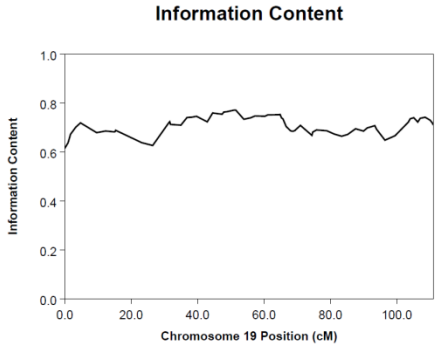

Information Content
Information Content
1.0
0.8
0.6
0.4
0.2
0.0
40.0
80.0
120.0
Chromosome 14 Position (cM)

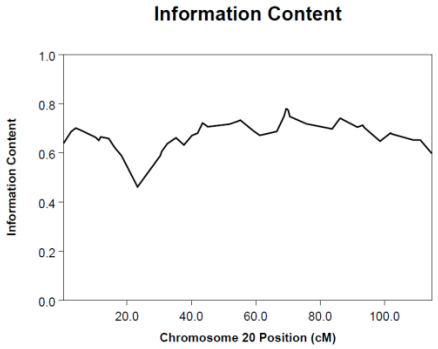

Information Content
Information Content
1.0
0.8
0.6
0.4
0.2
0.0
0.0
40.0
80.0
120.0
Chromosome 15 Position (cM)

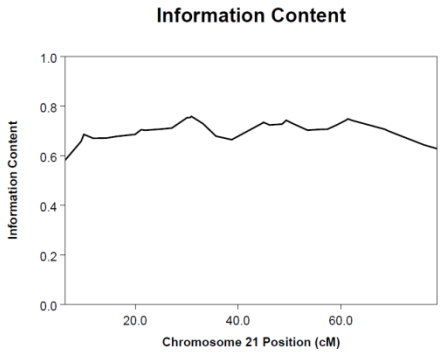

Information Content
Information Content
1.0
0.8
0.6
0.4
0.2
0.0
40.0
80.0
120.0
Chromosome 16 Position (cM)

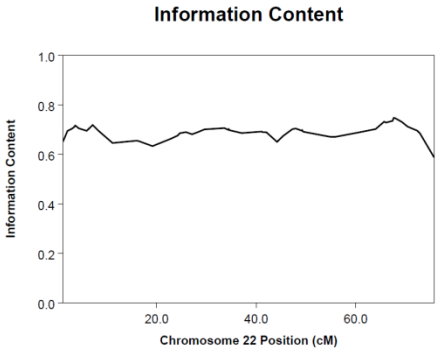

Information Content
Information Content
1.0
0.8
0.6
0.4
0.2
0.0
40.0
80.0
120.0
Chromosome 17 Position (cM)


Information Content
Information Content
1.0
0.8
0.6
0.4
0.2
0.0
0.0
20.0
40.0
60.0
80.0
100.0
Chromosome 18 Position (cM)

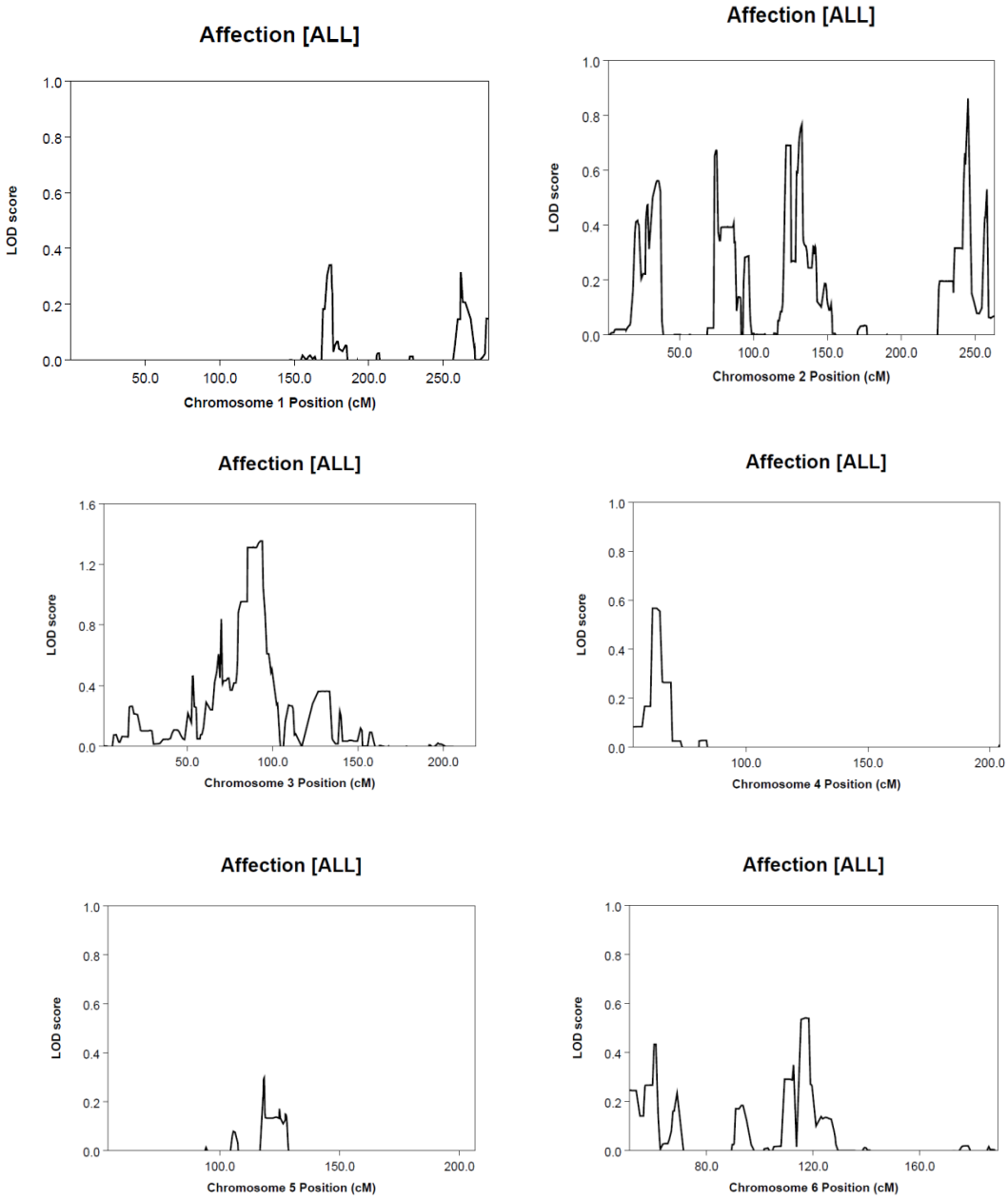

Information Content
Information Content
1.0
0.8
0.6
0.4
0.2
0.0
0.0
20.0
40.0
60.0
80.0
100.0
Chromosome 19 Position (cM)

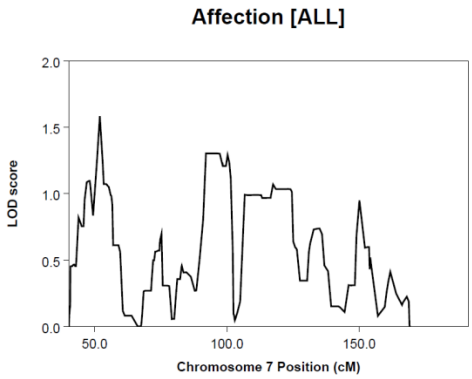

Information Content
Information Content
1.0
0.8
0.6
0.4
0.2
0.0
20.0
40.0
60.0
80.0
100.0
Chromosome 20 Position (cM)

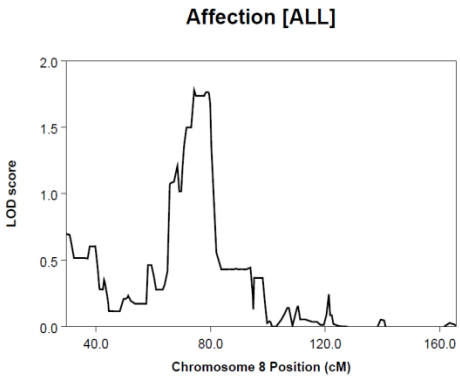

Information Content
Information Content
1.0
0.8
0.6
0.4
0.2
0.0
20.0
40.0
60.0
Chromosome 21 Position (cM)

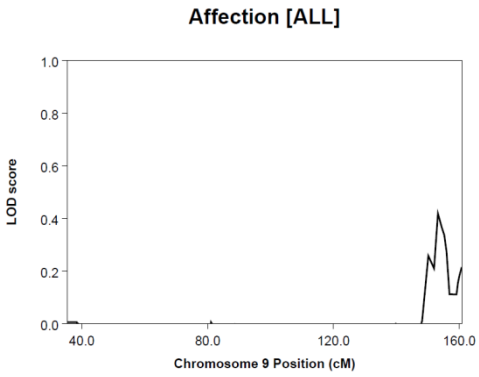

Information Content
Information Content
1.0
0.8
0.6
0.4
0.2
0.0
20.0
40.0
60.0
Chromosome 22 Position (cM)

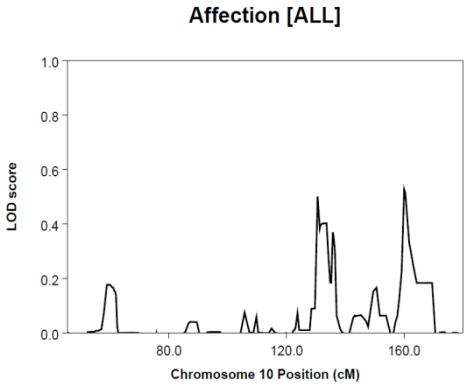
Appendix 1.4 Nonparametric linkage analysis results for the 22 autosomes

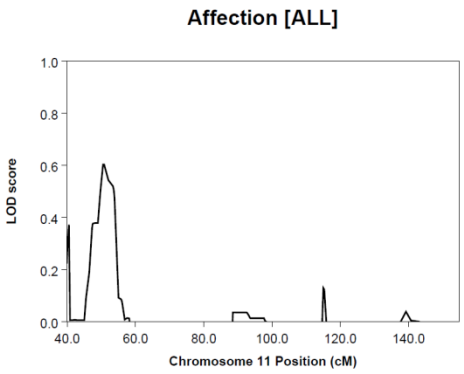


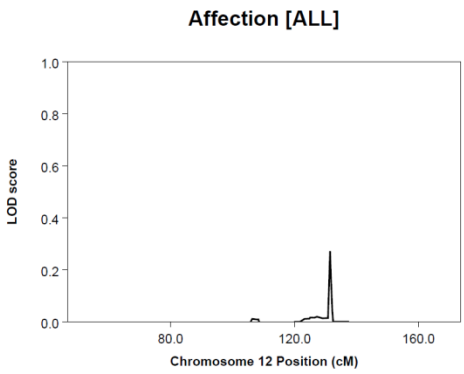

Affection [ALL]
LOD score
2.0
1.5
1.0
0.5
0.0
50.0
100.0
150.0
Chromosome 7 Position (cM)

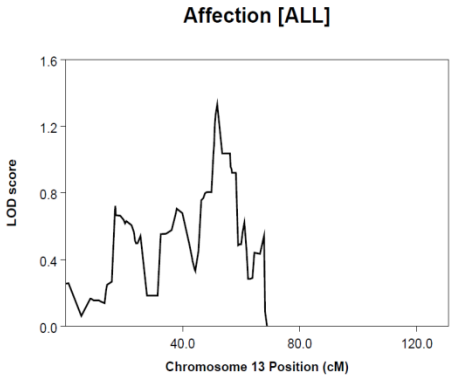

Affection [ALL]
LOD score
2.0
1.5
1.0
0.5
0.0
40.0
80.0
120.0
160.0
Chromosome 8 Position (cM)

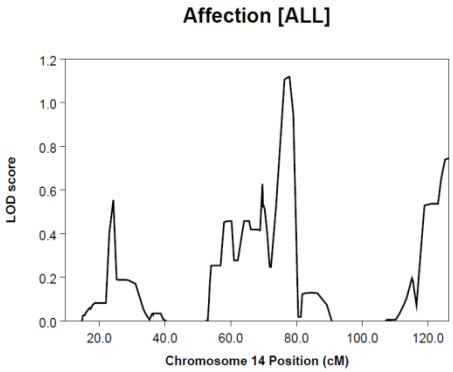

Affection [ALL]
LOD score
1.0
0.8
0.6
0.4
0.2
0.0
40.0
80.0
120.0
160.0
Chromosome 9 Position (cM)

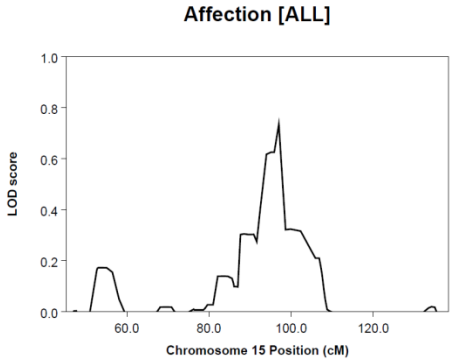

Affection [ALL]
LOD score
1.0
0.8
0.6
0.4
0.2
0.0
80.0
120.0
160.0
Chromosome 10 Position (cM)

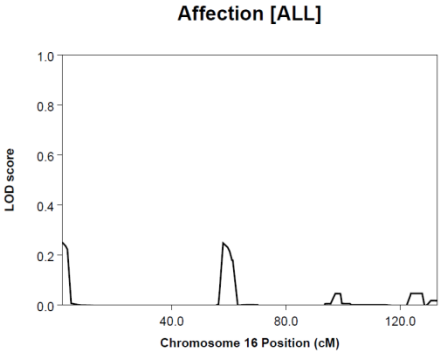

Affection [ALL]
LOD score
1.0
0.8
0.6
0.4
0.2
0.0
40.0
60.0
80.0
100.0
120.0
140.0
Chromosome 11 Position (cM)

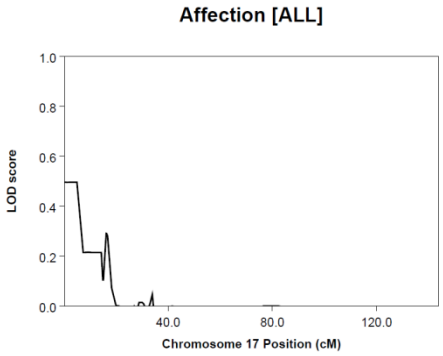

Affection [ALL]
LOD score
1.0
0.8
0.6
0.4
0.2
0.0
80.0
120.0
160.0
Chromosome 12 Position (cM)

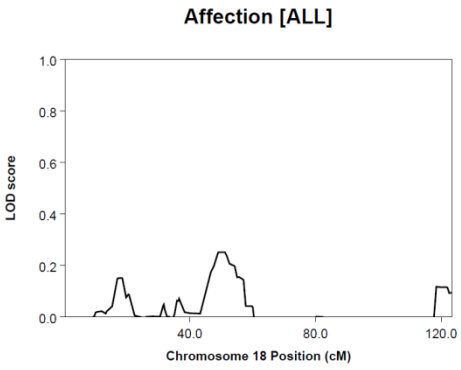

Affection [ALL]
LOD score
1.6
1.2
0.8
0.4
0.0
40.0
80.0
120.0
Chromosome 13 Position (cM)

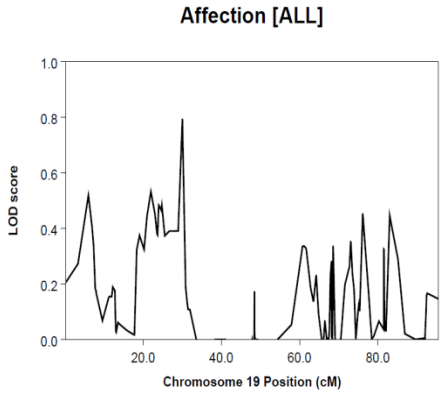

Affection [ALL]
LOD score
1.2
1.0
0.8
0.6
0.4
0.2
0.0
20.0
40.0
60.0
80.0
100.0
120.0
Chromosome 14 Position (cM)

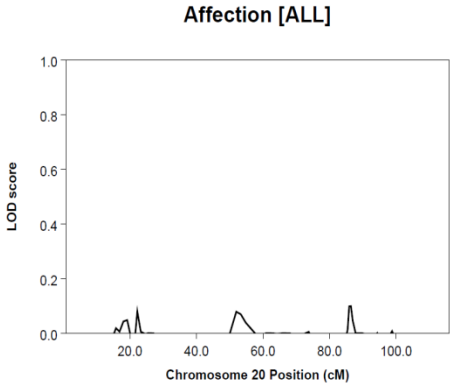

Affection [ALL]
LOD score
1.0
0.8
0.6
0.4
0.2
0.0
60.0
80.0
100.0
120.0
Chromosome 15 Position (cM)

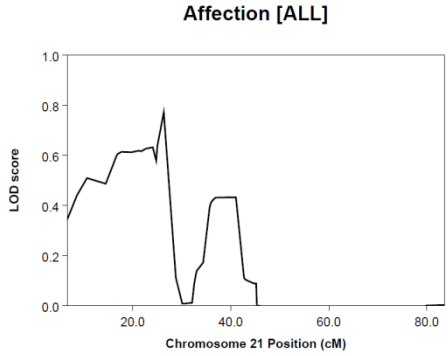

Affection [ALL]
LOD score
1.0
0.8
0.6
0.4
0.2
0.0
40.0
80.0
120.0
Chromosome 16 Position (cM)

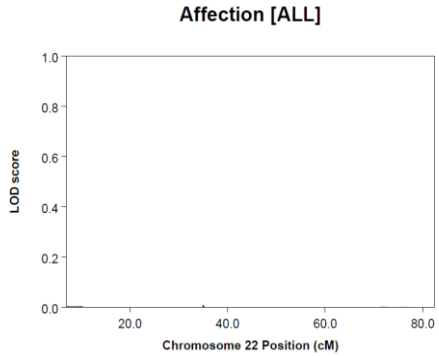

Affection [ALL]
LOD score
1.0
0.8
0.6
0.4
0.2
0.0
40.0
80.0
120.0
Chromosome 17 Position (cM)


Affection [ALL]
LOD score
1.0
0.8
0.6
0.4
0.2
0.0
40.0
80.0
120.0
Chromosome 18 Position (cM)


Affection [ALL]
LOD score
1.0
0.8
0.6
0.4
0.2
0.0
20.0
40.0
60.0
80.0
Chromosome 19 Position (cM)

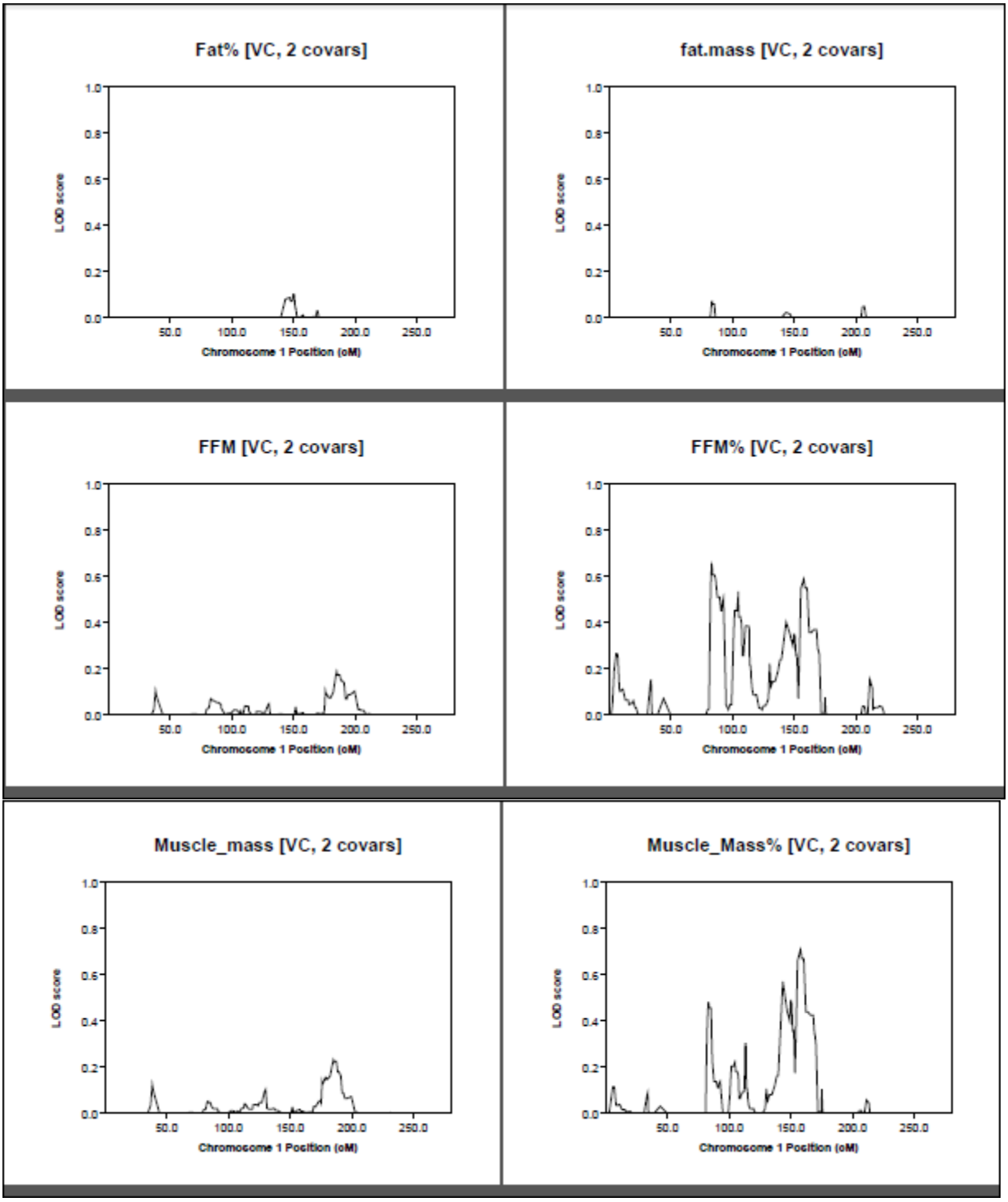

Affection [ALL]
LOD score
1.0
0.8
0.6
0.4
0.2
0.0
20.0
40.0
60.0
80.0
100.0
Chromosome 20 Position (cM)

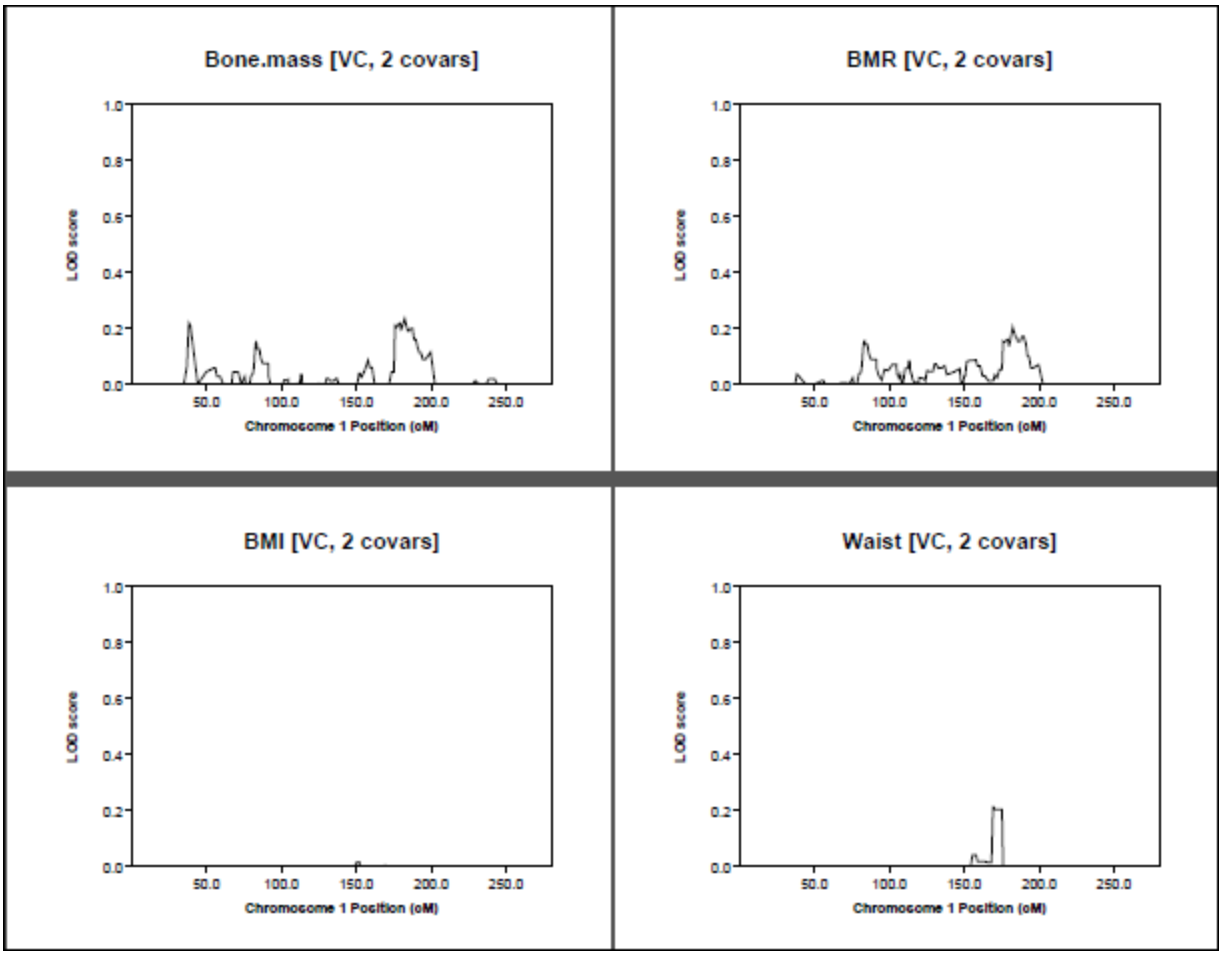

Affection [ALL]
LOD score
1.0
0.8
0.6
0.4
0.2
0.0
20.0
40.0
60.0
80.0
Chromosome 21 Position (cM)


Affection [ALL]
LOD score
1.0
0.8
0.6
0.4
0.2
0.0
20.0
40.0
60.0
80.0
Chromosome 22 Position (cM)

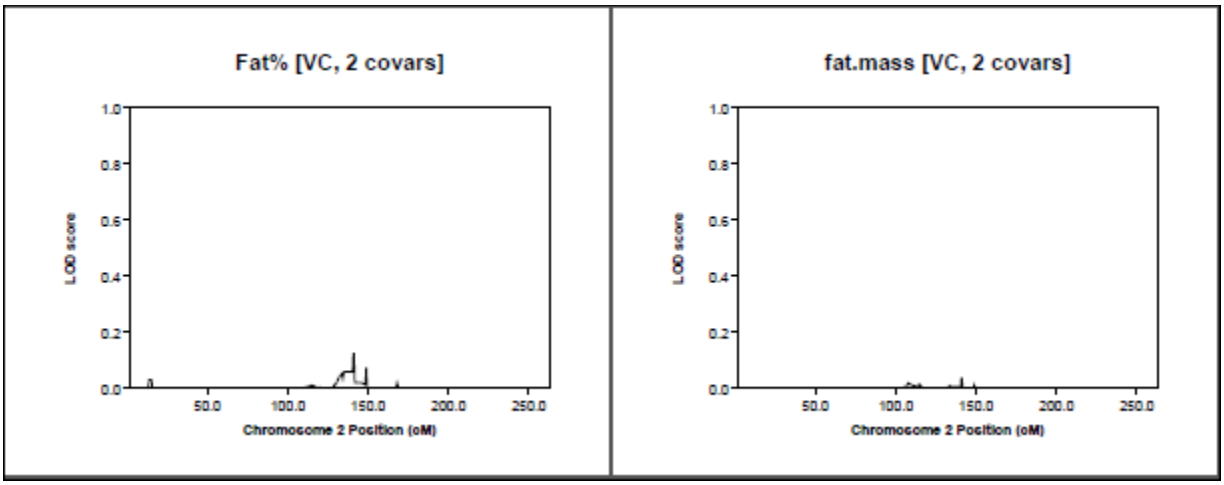
Appendix 1.5 Variance component linkage analysis results for the 22 autosomes

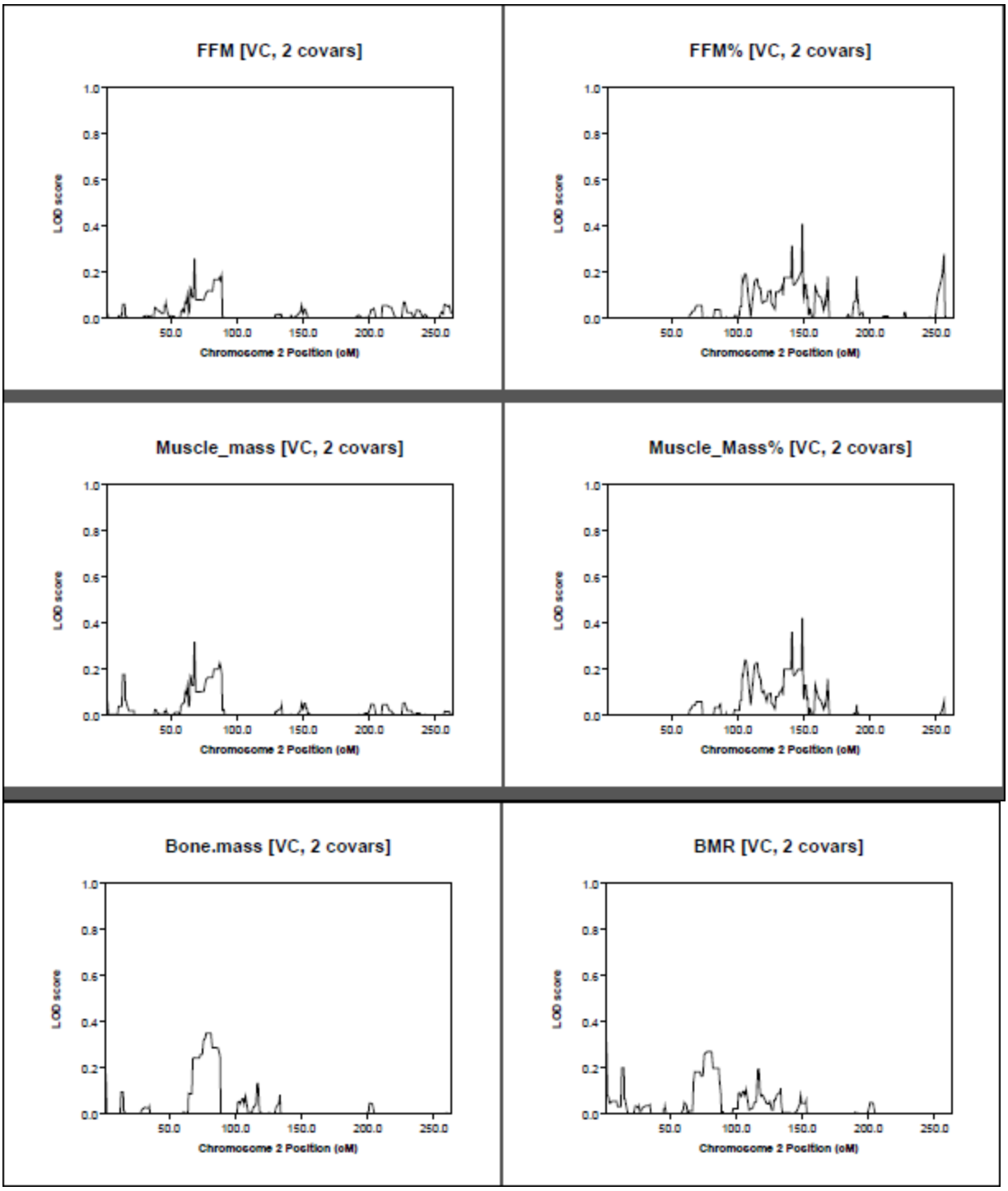
Chromosome 1

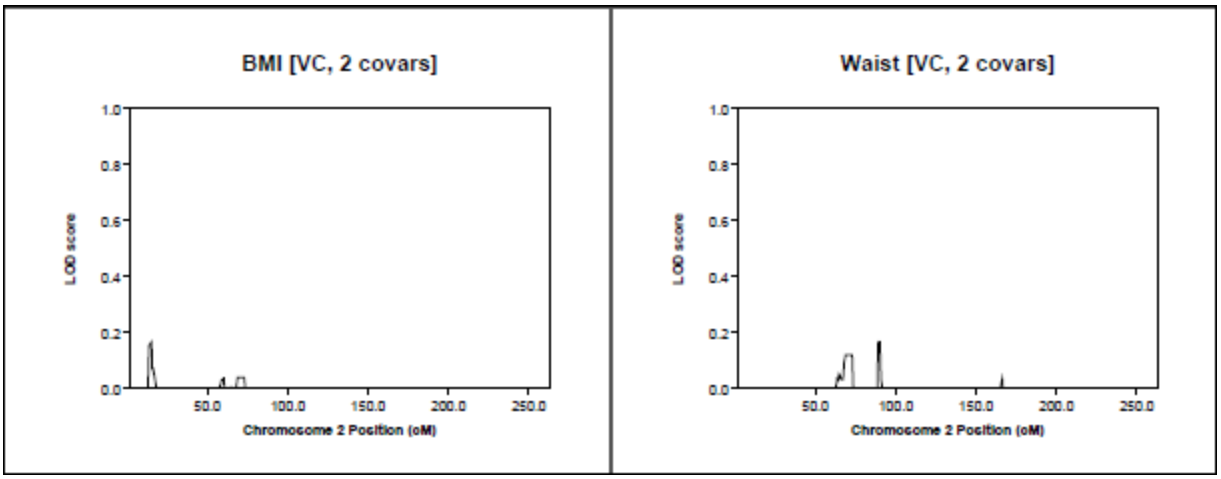




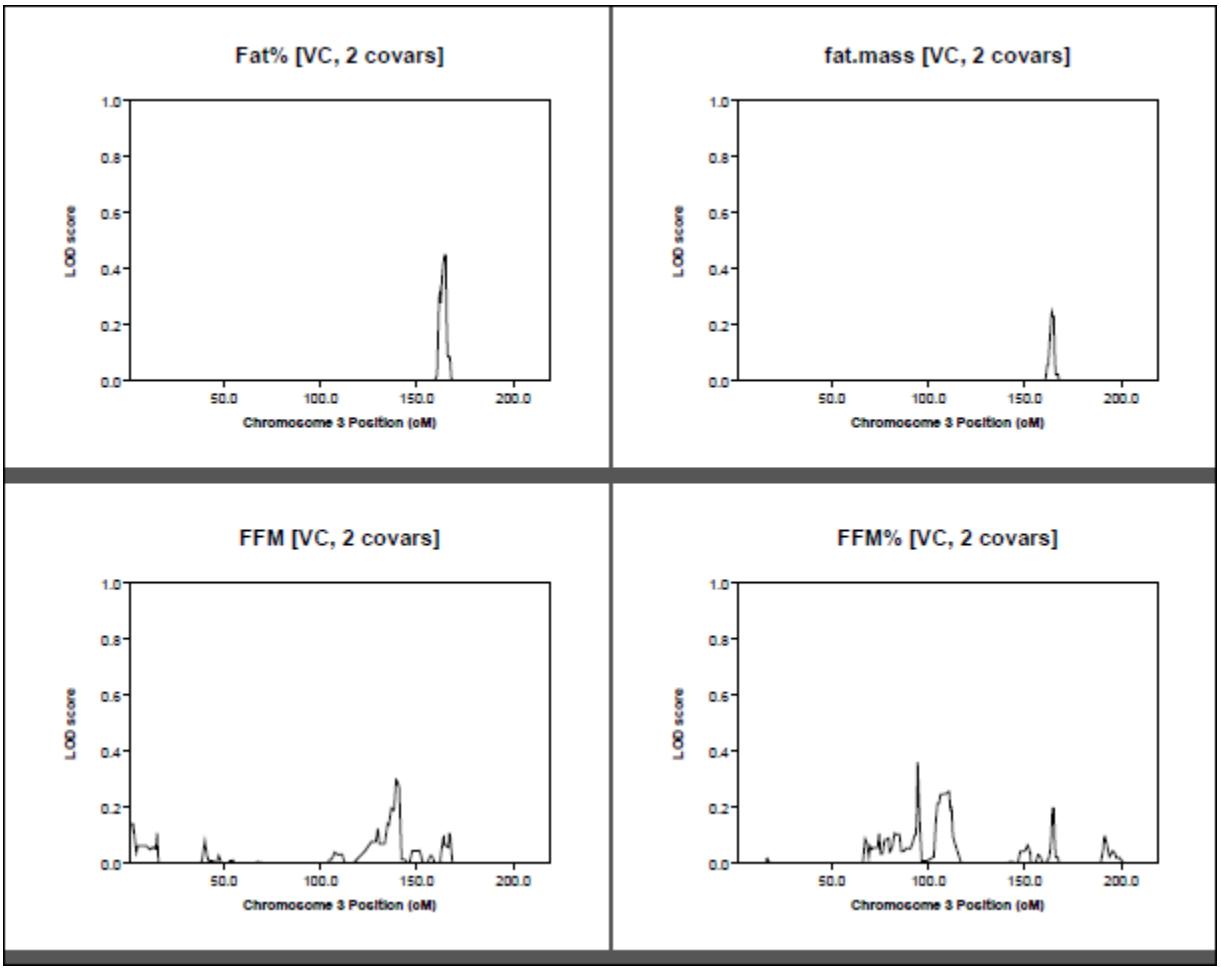
Chromosome 2

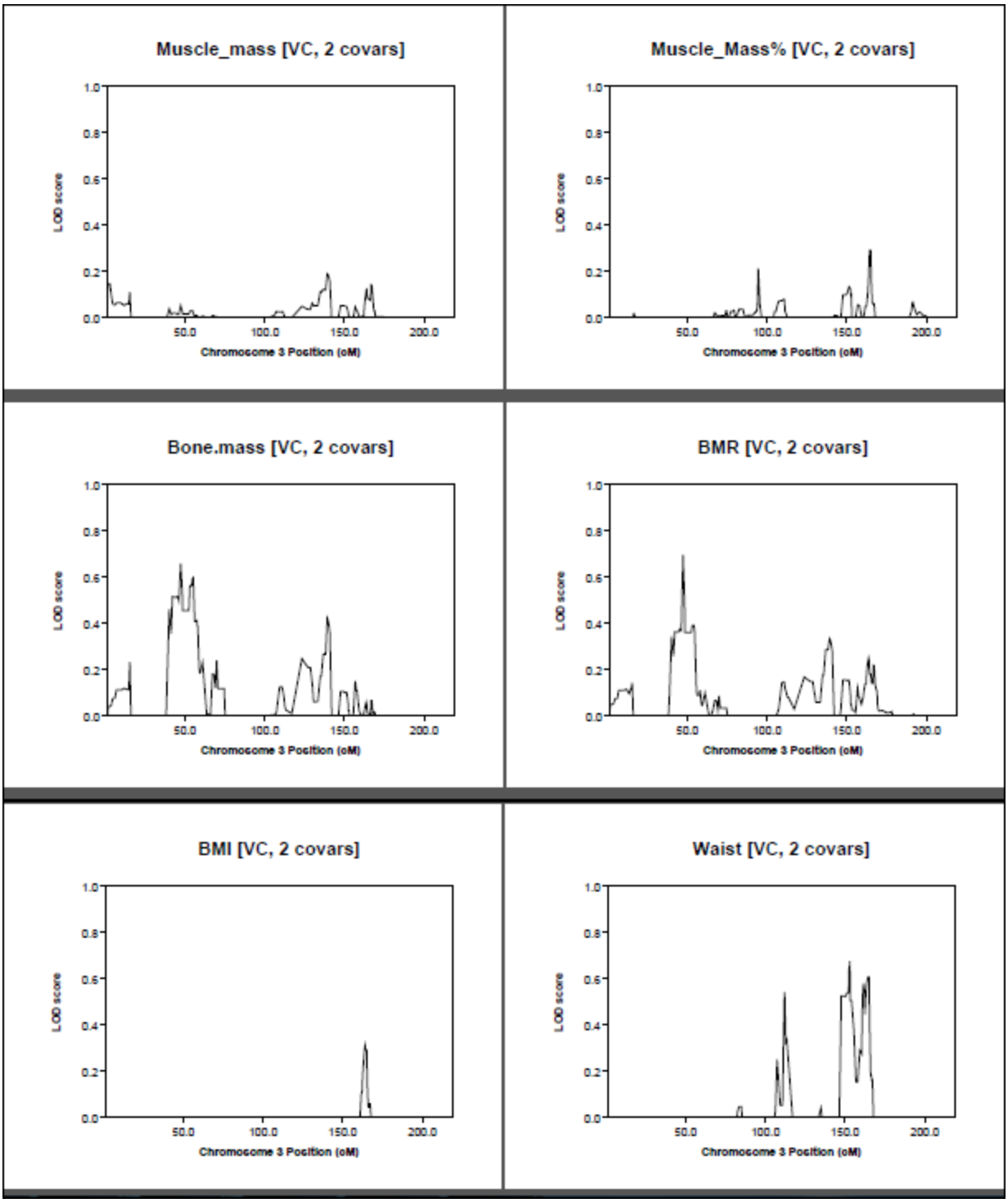



FFM [VC, 2 covars]
FFM% [VC, 2 covars]
Muscle_mass [VC, 2 covars]
Muscle_Mass% [VC, 2 covars]
Bone.mass [VC, 2 covars]
BMR [VC, 2 covars]
LOD score
Chromosome 2 Position (cM)
0.0
0.2
0.4
0.6
0.8
1.0
50.0
100.0
150.0
200.0
250.0

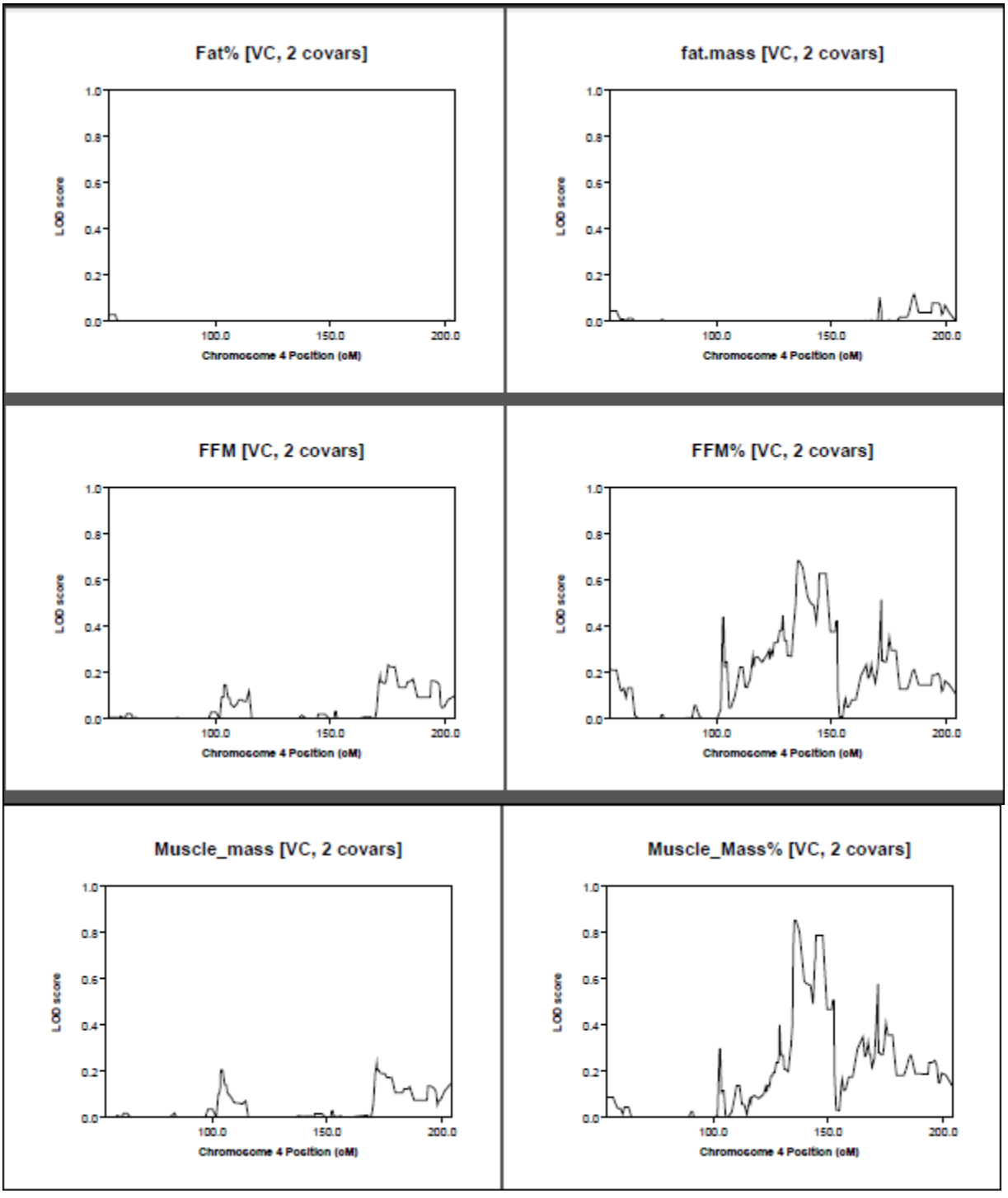


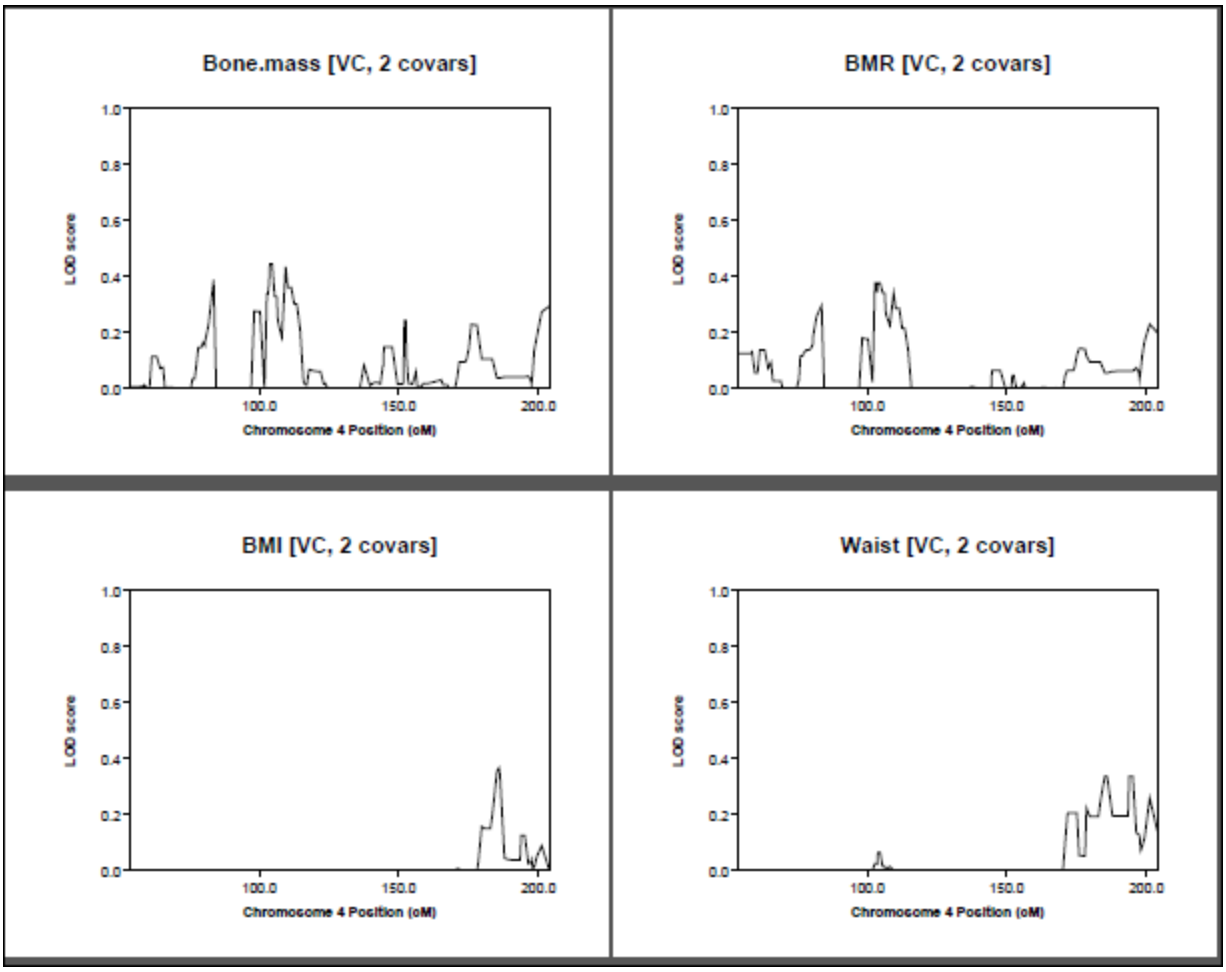
Chromosome 3



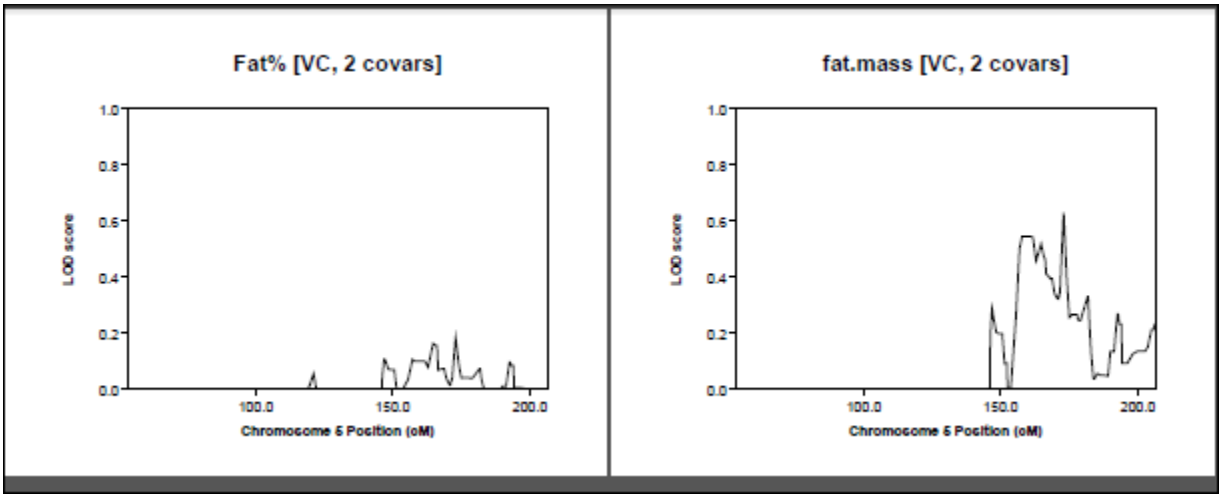

Muscle_mass [VC, 2 covars]
Muscle_Mass% [VC, 2 covars]
Bone.mass [VC, 2 covars]
BMR [VC, 2 covars]
BMI [VC, 2 covars]
Waist [VC, 2 covars]
LOD score
Chromosome 3 Position (cM)
0.0
0.2
0.4
0.6
0.8
1.0
50.0
100.0
150.0
200.0

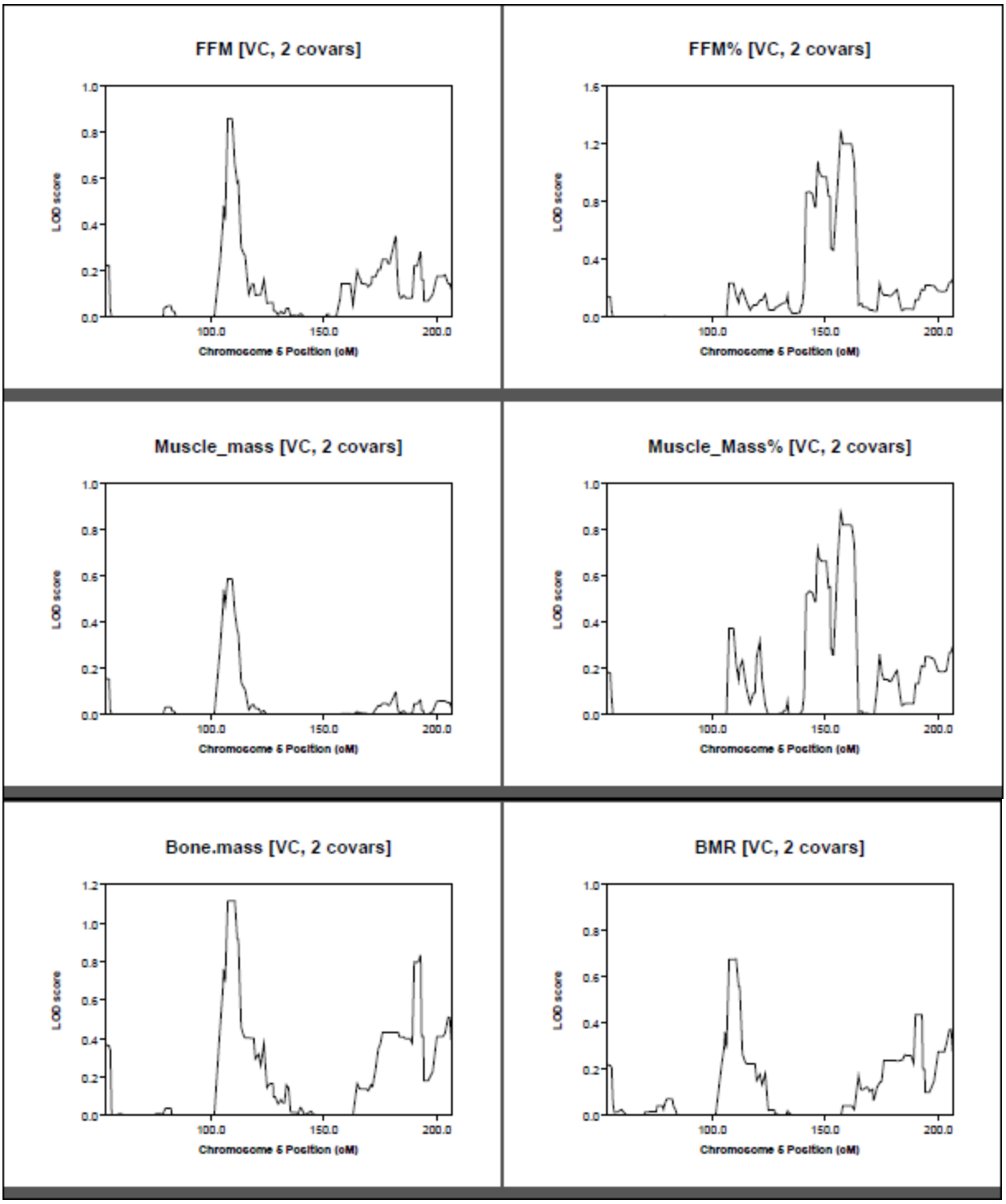
Chromosome 4

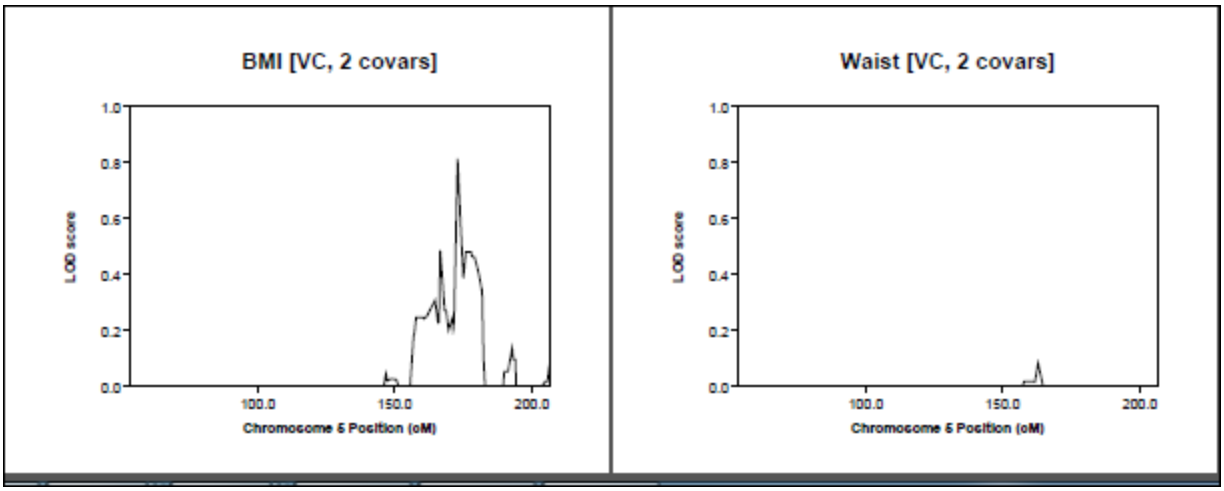




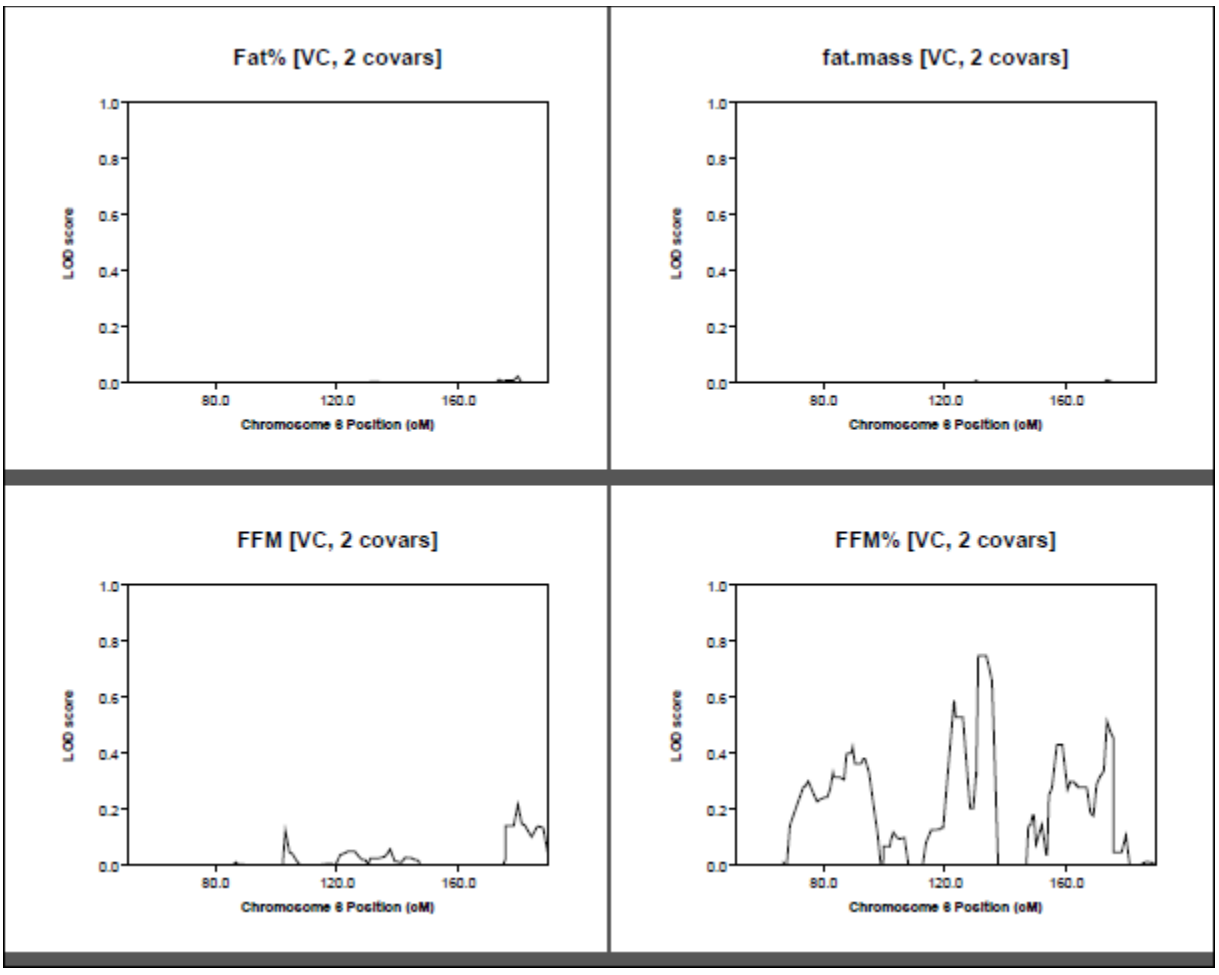
Chromosome 5

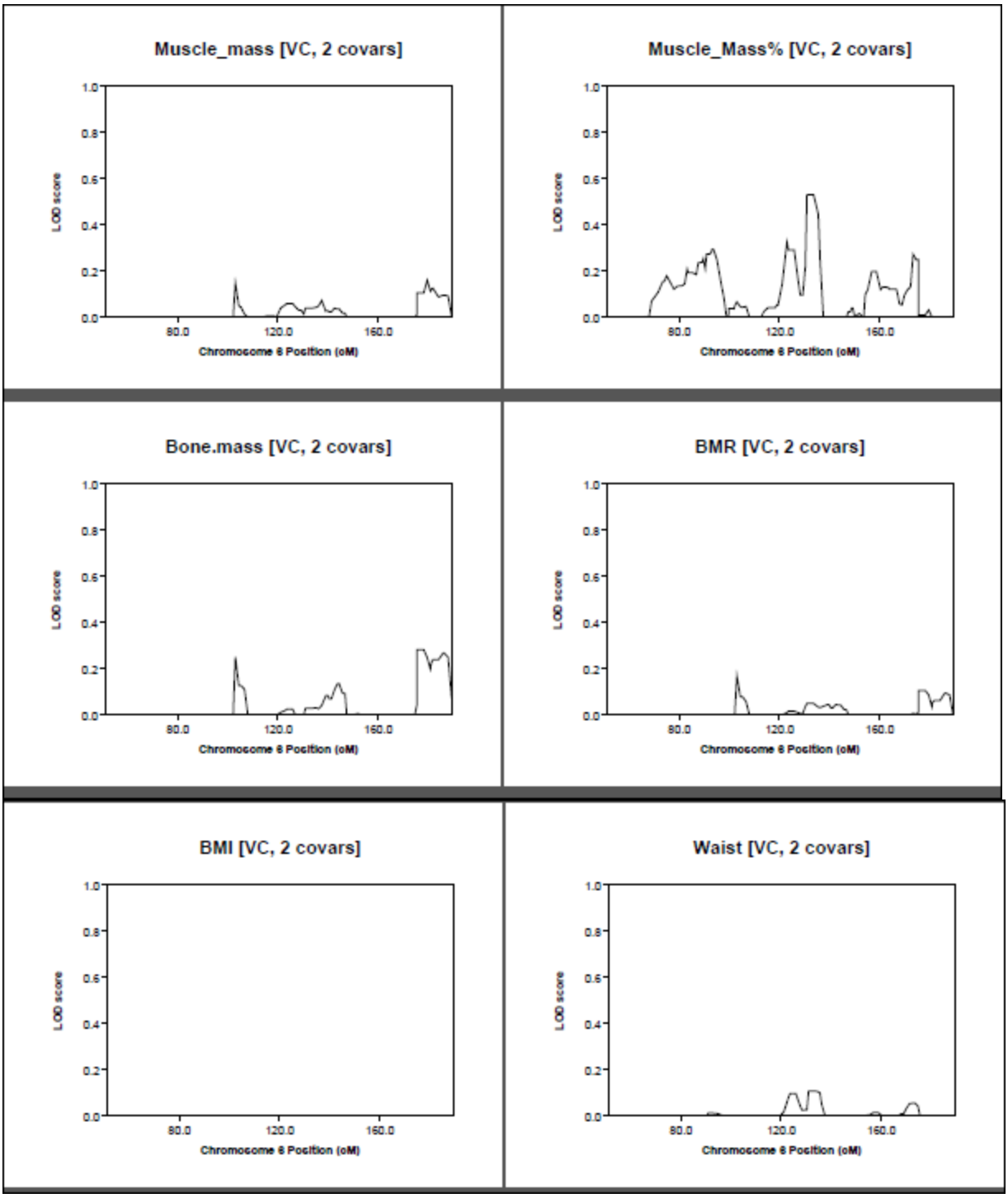



FFM [VC, 2 covars]
LOD score
Chromosome 6 Position (cM)
FFM% [VC, 2 covars]
LOD score
Chromosome 6 Position (cM)
Muscle_mass [VC, 2 covars]
LOD score
Chromosome 6 Position (cM)
Muscle_Mass% [VC, 2 covars]
LOD score
Chromosome 6 Position (cM)
Bone.mass [VC, 2 covars]
LOD score
Chromosome 6 Position (cM)
BMR [VC, 2 covars]
LOD score
Chromosome 6 Position (cM)

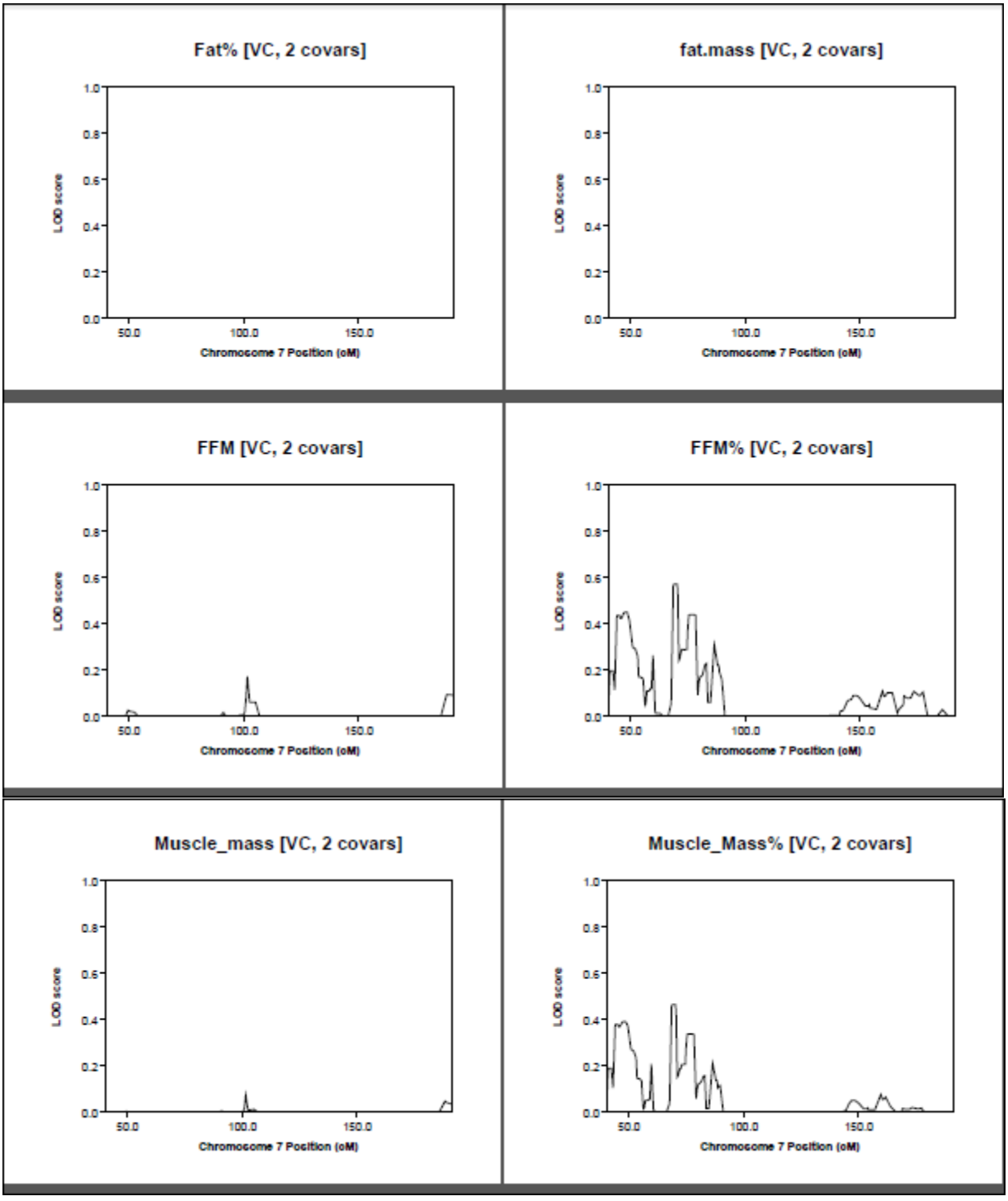


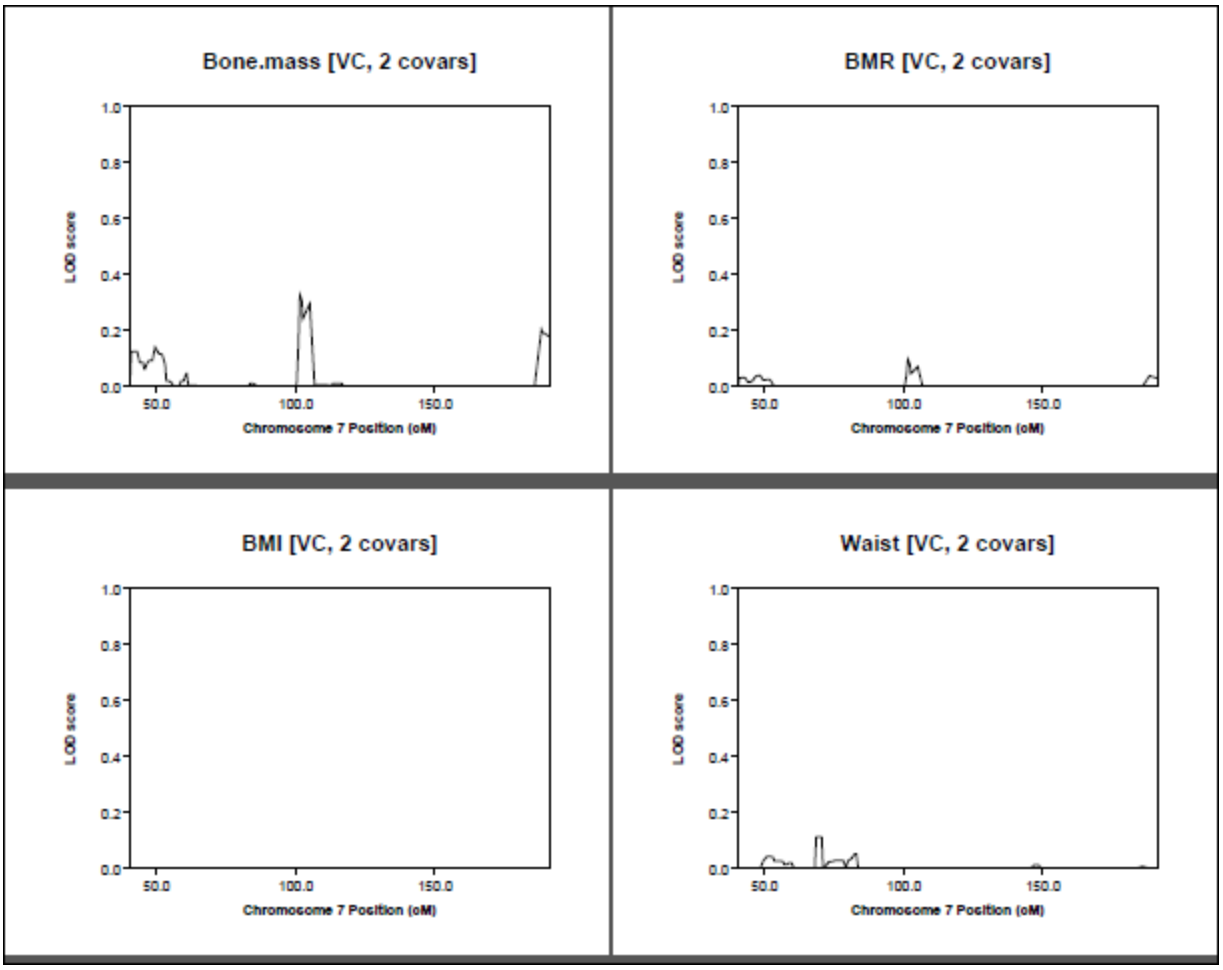
Chromosome 6



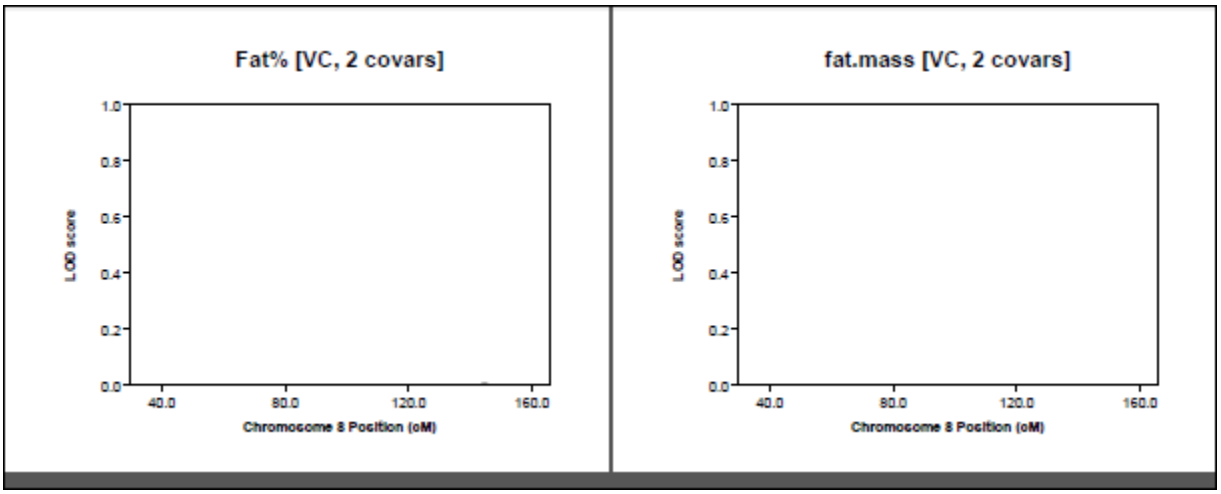

Muscle_mass [VC, 2 covars]
Muscle_Mass% [VC, 2 covars]
Bone.mass [VC, 2 covars]
BMR [VC, 2 covars]
BMI [VC, 2 covars]
Waist [VC, 2 covars]
LOD score
0.0
0.2
0.4
0.6
0.8
1.0
80.0
120.0
160.0
Chromosome 6 Position (cM)

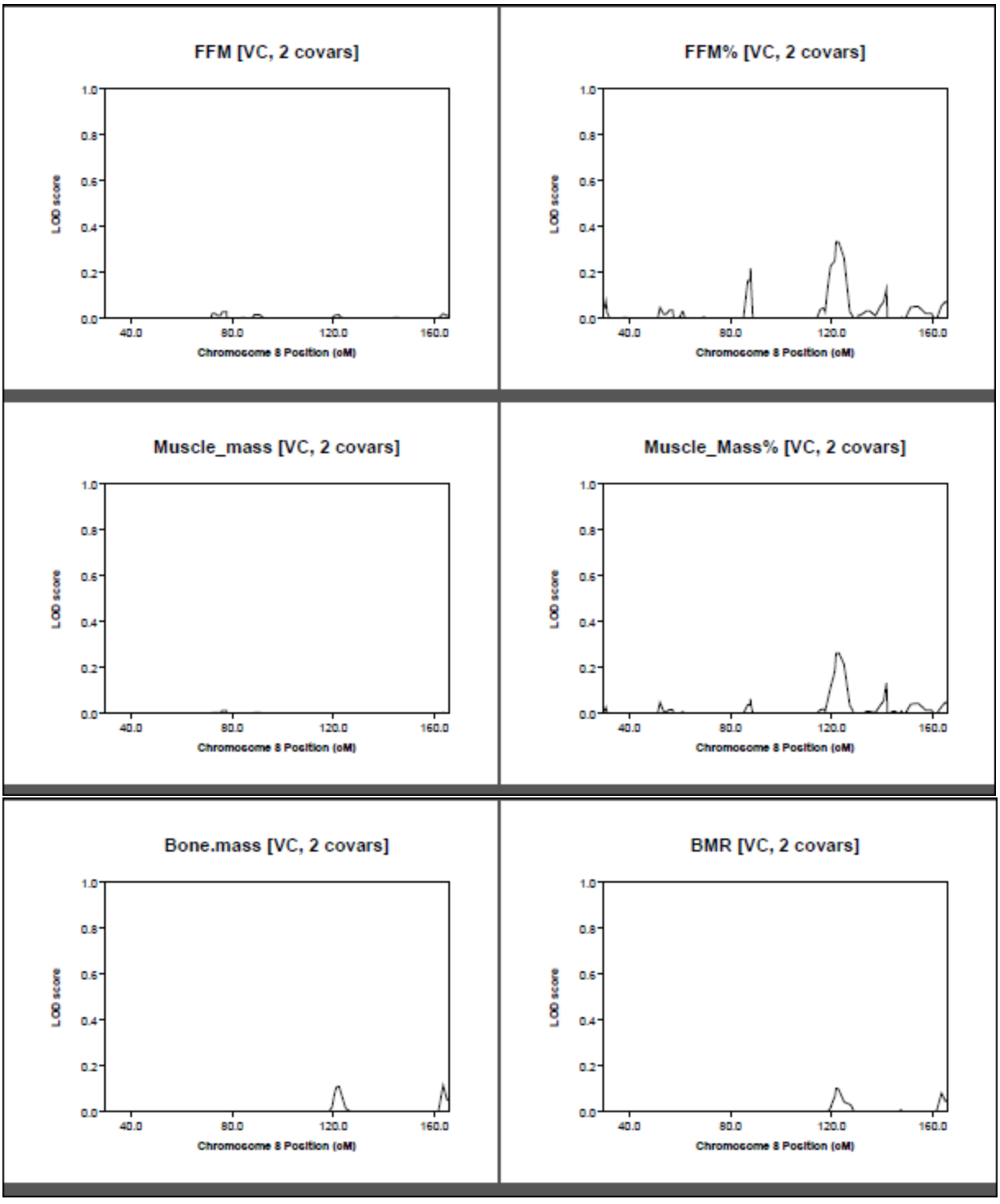
Chromosome 7

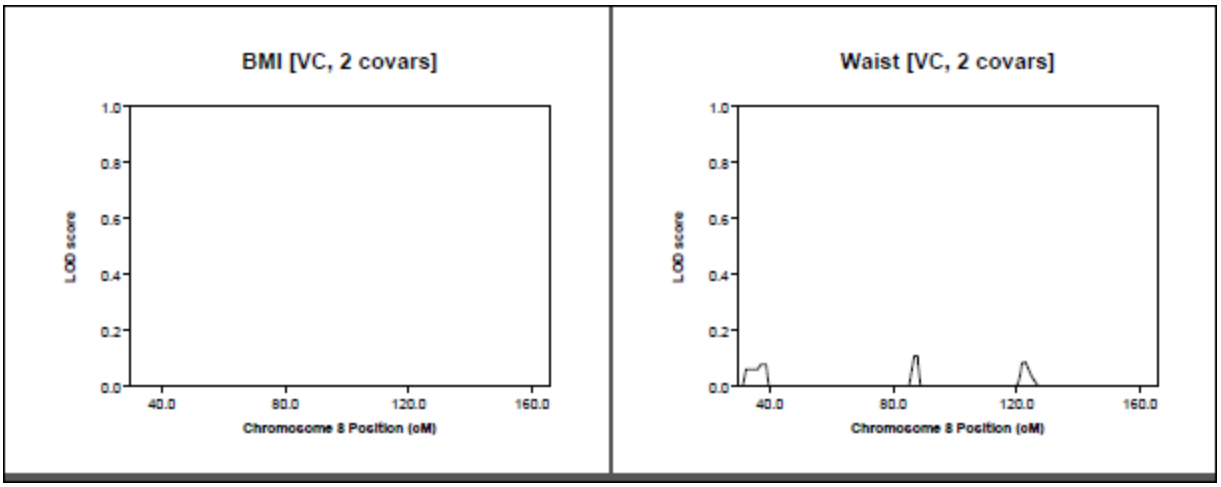




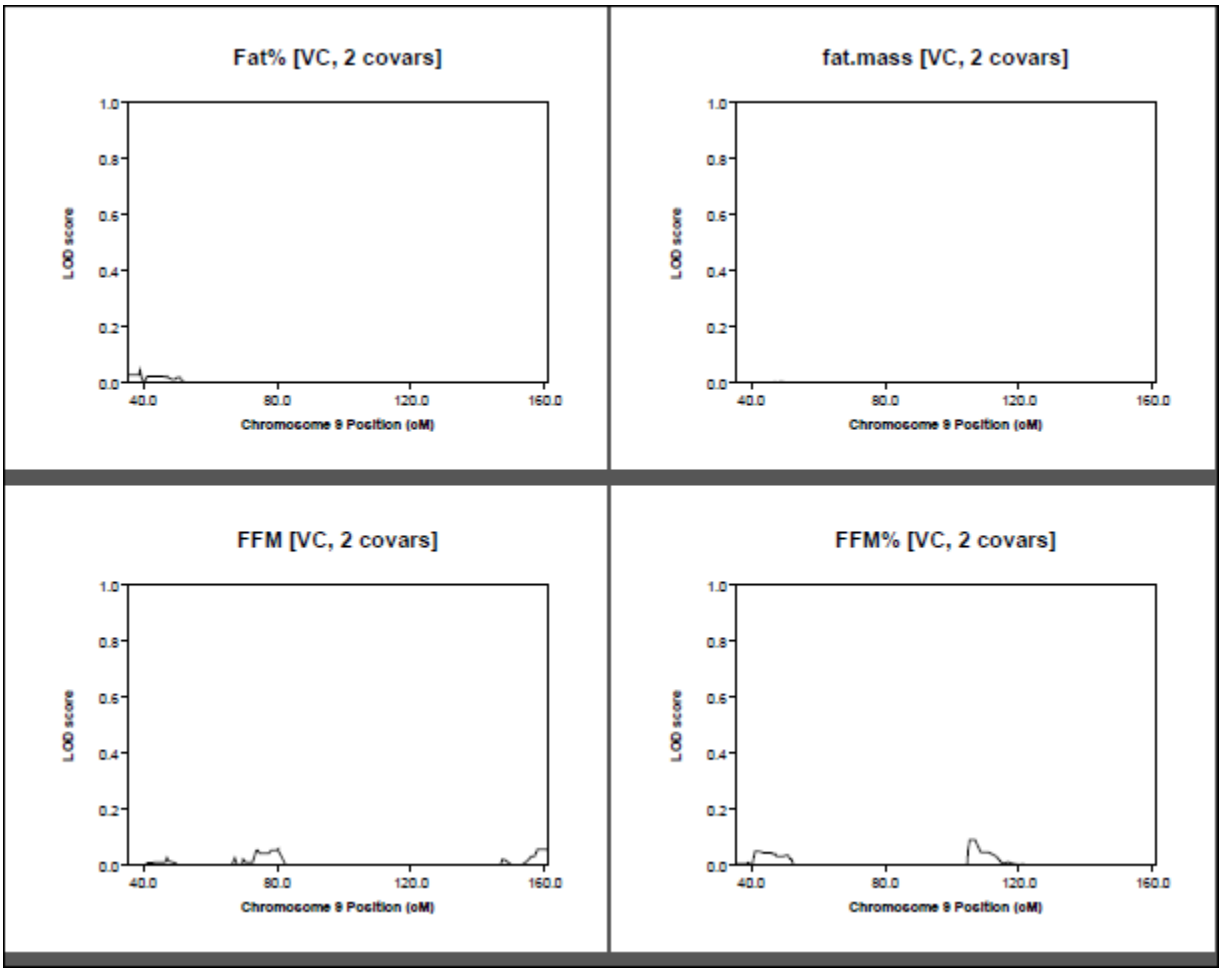
Chromosome 8

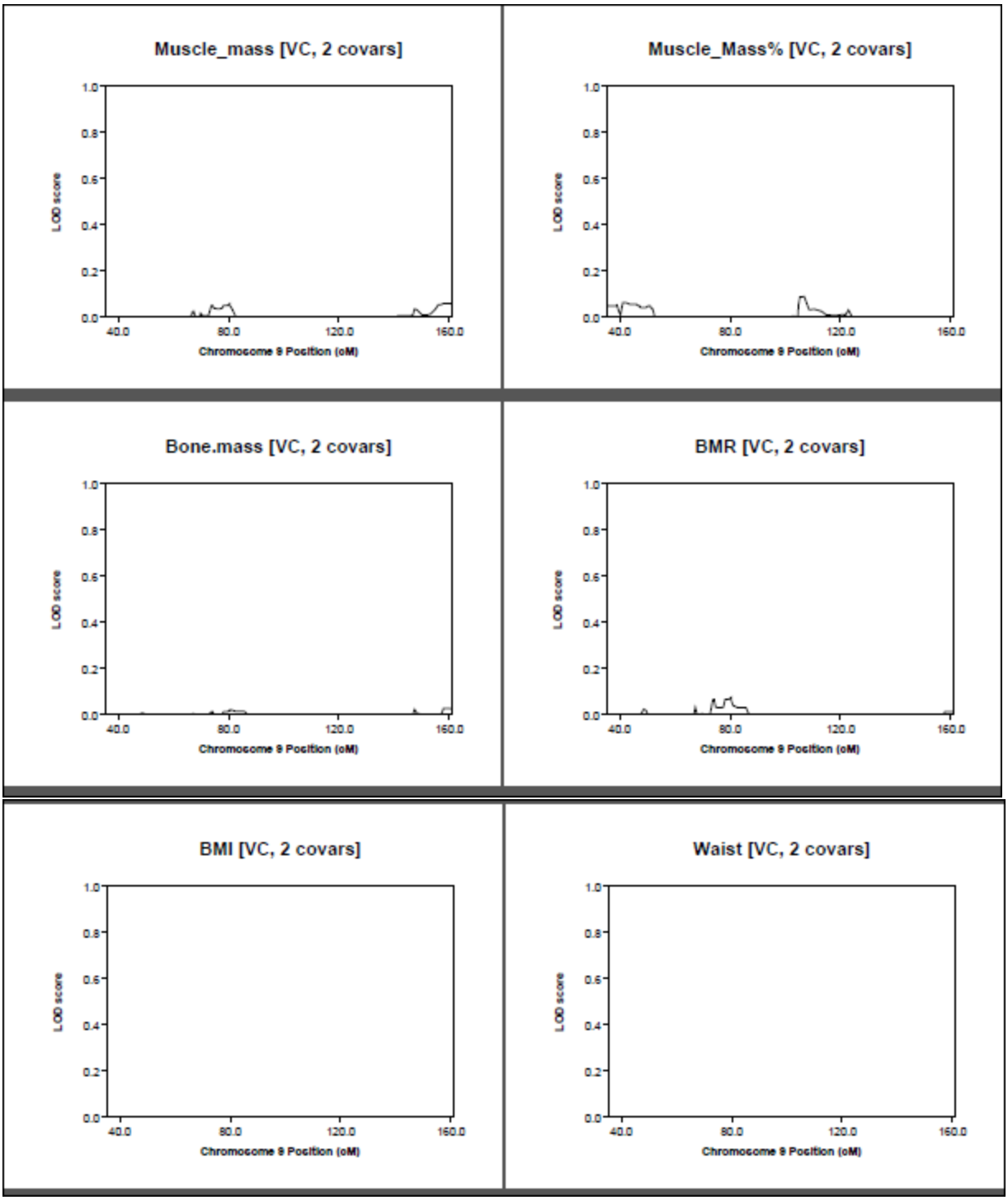



FFM [VC, 2 covars]
FFM% [VC, 2 covars]
Muscle_mass [VC, 2 covars]
Muscle_Mass% [VC, 2 covars]
Bone.mass [VC, 2 covars]
BMR [VC, 2 covars]
LOD score
Chromosome 8 Position (cM)
0.0
0.2
0.4
0.6
0.8
1.0
40.0
80.0
120.0
160.0

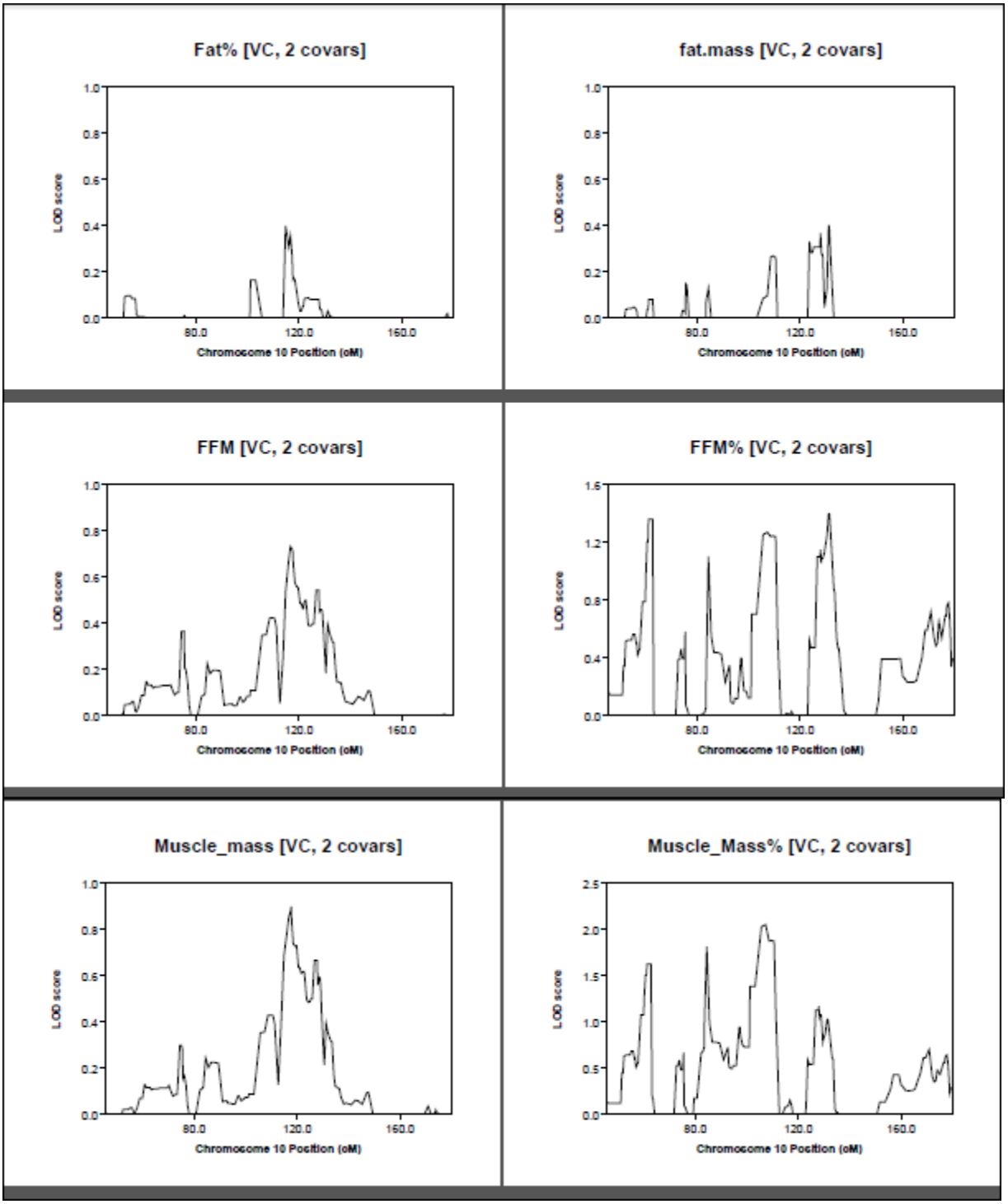


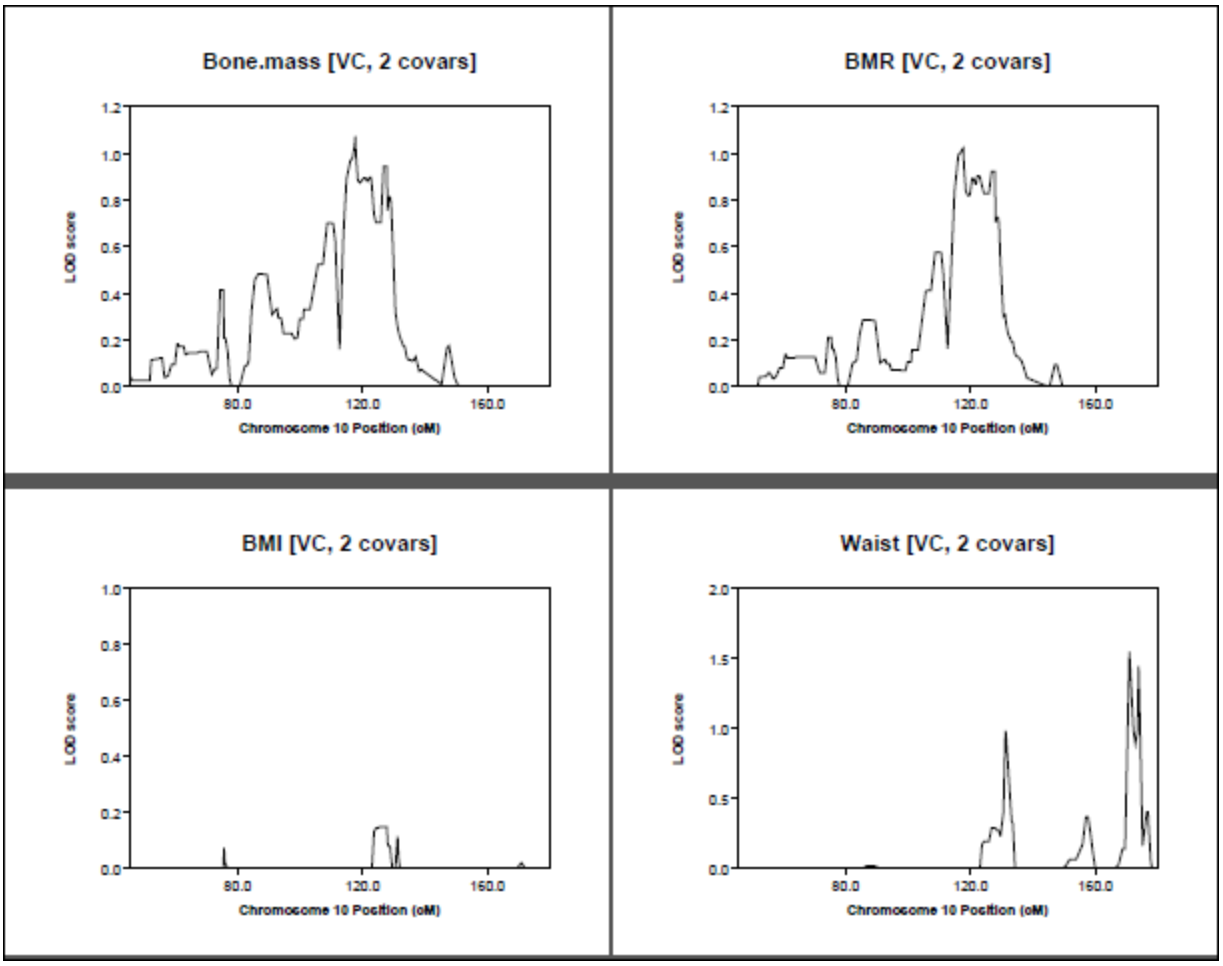
Chromosome 9



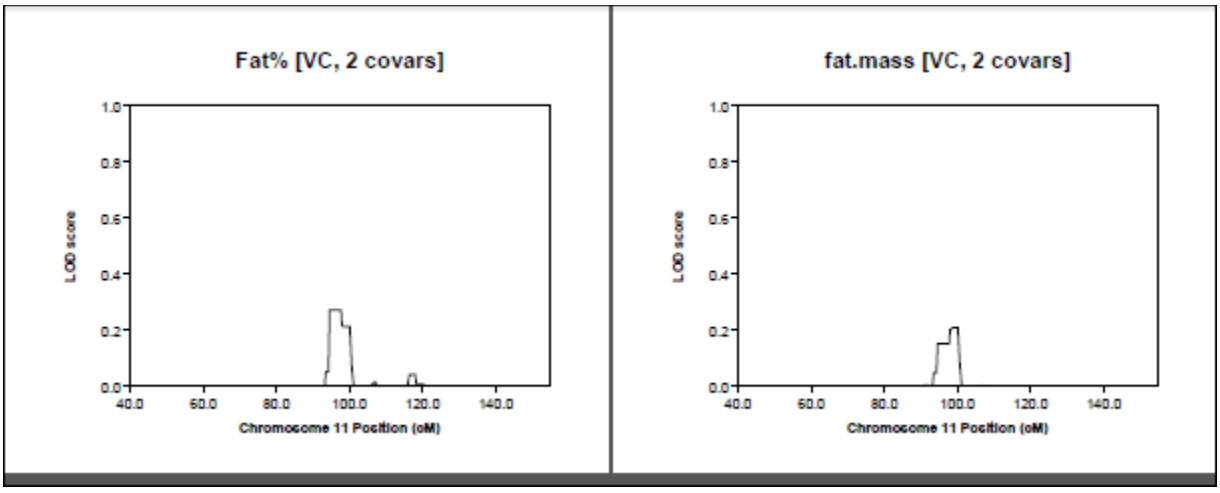

Muscle_mass [VC, 2 covars]
Muscle_Mass% [VC, 2 covars]
Bone.mass [VC, 2 covars]
BMR [VC, 2 covars]
BMI [VC, 2 covars]
Waist [VC, 2 covars]
LOD score
Chromosome 9 Position (cM)
0.0
0.2
0.4
0.6
0.8
1.0
40.0
80.0
120.0
160.0

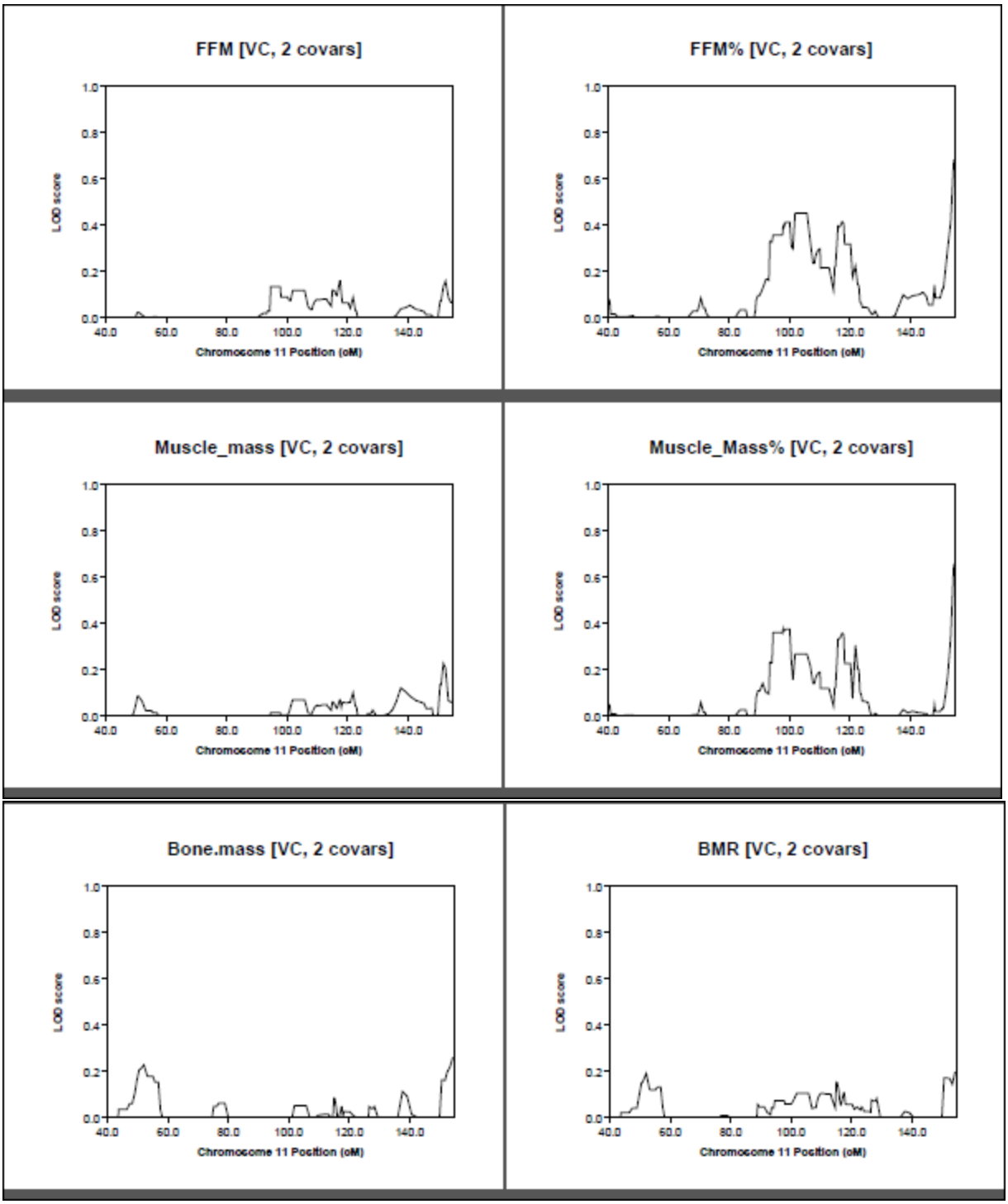
Chromosome 10

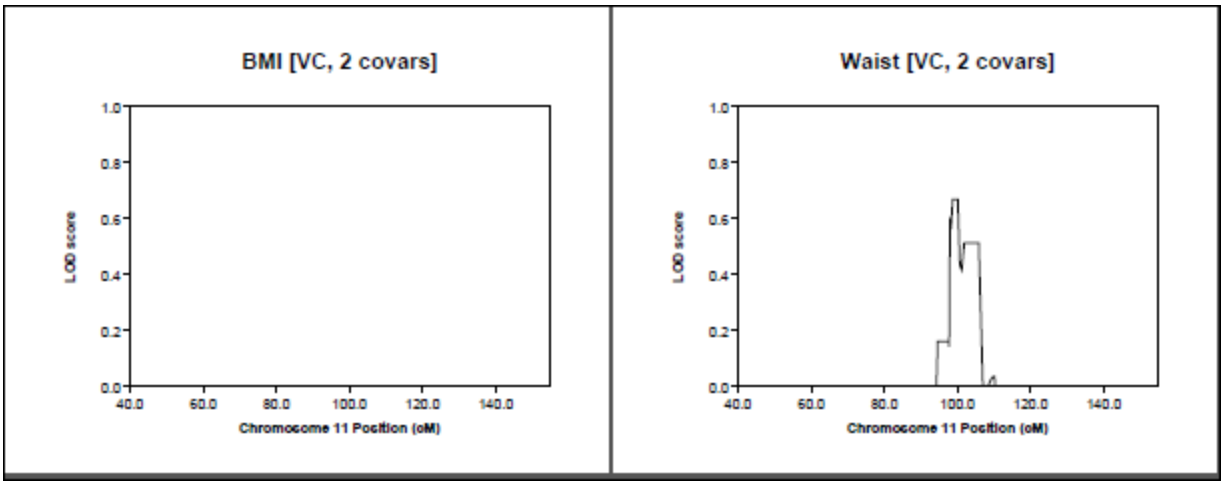




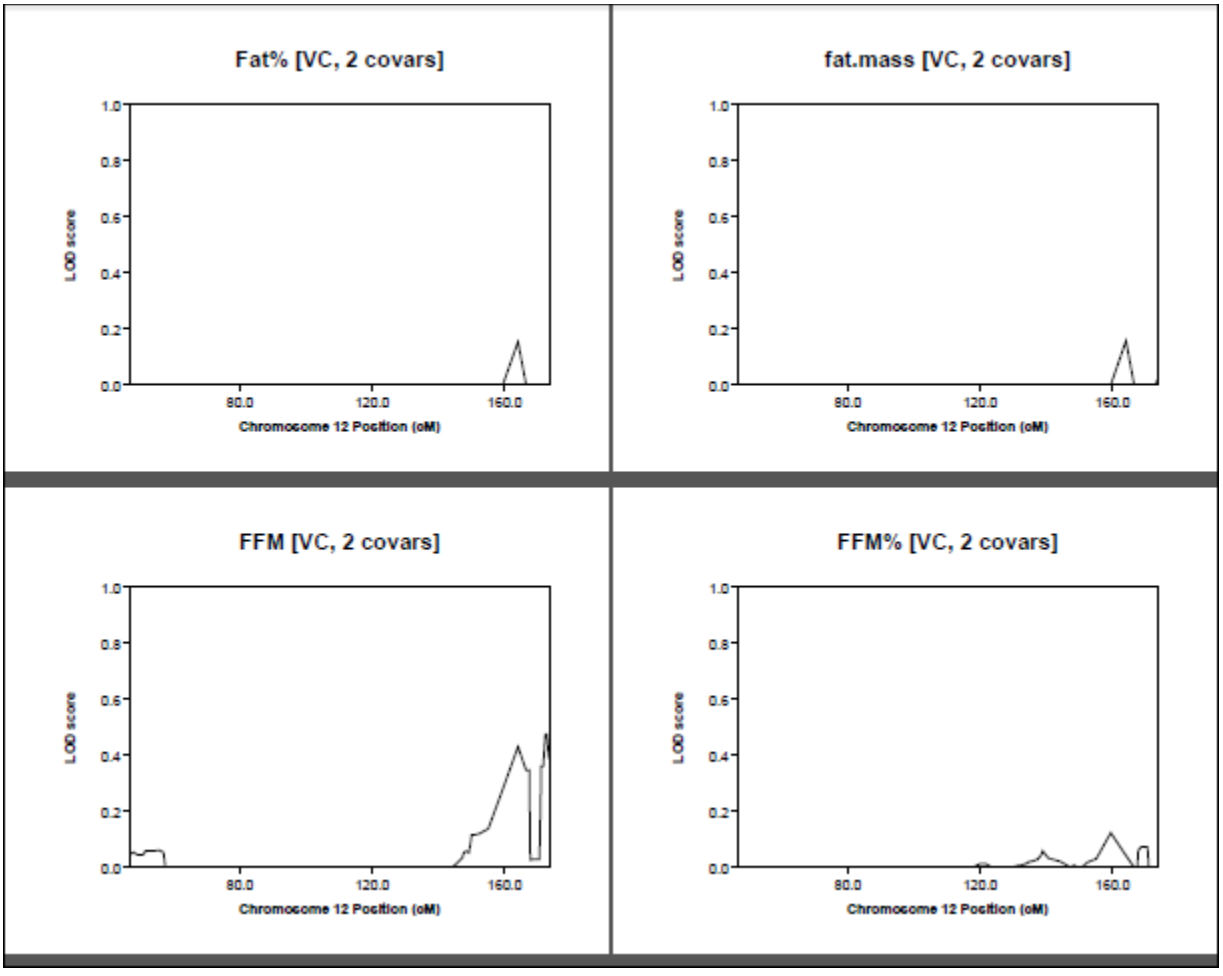
Chromosome 11

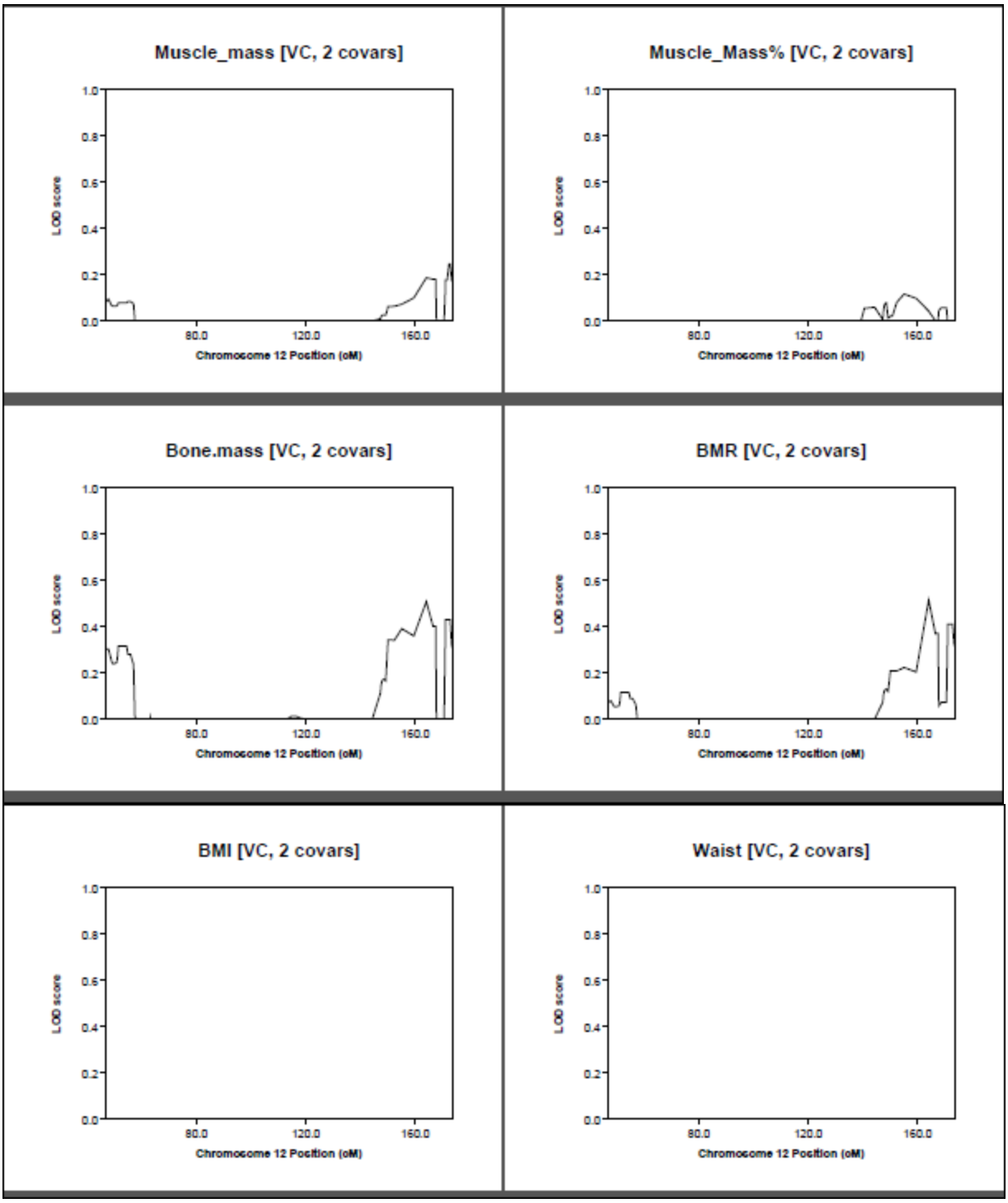



FFM [VC, 2 covars]
FFM% [VC, 2 covars]
Muscle_mass [VC, 2 covars]
Muscle_Mass% [VC, 2 covars]
Bone.mass [VC, 2 covars]
BMR [VC, 2 covars]
LOD score
Chromosome 11 Position (cM)
0.0
0.2
0.4
0.6
0.8
1.0
40.0
60.0
80.0
100.0
120.0
140.0

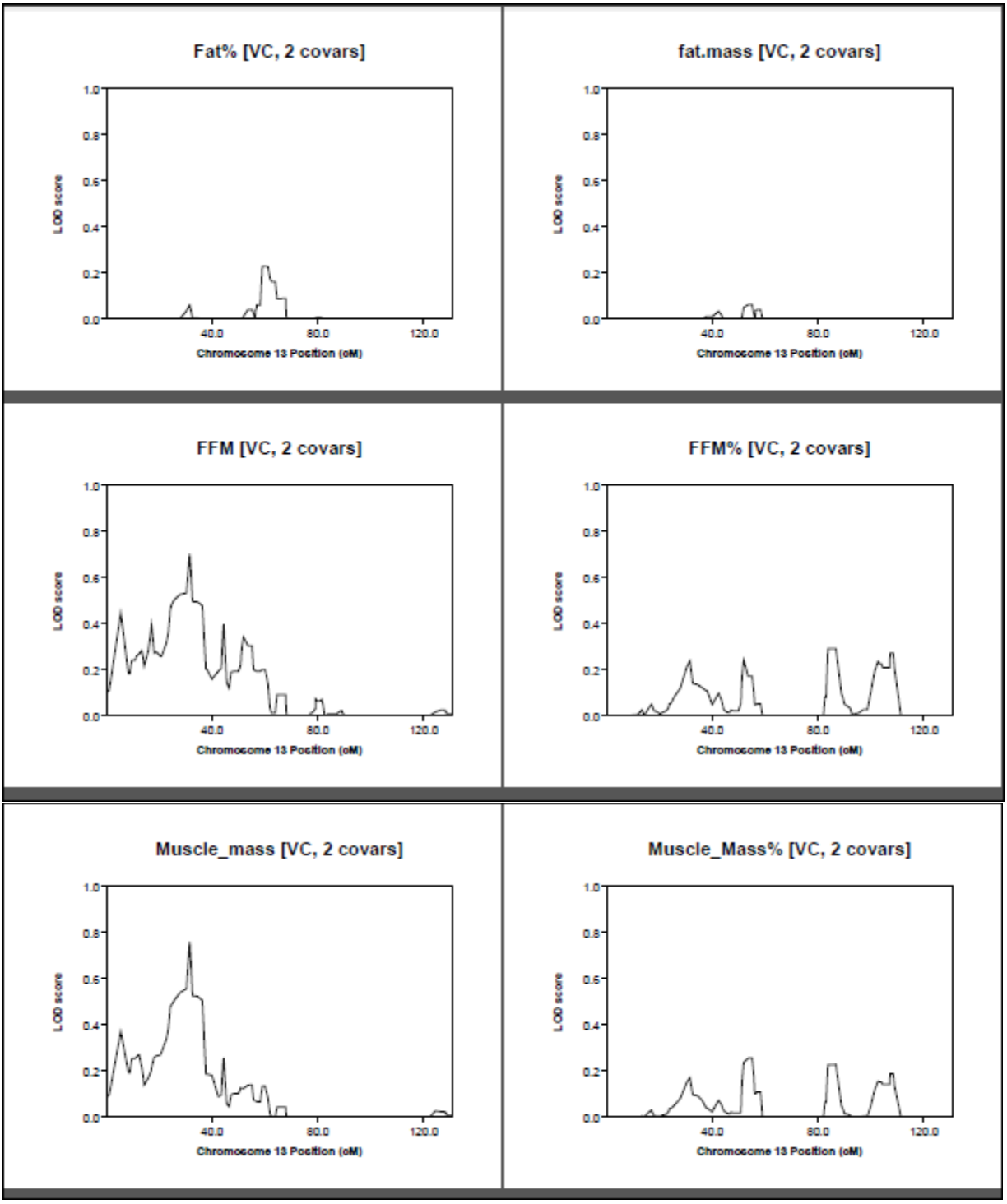


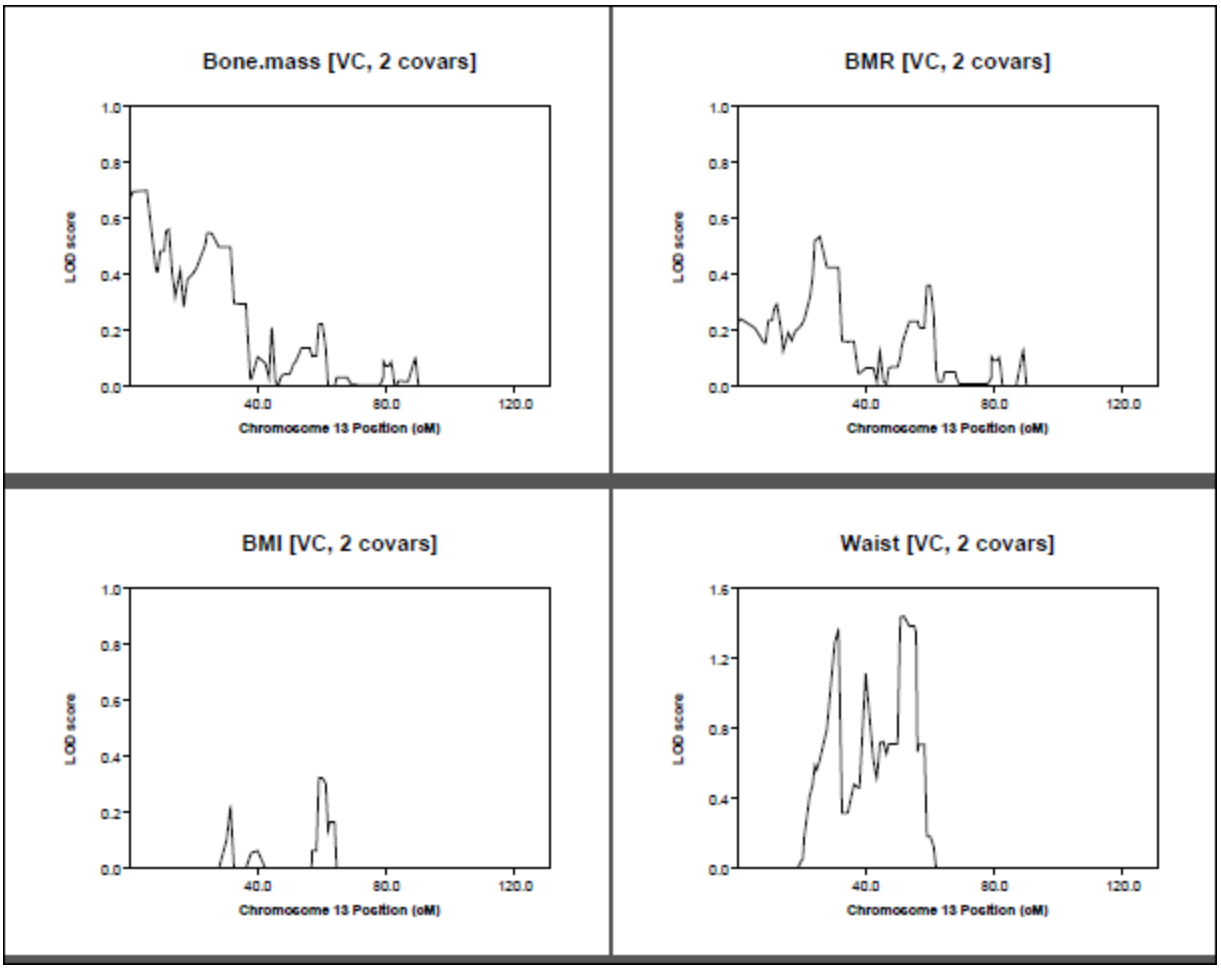
Chromosome 12



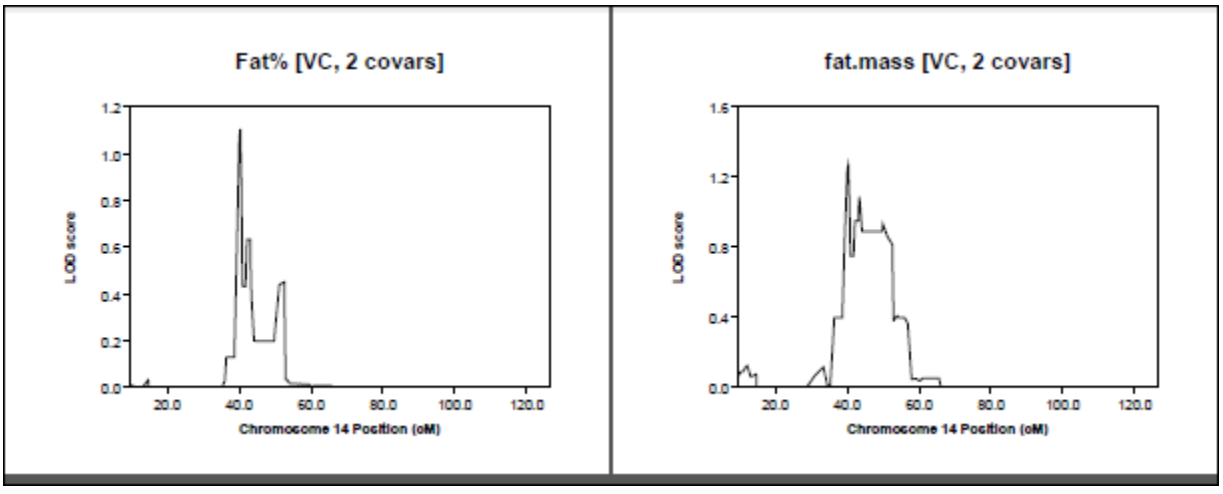

Muscle_mass [VC, 2 covars]
Muscle_Mass% [VC, 2 covars]
Bone.mass [VC, 2 covars]
BMR [VC, 2 covars]
BMI [VC, 2 covars]
Waist [VC, 2 covars]
LOD score
Chromosome 12 Position (cM)
0.0
0.2
0.4
0.6
0.8
1.0
80.0
120.0
160.0

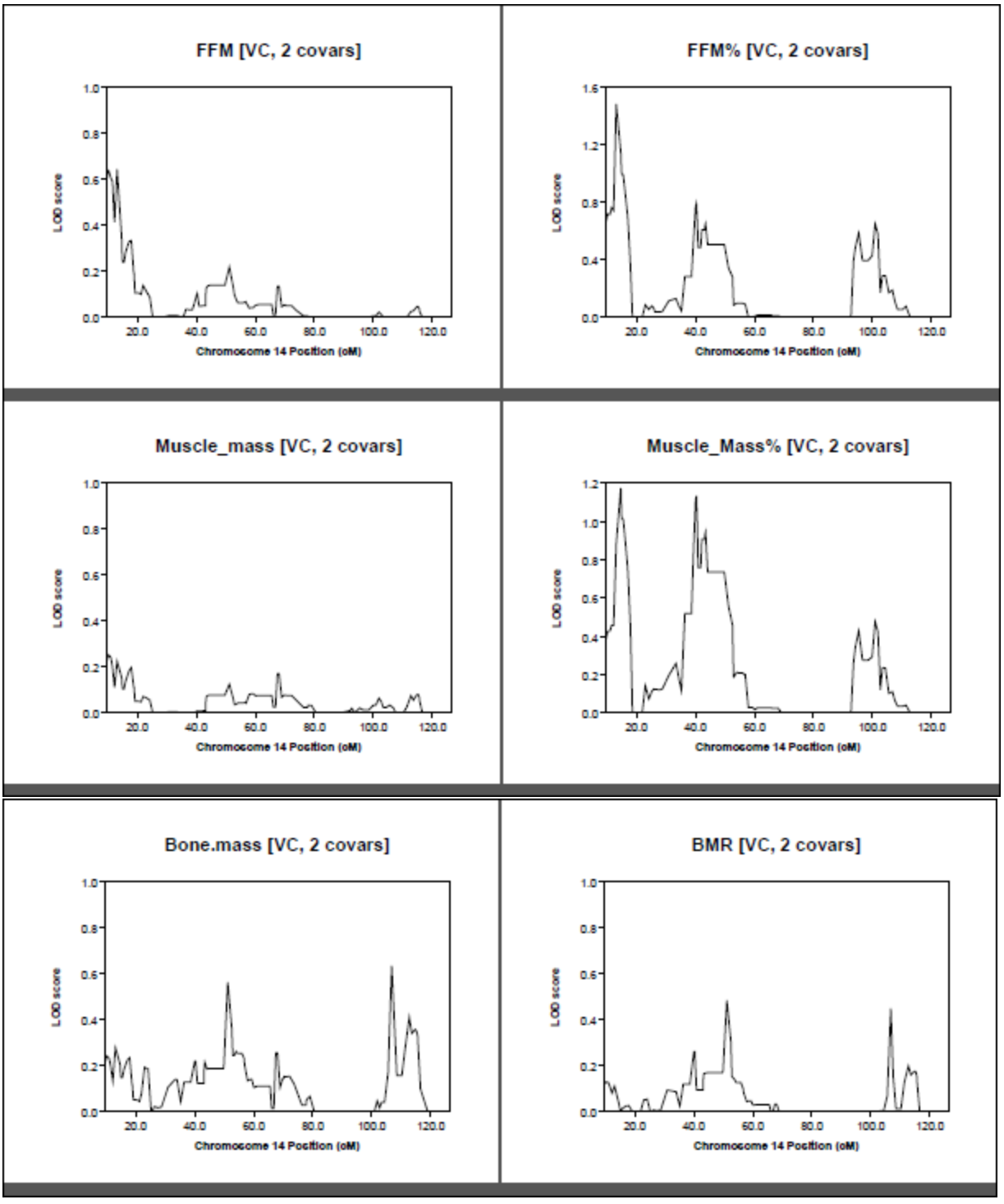
Chromosome 13

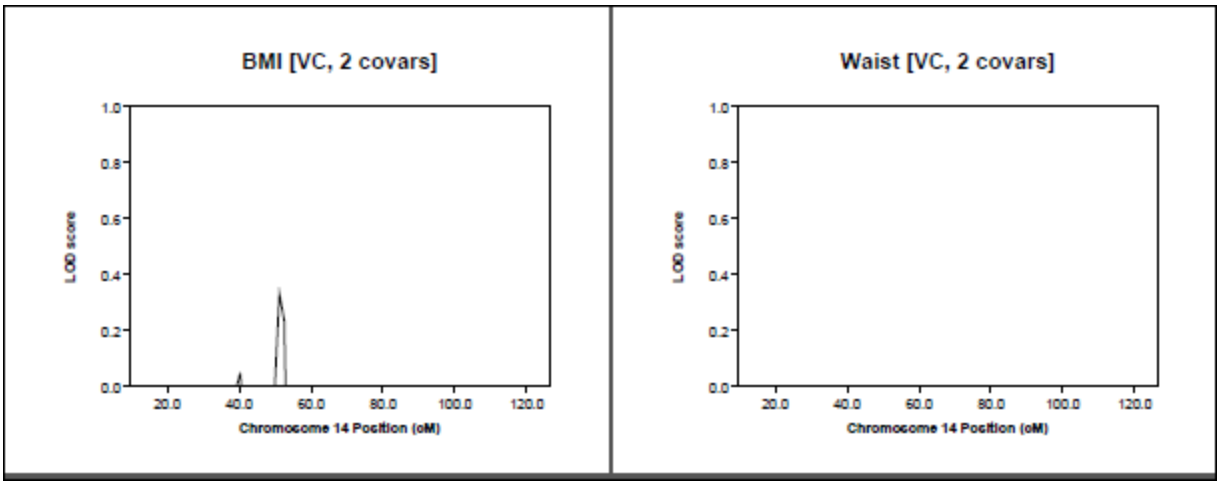




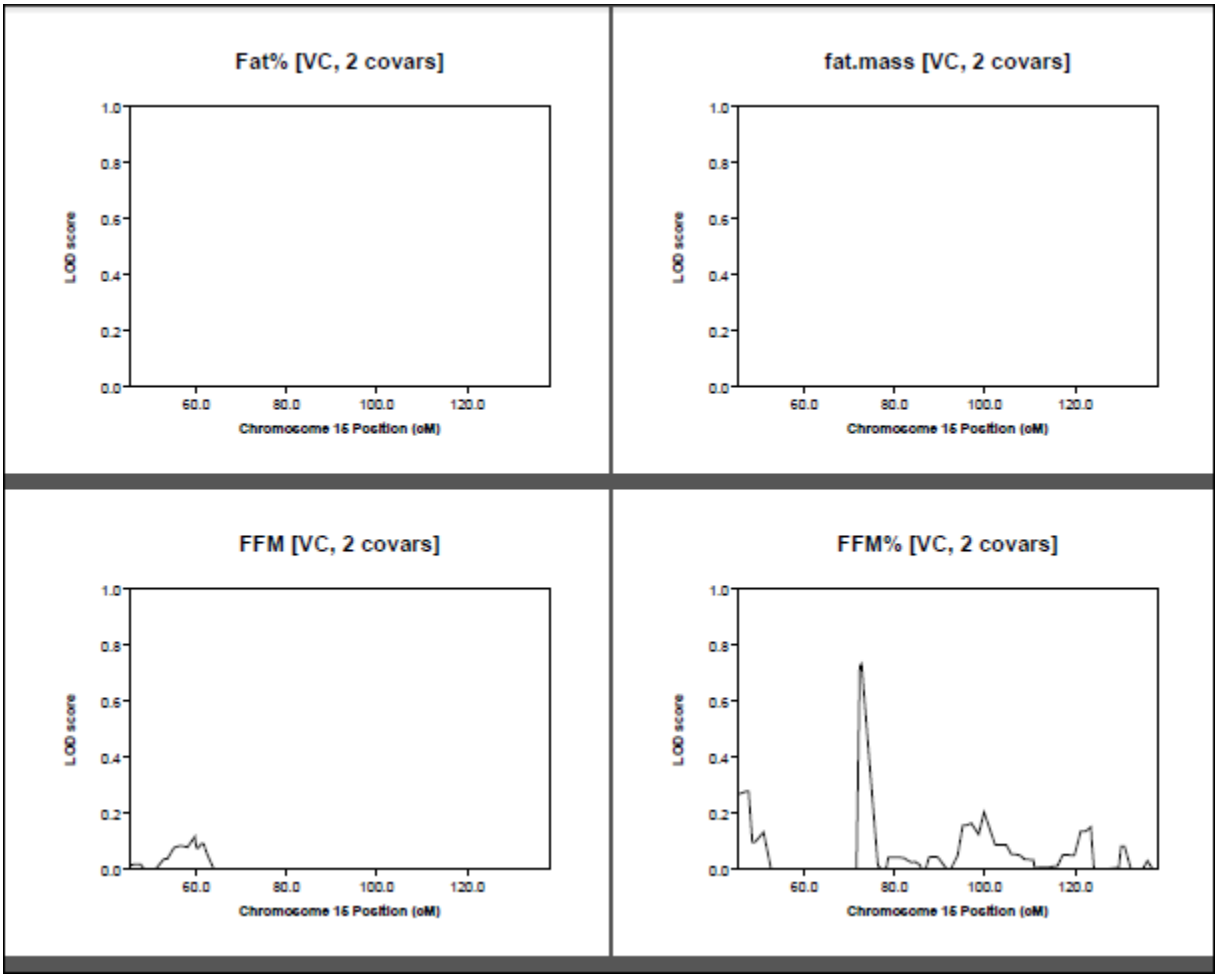
Chromosome 14

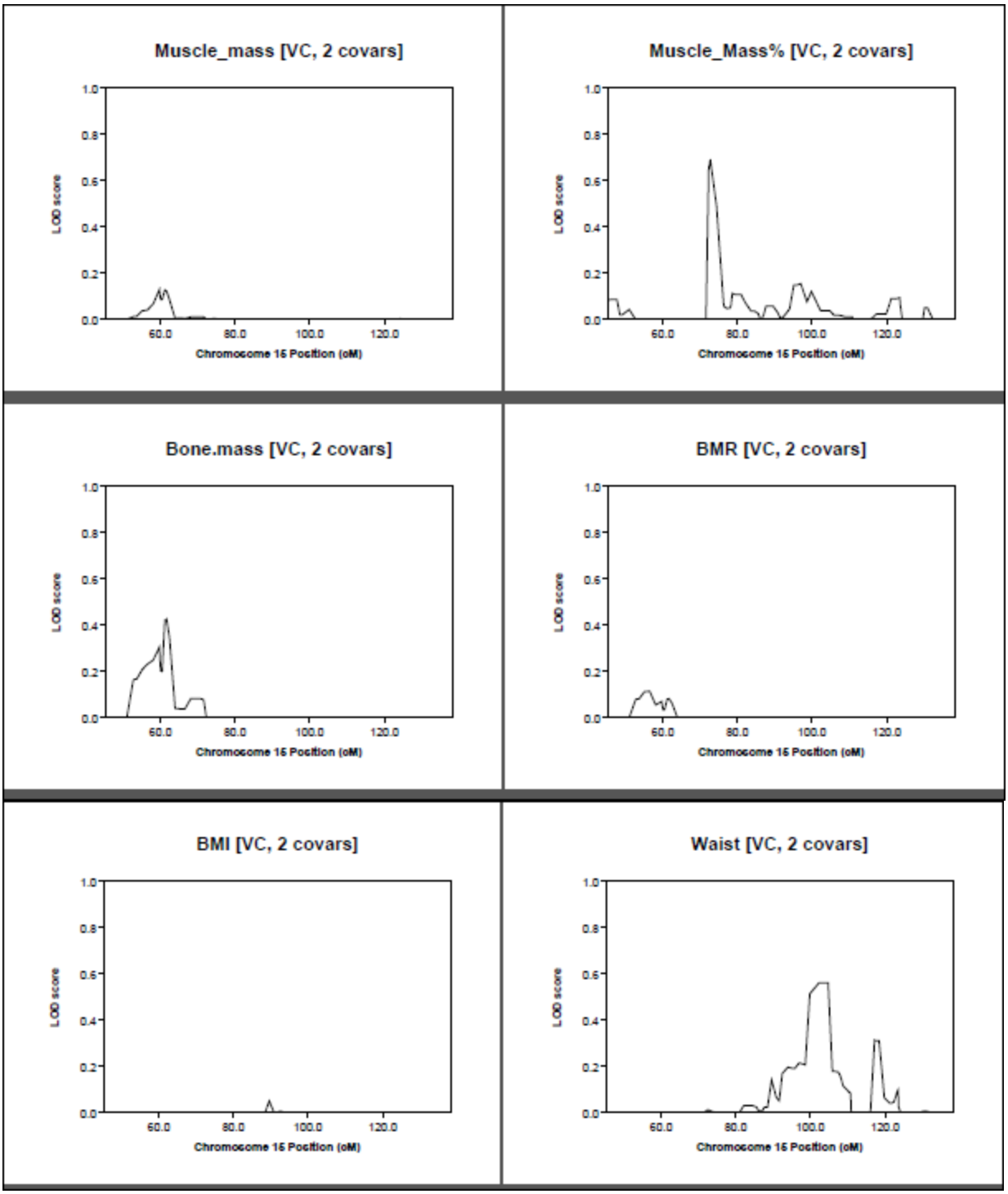



FFM [VC, 2 covars]
LOD score
Chromosome 14 Position (cM)
FFM% [VC, 2 covars]
LOD score
Chromosome 14 Position (cM)
Muscle_mass [VC, 2 covars]
LOD score
Chromosome 14 Position (cM)
Muscle_Mass% [VC, 2 covars]
LOD score
Chromosome 14 Position (cM)
Bone.mass [VC, 2 covars]
LOD score
Chromosome 14 Position (cM)
BMR [VC, 2 covars]
LOD score
Chromosome 14 Position (cM)

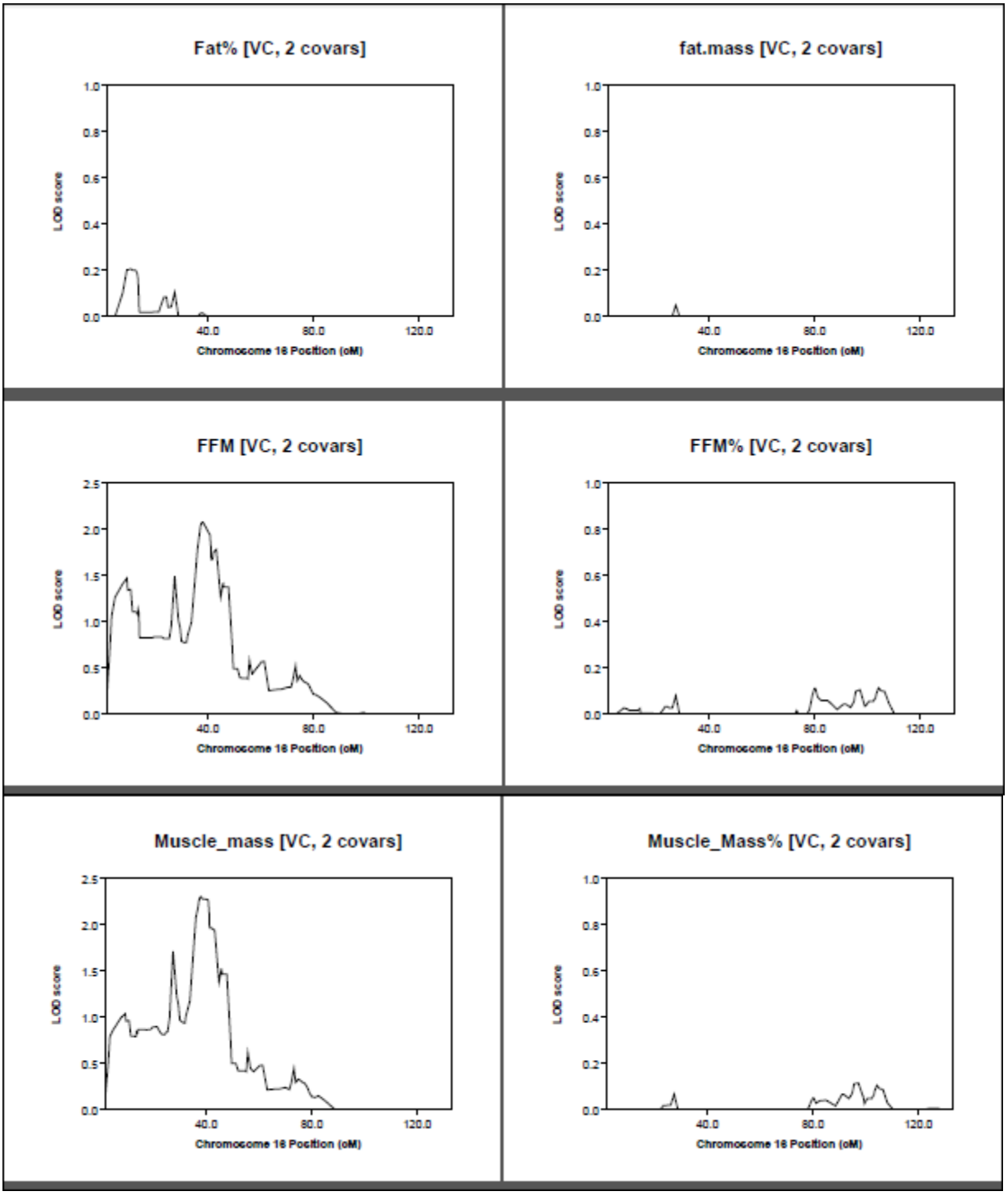


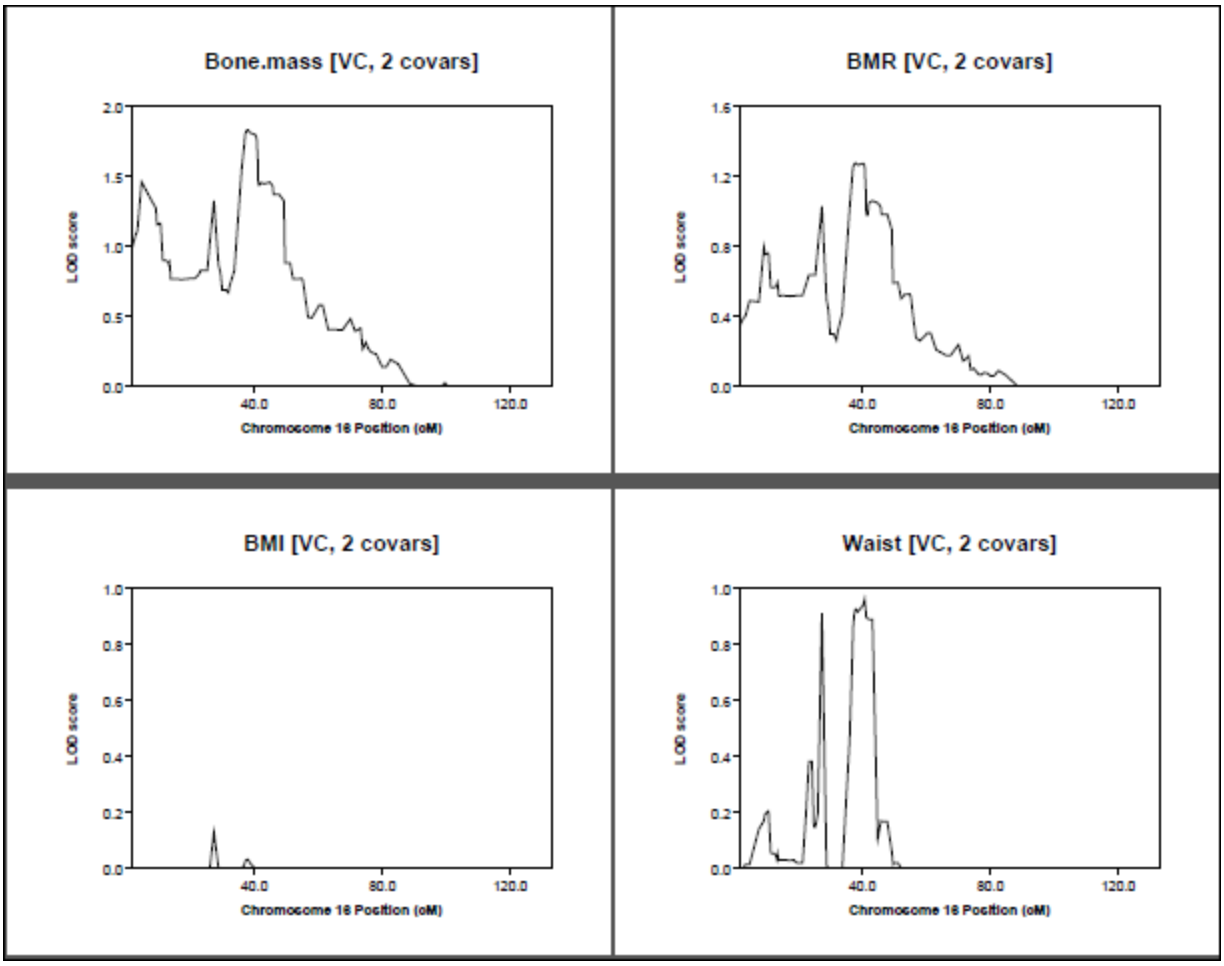
Chromosome 15



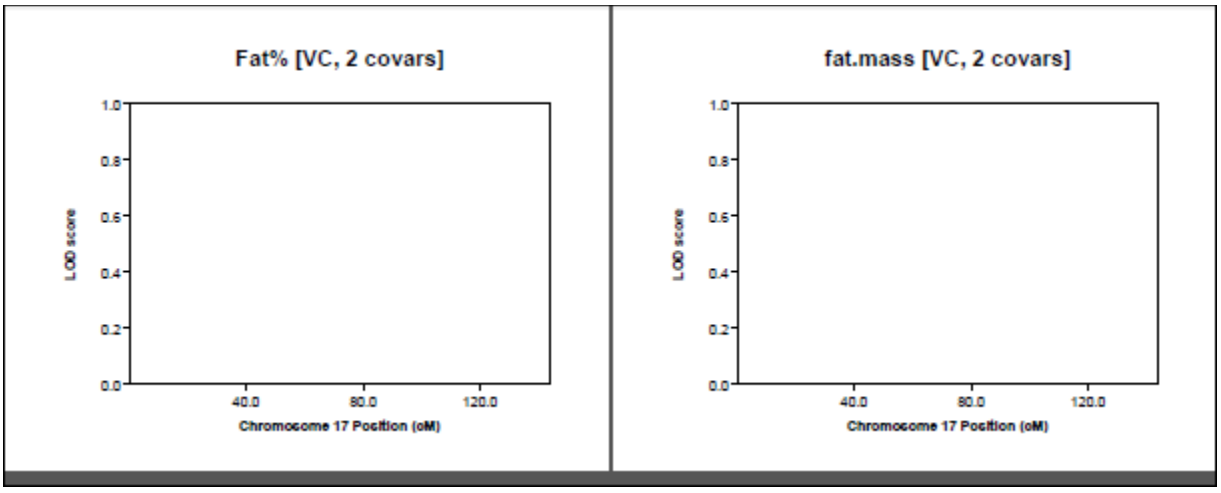

Muscle_mass [VC, 2 covars]
Muscle_Mass% [VC, 2 covars]
Bone.mass [VC, 2 covars]
BMR [VC, 2 covars]
BMI [VC, 2 covars]
Waist [VC, 2 covars]
LOD score
Chromosome 16 Position (cM)
0.0
0.2
0.4
0.6
0.8
1.0
60.0
80.0
100.0
120.0

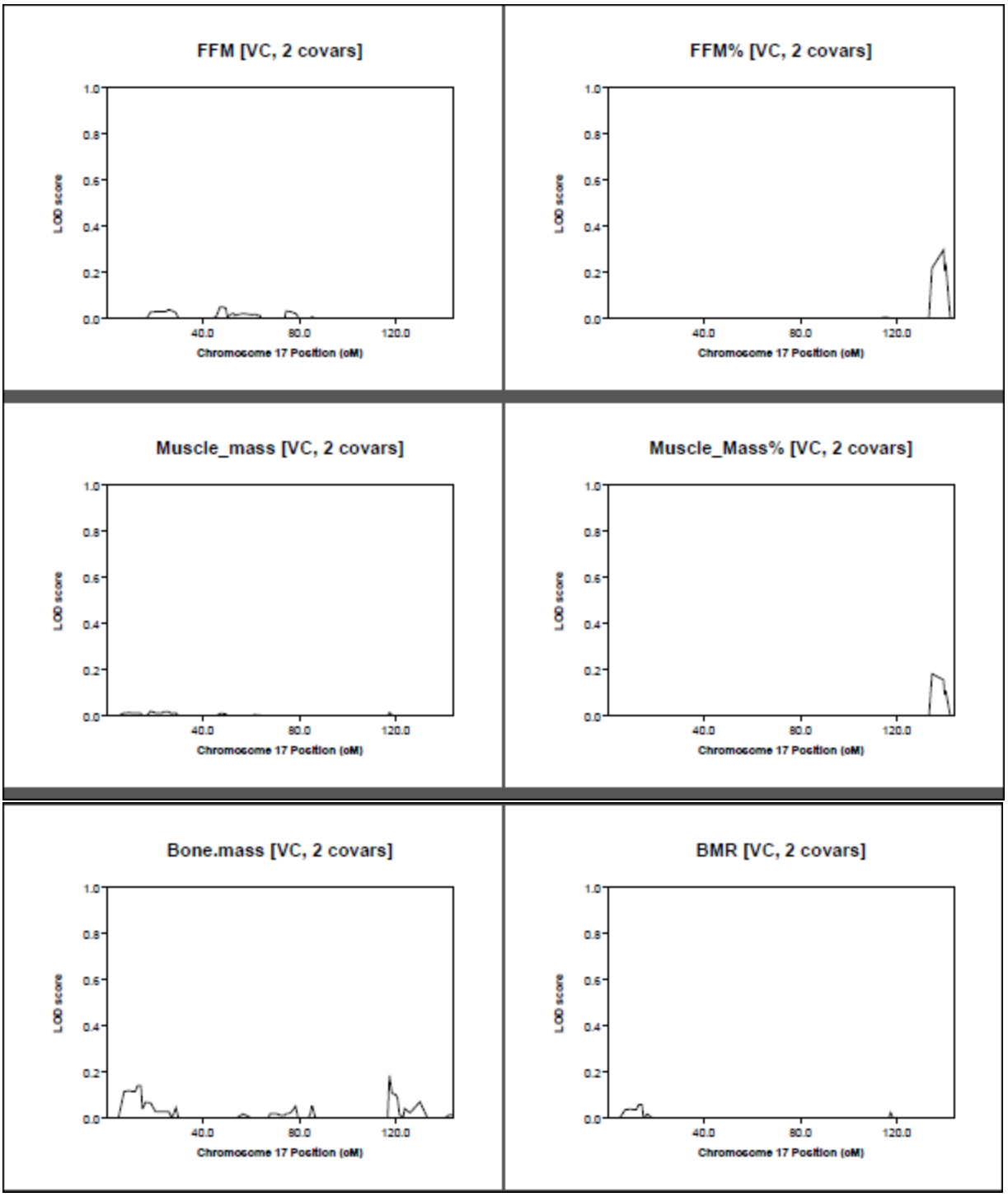
Chromosome 16

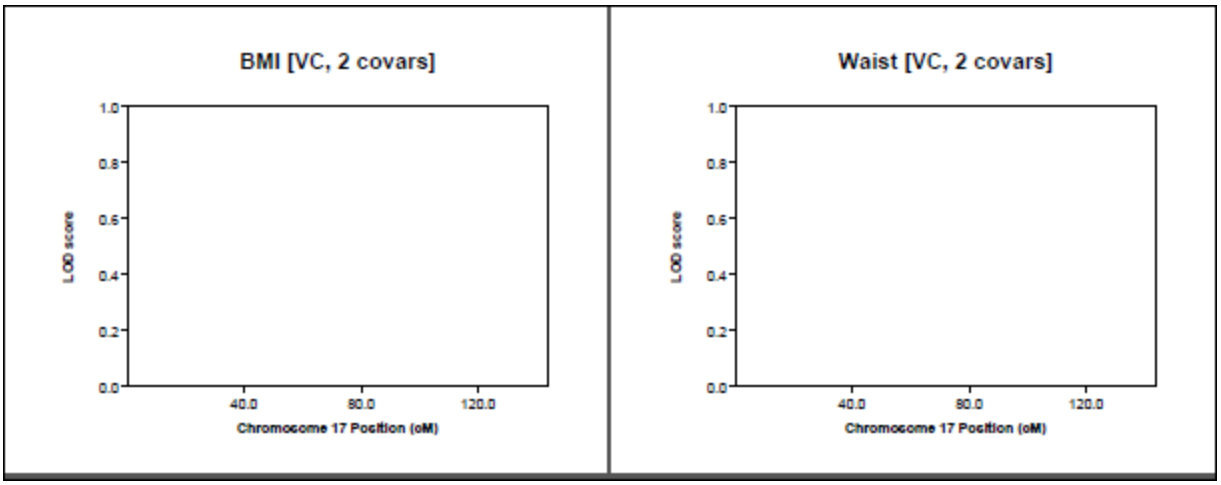




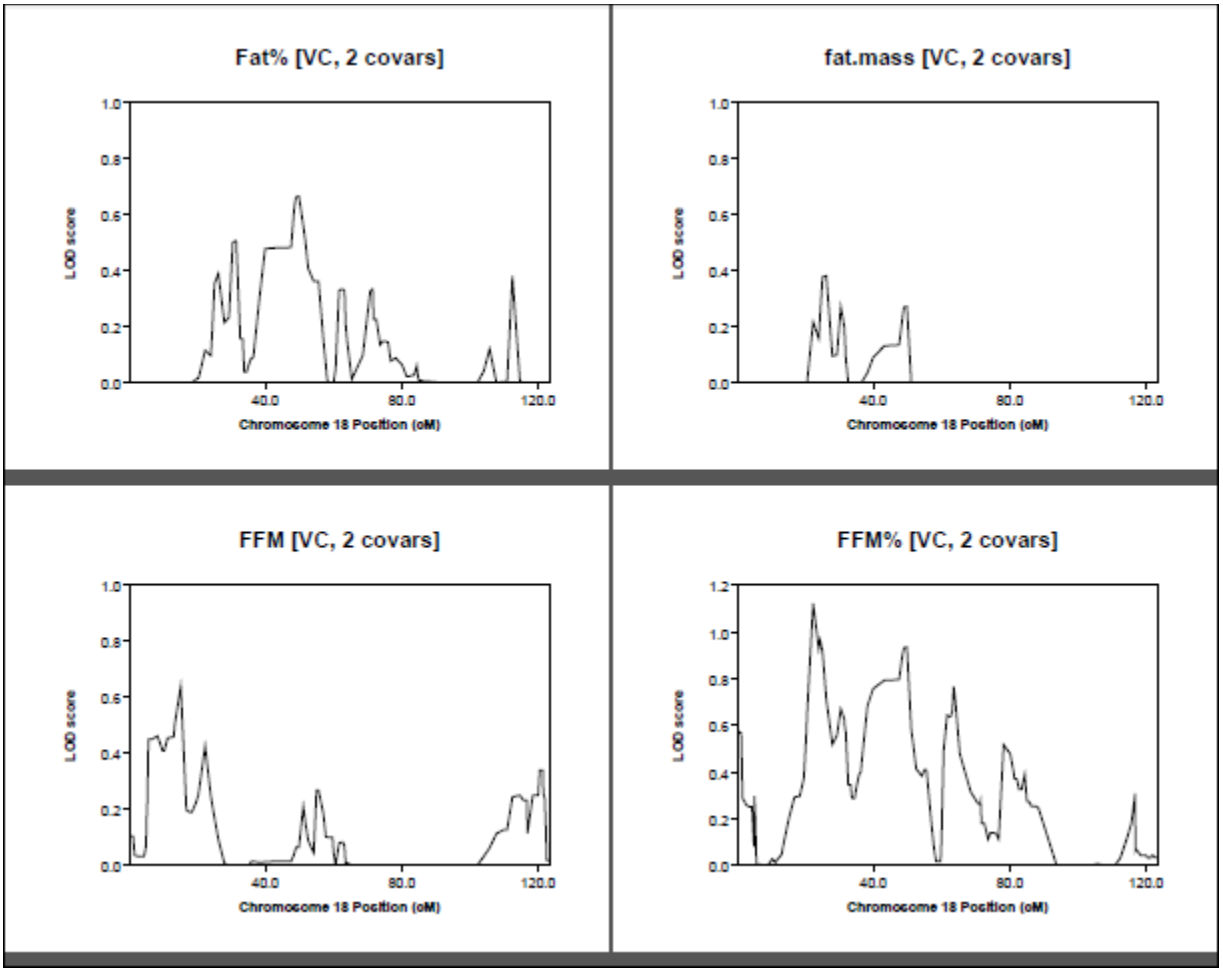
Chromosome 17

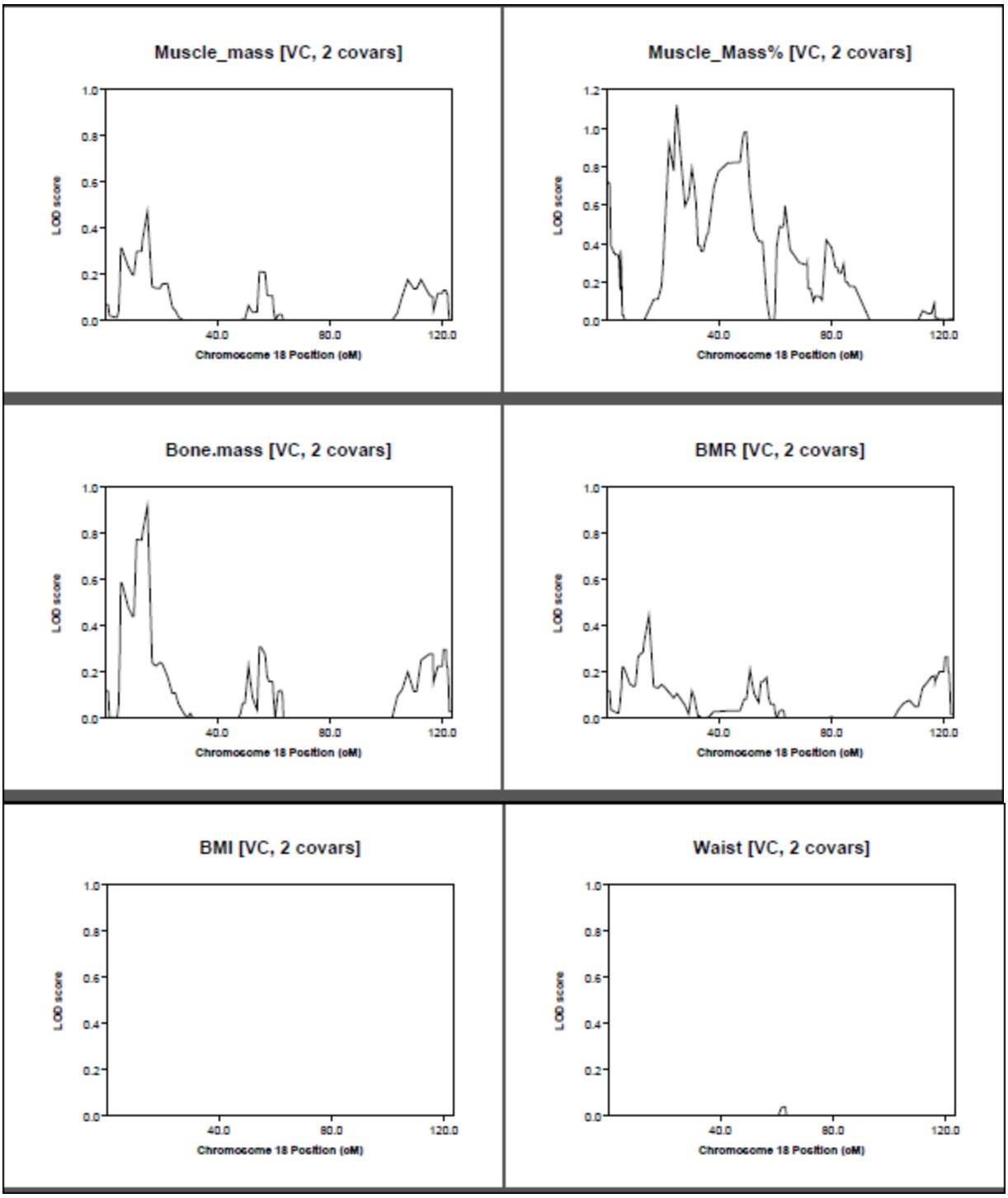



FFM [VC, 2 covars]
FFM% [VC, 2 covars]
Muscle_mass [VC, 2 covars]
Muscle_Mass% [VC, 2 covars]
Bone.mass [VC, 2 covars]
BMR [VC, 2 covars]
LOD score
0.0
0.2
0.4
0.6
0.8
1.0
40.0
80.0
120.0
Chromosome 17 Position (cM)

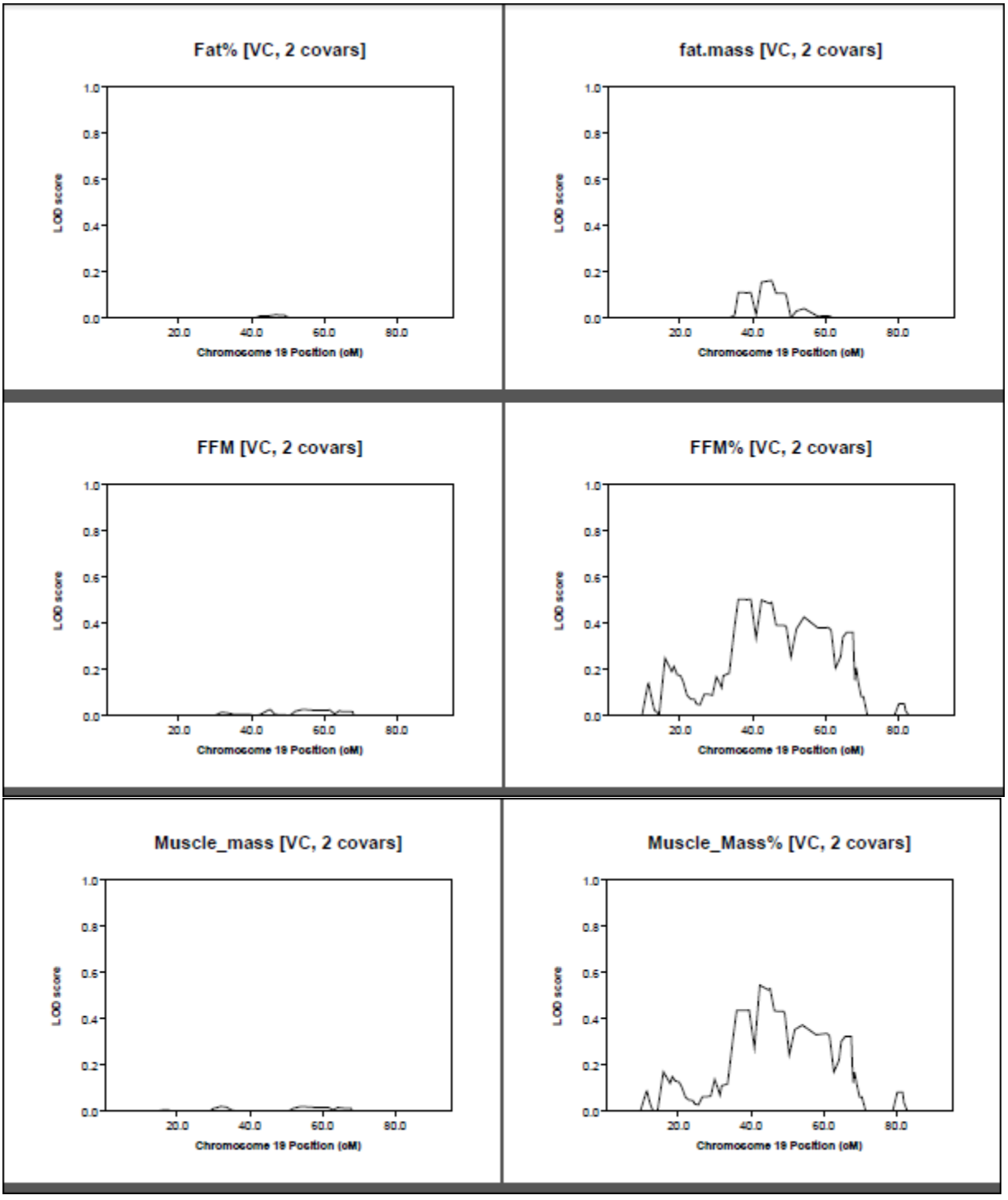


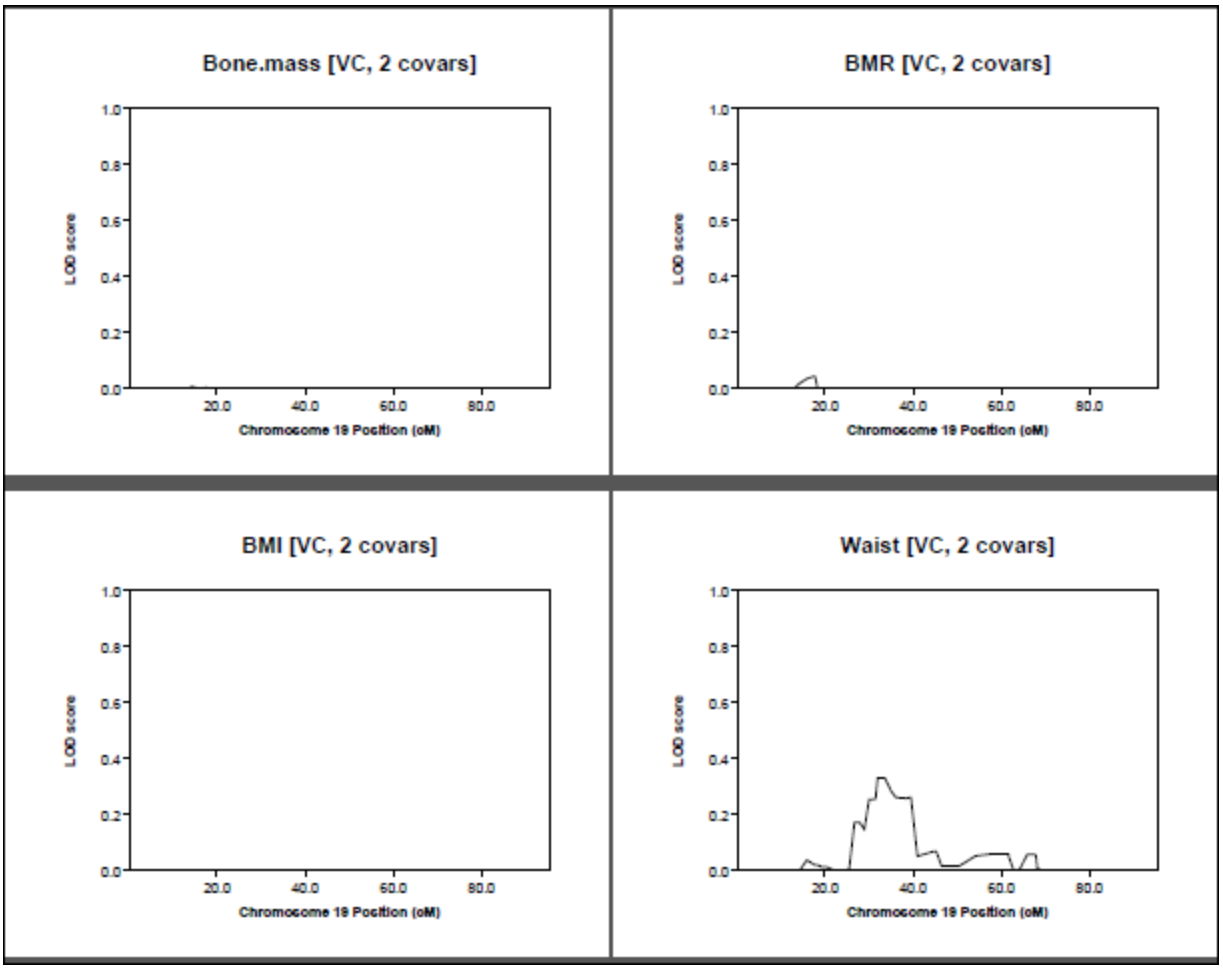
Chromosome 18



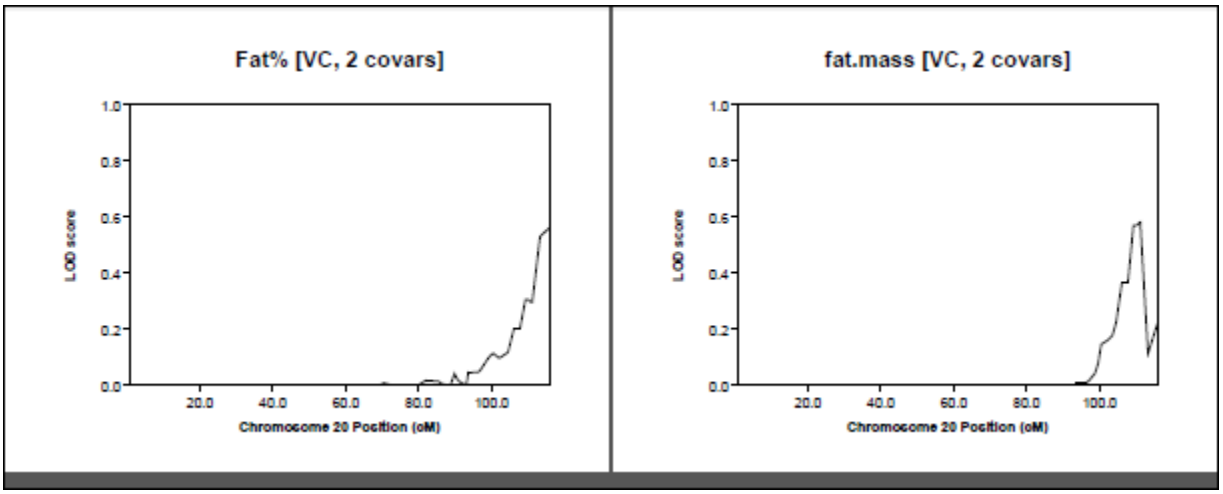

Muscle_mass [VC, 2 covars]
LOD score
Chromosome 18 Position (cM)
Muscle_Mass% [VC, 2 covars]
LOD score
Chromosome 18 Position (cM)
Bone.mass [VC, 2 covars]
LOD score
Chromosome 18 Position (cM)
BMR [VC, 2 covars]
LOD score
Chromosome 18 Position (cM)
BMI [VC, 2 covars]
LOD score
Chromosome 18 Position (cM)
Waist [VC, 2 covars]
LOD score
Chromosome 18 Position (cM)

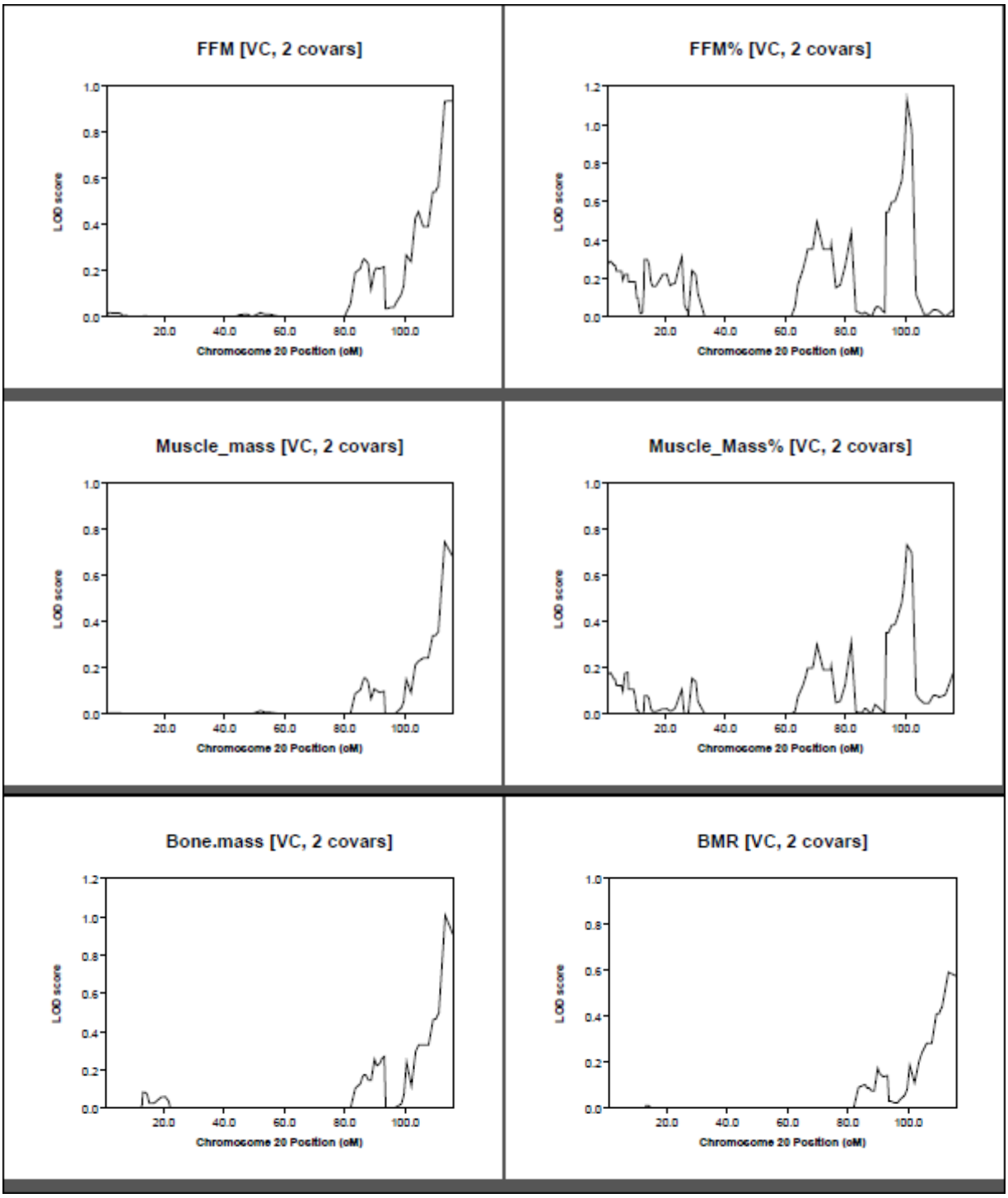
Chromosome 19

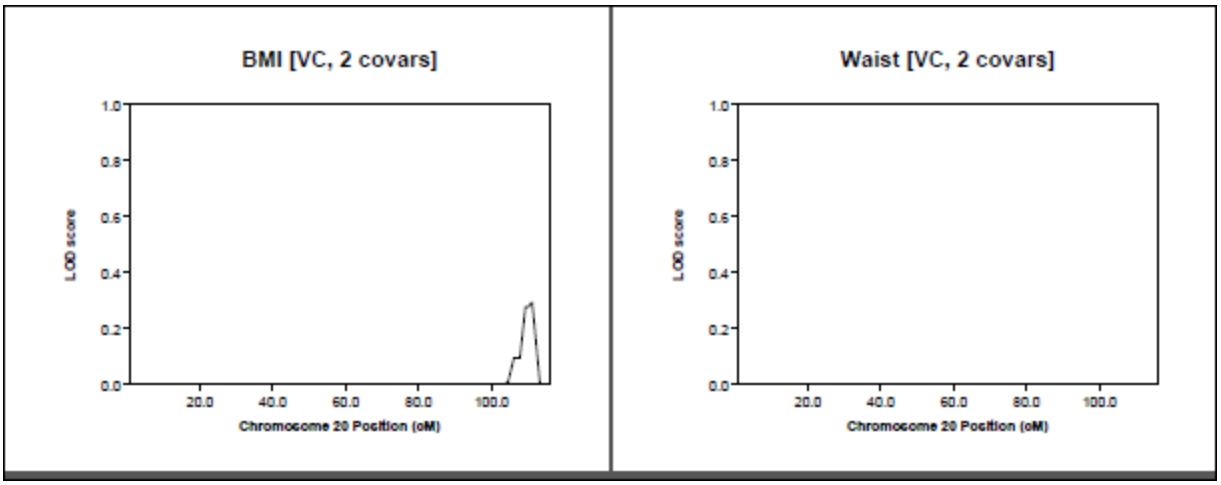




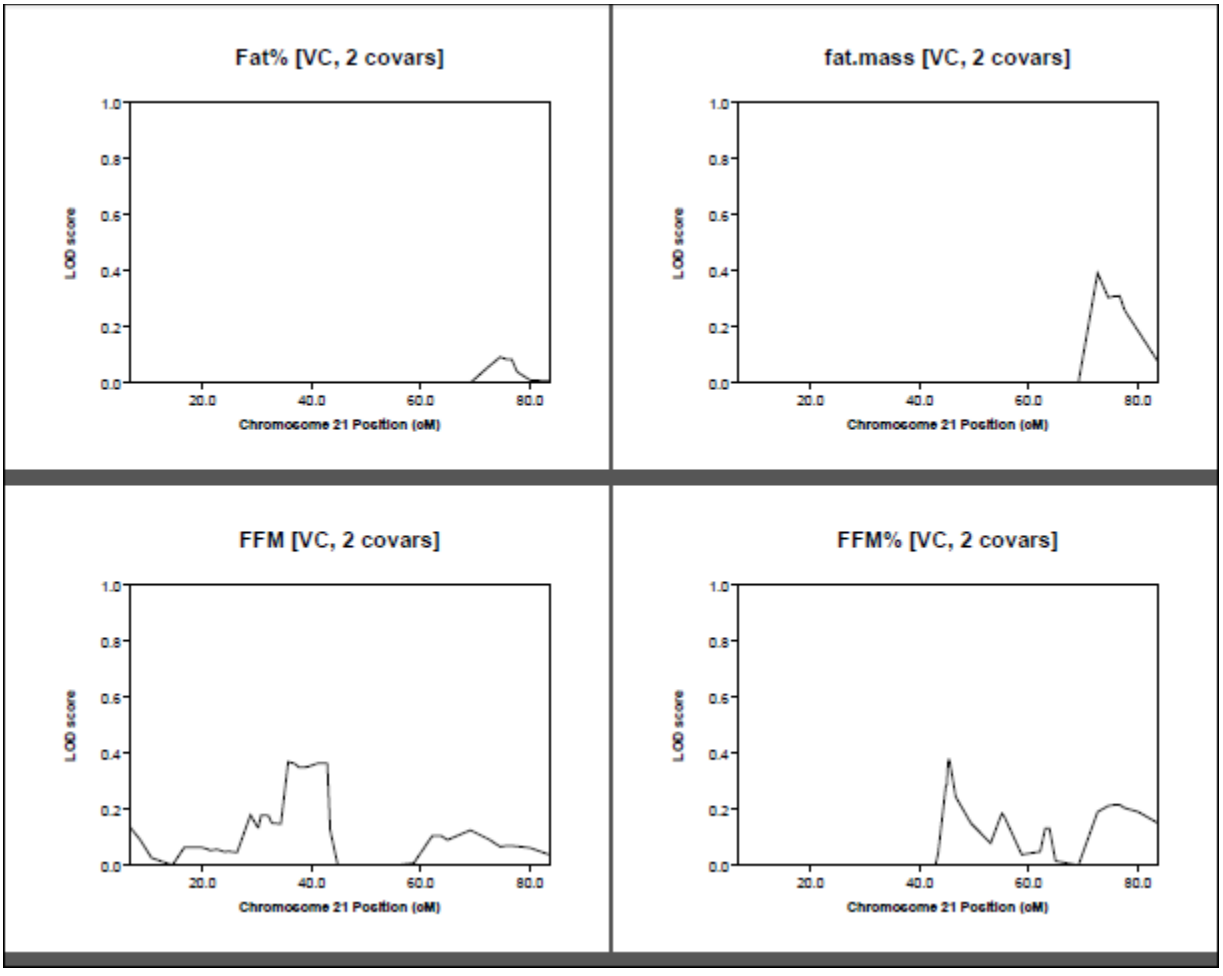
Chromosome 20

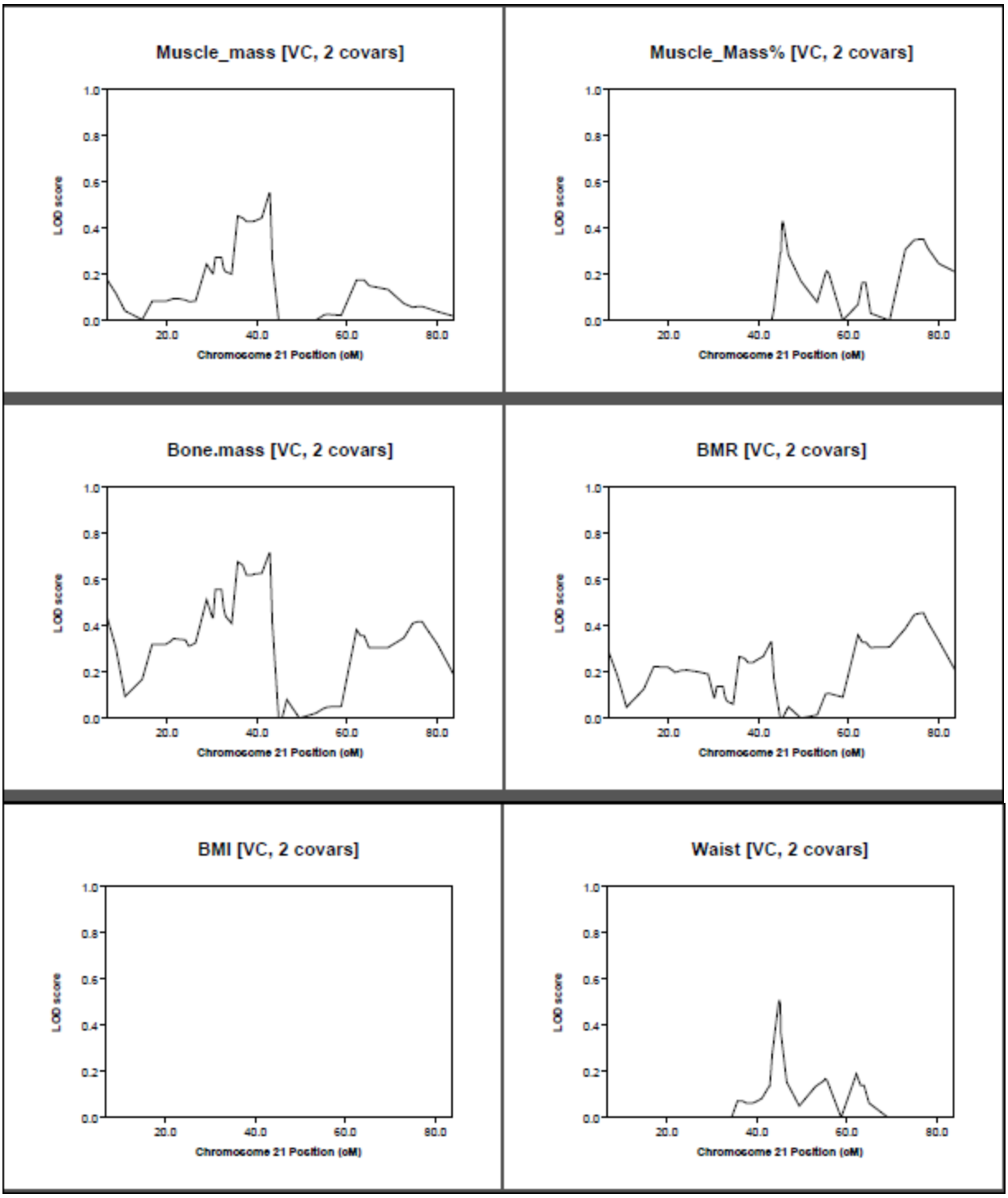



FFM [VC, 2 covars]
LOD score
Chromosome 20 Position (cM)
FFM% [VC, 2 covars]
LOD score
Chromosome 20 Position (cM)
Muscle_mass [VC, 2 covars]
LOD score
Chromosome 20 Position (cM)
Muscle_Mass% [VC, 2 covars]
LOD score
Chromosome 20 Position (cM)
Bone.mass [VC, 2 covars]
LOD score
Chromosome 20 Position (cM)
BMR [VC, 2 covars]
LOD score
Chromosome 20 Position (cM)

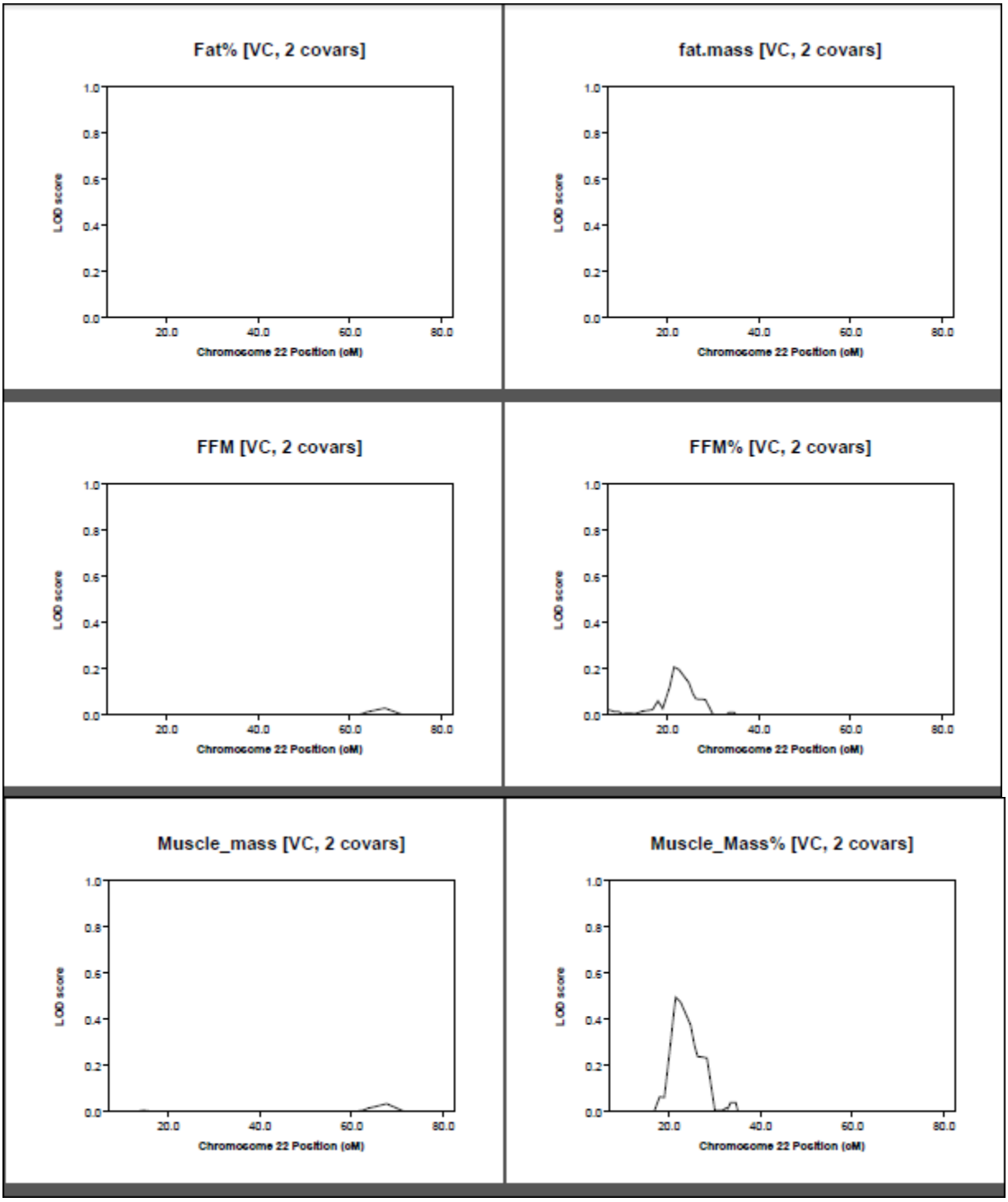


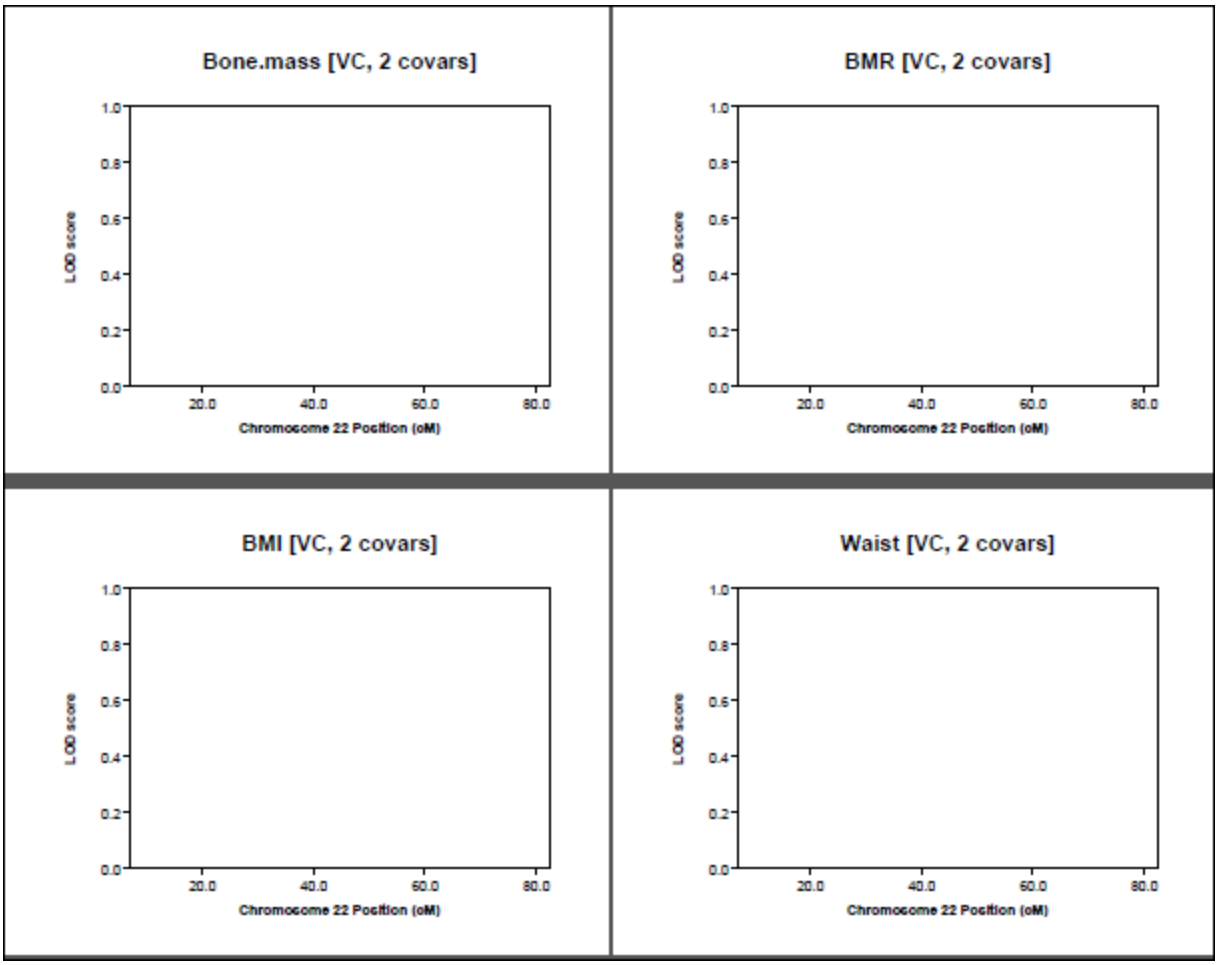
Chromosome 21




Muscle_mass [VC, 2 covars]
Muscle_Mass% [VC, 2 covars]
Bone.mass [VC, 2 covars]
BMR [VC, 2 covars]
BMI [VC, 2 covars]
Waist [VC, 2 covars]
LOD score
Chromosome 21 Position (cM)
0.0
0.2
0.4
0.6
0.8
1.0
20.0
40.0
60.0
80.0

Chromosome 22




Bone.mass [VC, 2 covars]
BMR [VC, 2 covars]
BMI [VC, 2 covars]
Waist [VC, 2 covars]
LOD score
0.0
0.2
0.4
0.6
0.8
1.0
20.0
40.0
60.0
80.0
Chromosome 22 Position (cM)

Appendix 1.6Summary of overlapping regions between linkage and ROH regions including information of the genes present within linkage regions.

Chr	Linkage Trait	Linkage Region Start	Linkage Region End	Overlapping ROH pools	Genes within linkage region
3	Obesity	54,807,240	73,964,127	S1(6), S4(2), S5(3)	*ARF4,DNASE1L3,FHIT,FLNB,GPR27,MITF,PDHB,PPIAP16,PTPRG,ATXN7,TMF1,WNT5A,SLMAP,ACOX2,CADPS,SUCLG2,HESX1,UBA3,MAGI1,PSMD6,ARL6IP5,RPP14, FAM107A,PDZRN3,FRMD4B,C3orf63,RYBP,LRIG1,ERC2,APPL1,FOXP1,ARHGEF3,IL17RD,PXK,FEZF2,FLJ10213,SHQ1,LMOD3,ADAMTS9,ABHD6,LRTM1,C3orf14,PROK2,THOC7,ID2B,KBTBD8,SLC25A26,FAM3D,LOC131691,C3orf49,S100A1L,SYNPR,LOC132205,ASB14,PSMC1P,FAM19A4,PPP4R2,PRICKLE2,C3orf67,KCTD6,DNHD2, PDE12,FAM116A,NPCDR1,C3orf64,CCDC66,EIF4E3,PSMD12P,hCG_1813818,LOC344709,SPATA12,DNAH12L,LOC375350,LOC389127,LOC391540,LOC391541, FAM19A1,LOC642487,LOC642567,CCDC137P,LOC654340,GLT8D4,LOC730057,LOC100128160,LOC100128209,LOC100128353,LOC100128448,LOC100128936, LOC100129031,LOC100129282,LOC100129765,LOC100130345,LOC100131083,LOC100131151,LOC100131568,LOC100132534,LOC100132946.*
5	FFM %	140,372,222	158,131,574	S12(2), S13(2)	*ADRB2,ANXA6,ASSP10,ATOX1,CAMK2A,CD74,CDX1,CSF1R,CSNK1A1,DIAPH1,DPYSL3,SLC26A2,ERPL2,FAT2,FGF1,GLRA1,GLULL1,GM2A,GPX3,GRIA1,NR3C1,NDST1, HTR4,ITK,MFAP3,PCDH1,PCDHGC3,PDE6A,PDGFRB,POU4F3,PPP2R2B,RPL7P,RPS14,RPS20P4,SGCD,SLC6A7,SPARC,SPINK1,TAF7,TCOF1,PCDHGB4,ADAM19,HDAC3, CNOT8,HAND1,MED7,RNF14,CLINT1,PCDHGA8,KIAA0141,JAKMIP2,GNPDA1,G3BP1,TNIP1,C5orf4,FAM114A2,TCERG1,SPINK5,SOX30,SYNPO,ABLIM3,KIAA0194, ARHGAP26,LARP1,GEMIN5,PCDHGA12,CCDC69,PCDHB5,HAVCR1,CYFIP2,IL17B,MRPL22,PCDHB1,DCTN4,PCDH12,LARS,RBM27,PCDHB18,PCDHB17,THG1L,GALNT10,RBM22,PCDHGC5,PCDHGC4,PCDHGB7,PCDHGB6,PCDHGB5,PCDHGB3,PCDHGB2,PCDHGB1,PCDHGA11,PCDHGA10,PCDHGA9,PCDHGA7,PCDHGA6,PCDHGA5, PCDHGA4,PCDHGA3,PCDHGA2,PCDHGA1,PCDHGB8P,PCDHB15,PCDHB14,PCDHB13,PCDHB12,PCDHB11,PCDHB10,PCDHB9,PCDHB8,PCDHB7,PCDHB6,PCDHB4, PCDHB3,PCDHB2,NMUR2,KCTD16,PCDHB16,HMHB1,ARAP3,PCYOX1L,SH3TC2,SAP30L,NDFIP1,FBXO38,YIPF5,TIGD6,SPRY4,SLC25A2,PCDHB19P,PCDHGB9P,SPINK7,HAVCR2,MST150,FCHSD1,TIMD4,ZNF300,MYOZ3,SCGB3A2,PPARGC1B,MRPL11P2,AFAP1L1,GRPEL2,LSM11,GPR151,LOC134466,SLC36A2,SPINK5L3,LOC153469, PPP1R2P3,FAM71B,PRELID2,SH3RF2,PLAC8L1,STK32A,SLC36A1,hCG_1980447,RELL2,SLC36A3,KIF4B,ARSI,LOC345471,IRGM,ICHTHYIN,C5orf46,FLJ41603, LOC402233,SPINK6,SPINK5L2,C5orf40,LOC442142,GRXCR2,SPINK9,LOC644198,LOC644762,LOC650866,LOC727846,LOC728264,LOC728287,LOC729080,LOC729123,hCG_1641617,LOC729421,LOC100128116,LOC100128121,LOC100128543,LOC100128833,LOC100128898,LOC100130088,LOC100130177,LOC100130260, LOC100131033,LOC100131520,LOC100131740,LOC100132712.*
5	Bone mass	82,904,528	123,449,266	S6(2), S7(2), S8(1), S9(1), S17(2), S20(2), S35(1)	*APC,CAMK4,CAST,CCNH,CDO1,CETN3,CHD1,AP3S1,COX7C,HAPLN1,CSNK1G3,CTNNAP1,DMXL1,EFNA5,FER,GJA1P1,GLRX,HSD17B4,KCNN2,LNPEP,LOX,MAN2A1, MCC,MEF2C,PAM,PCSK1,PGGT1B,PPIC,PTMAP2,RASA1,RPS17P2,RPS20P3,SNX2,SRP19,NR2F1,ZRSR1,XBPP1,ST8SIA4,REEP5,ATG12,C5orf13,RAB9P1,SNCAIP,TTC37,PJA2,EDIL3, POLR3G,RHOBTB3,ELL2,KIAA0825,HISPPD1,TNFAIP8,SNX24,TMED7,PRR16,COMMD10,ERAP1,FLJ11235,GIN1,FLJ11292,TRIM36,RIOK2,FEM1C,SEMA6A,ARRDC3,EPB41L4A,ERAP2,FBXL17,YTHDC2,MCTP1,FLJ12078,NUDT12,SPATA9,TSSK1B,FAM172A,GPR98,ANKRD32,TSLP,C5orf30,SLC25A46,PRDM6,FTMT,C5orf26, LYSMD3,SLCO6A1,LOC133789,LOC133790,ZNF474,POU5F2,STARD4,WDR36,LOC134505,CEP120,LOC153364,TMEM161B,SRFBP1,ARSK,FAM81B,CCDC112,DCP2, LIX1,C5orf27,LVRN,PSME2P1,C5orf36,GPR150,DTWD2,NBPF22P,RGMB,LOC285706,RFESD,FAM170A,LOC345571,LOC345576,FAM174A,LOC348958,SLCO4C1,TICAM2,MRPS35P2,LOC391807,CCT7P2,LOC391811,LOC391813,KRT18P42,LOC391825,KRT18P16,hCG_1640785,LOC402219,YTHDF1P,LOC402221,DDX18P4,LOC402226,LOC402287,OR7H2P,LOC441066,FLJ42709,LOC441097,LOC441098,FLJ35946,ARGFXP1,LOC642737,LOC642909,FLJ43080,LOC643031,LOC643534,LOC643992, LOC644100,LOC644146,LOC644285,LOC645125,LOC645181,LOC645261,LOC645323,LOC647954,SNORA13,LOC727938,LOC728093,LOC728104,LOC728217, LOC728254,LOC728342,LOC728366,LOC728460,LOC728682,LOC728711,LOC728979,LOC729011,LOC729040,LOC731157,LOC100127986,LOC100128407, LOC100128691,LOC100128694,LOC100129010,LOC100129099,LOC100129233,LOC100129283,LOC100129369,LOC100129374,LOC100129526,LOC100129563, LOC100129564,LOC100129716,LOC100129959,LOC100129962,LOC100130087,LOC100130089,LOC100130252,FABP5L6,LOC100130699,LOC100130703, LOC100130786,LOC100130882,LOC100130968,LOC100131040,LOC100131042,LOC100131066,LOC100131073,LOC100131236,LOC100131280,LOC100131573, LOC100132014,LOC100132020,LOC100132456,LOC100132811,LOC100133050,LOC100133054,LOC100133277.*

Chr	Linkage Trait	Linkage Region Start	Linkage Region End	Overlapping ROH pools	Genes within linkage region
5	FFM%	134,769,827	162,385,429	S12(2), S13(2)	*ADRA1B,ADRB2,ANXA6,ASSP10,ATOX1,CAMK2A,CD14,CD74,CDC25C,CDX1,CSF1R,CSNK1A1,CTNNA1,DIAPH1,DPYSL3,SLC26A2,HBEGF,EBF1,EGR1,ERPL2,ETF1, FABP6,FAT2,FGF1,GABRA1,GABRA6,GABRB2,GABRG2,GFRA3,GLRA1,GLULL1,GM2A,GPX3,GRIA1,NR3C1,HARS,HSPA9,NDST1,HTR4,IK,IL9,IL12B,ITK,LECT2,SMAD5, MFAP3,NDUFA2,NEUROG1,NPY6R,PCDH1,PCDHGC3,PDE6A,PDGFRB,PFDN1,POU4F3,PPP2R2B,PURA,RPL7P,RPS14,RPS20P4,SGCD,SLC6A7,SPARC,SPINK1,SPOCK1, TAF7,TCOF1,TGFBI,TTC1,UBE2D2,WNT8A,NME5,EIF4EBP3,PCDHGB4,CDC23,ADAM19,HDAC3,PTTG1,CNOT8,HAND1,MED7,MYOT,NRG2,CXCL14,SMAD5OS,RNF14, CLINT1,PCDHGA8,PCDHA9,MATR3,KIAA0141,JAKMIP2,SLC23A1,GNPDA1,SRA1,KIF20A,G3BP1,APBB3,TNIP1,SLU7,C5orf4,FAM114A2,BRD8,TCERG1,HNRNPA0, SPINK5,SOX30,SYNPO,ABLIM3,KIAA0194,ARHGAP26,ATP10B,LARP1,HARS2,GEMIN5,PCDHGA12,LRRTM2,CCDC69,PCDHB5,FBXL21,KLHL3,PCDHA14,HAVCR1, SNORD63,SNORA74A,CYFIP2,PKD2L2,IL17B,MRPL22,PCDHB1,DCTN4,MGC29506,PAIP2,PCDH12,FAM13B1,FAM53C,REEP2,LARS,CXXC5,JMJD1B,RBM27,PCDHB18, PCDHB17,WDR55,ANKHD1,THG1L,TMCO6,GALNT10,RBM22,PCDHGC5,PCDHGC4,PCDHGB7,PCDHGB6,PCDHGB5,PCDHGB3,PCDHGB2,PCDHGB1,PCDHGA11, PCDHGA10,PCDHGA9,PCDHGA7,PCDHGA6,PCDHGA5,PCDHGA4,PCDHGA3,PCDHGA2,PCDHGA1,PCDHA@,PCDHACT,PCDHGB8P,PCDHB15,PCDHB14,PCDHB13, PCDHB12,PCDHB11,PCDHB10,PCDHB9,PCDHB8,PCDHB7,PCDHB6,PCDHB4,PCDHB3,PCDHB2,PCDHAC2,PCDHAC1,PCDHA13,PCDHA12,PCDHA11,PCDHA10,PCDHA8, PCDHA7,PCDHA6,PCDHA5,PCDHA4,PCDHA3,PCDHA2,PCDHA1,NMUR2,TRPC7,KCTD16,PCDHB16,HMHB1,LOC63920,SIL1,ARAP3,PCYOX1L,CCNJL,SH3TC2,SAP30L, NDFIP1,FBXO38,YIPF5,TIGD6,SPRY4,SLC4A9,SLC25A2,PCDHB19P,PCDHGB9P,PSD2,C5orf32,SPINK7,HAVCR2,MST150,FCHSD1,TIMD4,ZNF300,MYOZ3,SLC35A4, PWWP2A,C1QTNF2,SCGB3A2,PPARGC1B,MRPL11P2,AFAP1L1,GRPEL2,LSM11,GPR151,LOC134466,UBLCP1,C5orf20,SLC36A2,SPINK5L3,LOC153328,LOC153469, ZMAT2,PPP1R2P3,FAM71B,PRELID2,SH3RF2,PLAC8L1,RNF145,LOC202051,DNAJC18,STK32A,SLC36A1,hCG_1980447,RELL2,SLC36A3,KIF4B,TMEM173,ARSI, LOC345471,IRGM,ICHTHYIN,DND1,LOC389332,LOC389333,C5orf46,FLJ41603,LOC391834,LOC391836,LOC402233,SPINK6,ANKHD1.*
10	Muscle mass %	77,379,512	93,347,297	S1(5), S35(3)	*ACTA2,ANXA11,FAS,BMPR1A,EIF5AP1,GLUD1,GRID1,HTR7,IFIT2,IFIT1,IFIT3,KCNMA1,LIPA,MAT1A,PTEN,RGR,RPS24,SFTPA1B,SFTPA2B,SFTPD,SNCG,MBL1P1,LIPF, CH25H,PAPSS2,DLG5,MINPP1,KIF20B,PPIF,RPP30,NRG3,C10orf116,POLR3A,LDB3,WAPAL,IFIT5,LRIT1,ANKRD1,GHITM,PANK1,KIAA1128,FAM35A,C10orf59,ZMIZ1, STAMBPL1,MMRN2,C10orf57,TSPAN14,C10orf58,DYDC2,PCGF5,ATAD1,PCDH21,OPN4,ANKRD22,LOC118945,LOC119358,NUDT9P1,AGAP11,LIPJ,CFLP1,DYDC1, EIF5AL1,HECTD2,PLAC9,ZCCHC24,FAM22E,FLJ37201,LOC283050,LIPM,LRIT2,RPL7AP8,IMPDH1P5,SH2D4B,C10orf99,SLC16A12,LOC389990,LOC389992,LOC399783, LOC401646,BMS1P3,LOC439992,LOC439994,IFIT1L,TPRX1P1,TPRX1P2,LOC642300,LOC642361,LOC642413,LOC642521,LOC642538,EIF5AL3,LOC642666,LOC642934,FAM25A,LI PK,LIPN,hCG_2024094,LOC643599,LOC647532,LOC650623,SFTPA1,LOC727749,FAM22C,LOC727879,LOC727923,LOC727960,LOC728027,LOC728050, FAM22A,FAM22D,LOC728190,LOC728218,LOC729187,SFTPA2,FAM22B,LOC729815,LOC100128292,LOC100128309,LOC100128465,LOC100128756,LOC100128990,LOC100129 156,LOC100129552,LOC100130253,LOC100130698,LOC100130759,LOC100130879,LOC100131132,DDX18P6,LOC100131699,LOC100132116, LOC100132402,LOC100132477,LOC100132487,LOC100132592,LOC100132987,LOC100133190.*
10	Waist	130,744,348	133,198,982	S5(3), S11(3)	*MGMT,GLRX3,CTAGEP,EBF3,TCERG1L,LOC387723,LOC100129103.*
10	FFM %	105,727,659	117,718,290	S20(1), S1(5), S35(3), S14(2), S20(1)	*ADD3,ADRA2A,ADRB1,CASP7,COL17A1,DUSP5,HABP2,ABLIM1,MXI1,NRAP,TCF7L2,TECTB,XPNPEP1,SHOC2,SMC3,GSTO1,DCLRE1A,SMNDC1,SORCS3,ATRNL1, RNU4P5,PDCD4,ACSL5,C10orf118,TDRD1,GPAM,FAM160B1,ZDHHC6,hCG_1776259,C10orf81,C10orf79,AFAP1L2,ITPRIP,YWHAZP5,NCRNA00081,SORCS1,GSTO2, C10orf78,TRUB1,VTI1A,LOC143188,CCDC147,RBM20,LOC282997,VWA2,NHLRC2,AURKAPS2,GUCY2G,PPIAP19,LOC644511,LOC645318,LOC645360,LOC724065, LOC751602,LOC100128304,LOC100129732,LOC100129859,LOC100130175,LOC100130319,LOC100130409,LOC100131338,LOC100132397,LOC100132573, LOC100132839,LOC100132843,LOC100133269.*

chr	Linkage Trait	Linkage Region Start	Linkage Region End	Overlapping ROH pools	Genes within linkage region
10	BMR	63,043,987	121,480,161	S1(5), S35(3), S14(1), S20(5)	*ACTA2,ADD3,ADK,ADRA2A,ADRB1,ANXA2P3,ANXA7,ANXA11,FAS,ARL3,ATQL4,BMPR1A,CALM2P2,CAMK2G,CASP7,ENTPD1,CHUK,ABCC2,COL13A1,COL17A1,COX15,CPN1,CYP 2C19,CYP2C8,CYP2C9,CYP2C18,CYP17A1,CYP26A1,DNA2,DNTT,DUSP5,EGR2,EIF4EBP2,EIF5AP1,EMX2,FGF8,GFRA1,GLUD1,GLUDP3,GOT1,PRLHR,GRK5, GRID1,HABP2,HELLS,HHEX,HK1,HNRNPH3,TLX1,HPS1,HTR7,IDE,IFIT2,IFIT1,IFIT3,KCNMA1,KIF11,ABLIM1,LIPA,MAT1A,MXI1,NDUFB8,NFKB2,NODAL,NRAP,P4HA1, PAX2,PCBD1,PDE6C,PGAM1,PITX3,PLAU,PNLIP,PNLIPRP1,PNLIPRP2,PPA1,PPIAP13,PPP1R3C,PPP3CB,PRF1,SRGN,PSAP,PSD,PTEN,ALDH18A1,RBP4,RGR,RGS10, RPS24,SCD,SFRP5,SFTPA1B,SFTPA2B,SFTPD,FBXW4,SLC18A2,SLIT1,SNCG,SUPV3L1,TACR2,TAF5,TCF7L2,TECTB,TIAL1,TLL2,VCL,VDAC2,WNT8B,XPNPEP1,SLC25A16, SHOC2,NDST2,MBL1P1,LIPF,EIF3A,GBF1,LDB1,SGPL1,BTRC,CH25H,PKD2L1,BTAF1,PAPSS2,INA,PDLIM1,SMC3,NEURL,DDX21,LGI1,NOLC1,DLG5,GSTO1,CHST3,BAG3,VPS26A,MI NPP1,KIF20B,SEC24C,SH3PXD2A,SLK,SPOCK2,ZNF518A,DCLRE1A,FRAT1,PPIF,ACTR1A,SMNDC1,NPM3,CBARA1,RPP30,SORBS1,ERLIN1,LBX1,NRG3, MGEA5,PRDX3,C10orf116,VAX1,POLR3A,LDB3,ECD,RAB11FIP2,CPEB3,ZNF365,NT5C2,PDCD11,SORCS3,KIAA0913,WAPAL,PPRC1,RRP12,TBC1D12,DNAJC9,DNMBP, FRAT2,SIRT1,MYST4,TSPAN15,IFIT5, RP11.*
13	Waist	34,637,818	61,201,257	S8(5)	*ATP7B,CALM2P3,RCBTB2,CPB2,ELF1,ESD,FKBP1P3,FOXO1,MLNR,GTF2F2,GUCY1B2,HTR2A,KPNA3,LCP1,MAB21L1,SMAD9,NEK3,PCDH8,RB1,RFXAP,TPT1,TRPC4, TNFSF11,SUCLA2,DLEU2,TSC22D1,CCNA1,DCLK1,RAD17P2,COX7CP1,ITM2B,MTRF1,UTP14C,P2RY5,SLC25A15,LHFP,TRIM13,MRPS31,DLEU1,OLFM4,POSTN,SUGT1, LECT1,WBP4,AKAP11,EXOSC8,FNDC3A,KIAA0564,ZC3H13,SPG20,LRCH1,INTS6,CKAP2,OR7E37P,OR7E36P,NUFIP1,PCDH17,C13orf15,MED4,DNAJC15,ALG5,VPS36, PHF11,UFM1,SOHLH2,ENOX1,RCBTB1,NUDT15,KIAA1704,FAM48A,THSD1,FABP3P2,CYSLTR2,C13orf1,COG6,PCDH20,MGC5590,NARG1L,RNASEH2B,DHRS12, C13orf18,C13orf23,OR7E101P,TDRD3,CDADC1,CAB39L,DIAPH3,COX17P,CCDC70,COG3,SETDB2,KBTBD7,EBPL,KBTBD6,EPSTI1,ARL11,WDFY2,LOC121838,RP11.*
14	Obesity	54,221,757	78,870,408	S1(4), S2(1)	*ACTN1,ACYP1,PARP1P2,ARG2,ZFP36L1,ENTPD5,DLST,EIF2S1,ERH,ESR2,ESRRB,FDPSL3,FNTB,FOS,FUT8,GCH1,GPX2,GSTZ1,HIF1A,HSPA2,KTN1,LGALS3,LTBP2,MAX, MAP3K9,ALDH6A1,MNAT1,MTHFD1,SIX6,OTX2,PGF,PIGH,PPM1A,PPP2R5E,PRKCH,PSEN1,PSMA3,ABCD4,RAD51L1,RAP1AP,ARID4A,RTN1,SFRS5,SIX1,SLC8A3, SLC10A1,SNAPC1,SPTB,TGFB3,ZBTB25,DPF3,NUMB,ADAM21,ADAM20,WDR22,ALKBH1,EIF2B2,PNMA1,COX5AP1,AKAP5,SPTLC2,RGS6,KIAA0247,KIAA0586,DLGAP5,KIAA0317, MED6,GPHN,VTI1B,BATF,NPC2,AHSA1,EXOC5,ACOT2,TMED10,C14orf1,WDHD1,VASH1,KIAA0831,ZBTB1,SNW1,PCNX,DAAM1,TTLL5,SYNE2,ANGEL1 ,ZFYVE26,TTC9,PLEKHG3,SIPA1L1,WDR21A,PLEK2,TIMM9,MLH3,KCNH5,POMT2,YBX1P1,COQ6,FCF1,RDH11,COX16,DACT1,ATP6V1D,C14orf100,DHRS7,SIX4,ZFYVE1,C14orf101, FBXO34,C14orf105,EXDL2,C14orf115,SYNJ2BP,SLC39A9,FLVCR2,C14orf118,C14orf108,ACTR10,YLPM1,C14orf162,ADCK1,TMEM63C,PELI2,RHOJ,GALNTL1,PLEKHH1,TRMT5,ZNF 410,NGB,RBM25,C14orf133,SMOC1,C14orf4,MPP5,C14orf135,GPR135,FAM164C,C14orf169,C14orf45,SGPP1,C14orf156,DNAL1,RPS6KL1, SYT16,KIAA1737,C14orf56,PAPLN,MAGOHP,CHURC1,C14orf43,LIN52,NEK9,MAPK1IP1L,C14orf179,WDR89,RBM8B,C14orf149,RPS29P1,SOCS4,NAT12,UBA52P3, PPIAP5,LOC122867,GPHB5,PTBP1P,FRDAP,C14orf148,CYCSP1,JDP2,ISCA2,ACOT4,RDH12,RPL24P3,ADAM21P,NUP50P1,C14orf50,SLC38A6,C14orf37,PTGR2, FAM161B,C14orf166B,RPLP1P1,THSD3,RPSAP3,FAM71D,C14orf83,TMEM30B,C14orf174,UBE2L7P,RPL21P9,PROX2,ZDHHC22,TMED8,C14orf178,BLZF2P,C14orf55, C14orf39,COX7AP1,DDX18P1,EIF2S2P,PPIAP6,HMGB1P,HMGN1P1,HNRPCP,SNORD56B,GCATP,RPL21P10,RPL21P8,ABI1P,SRMP2,RPL12P7,SF3B4P,RPL22P2, HSBP1P1,HSPEP2,RPL36AP2,RPL36AP3,MOCS3P,RPL41P4,RPL7AP5,RPL7AP6,RPS12P1,PTTG4P,RPS24P2,RPS2P2,MAD2L1P,NCOA4P,NDUFB8P1,NT5CP1,NT5CP2, PSMA3P,VSX2,SLC35F4,LRRC9,RAB15,FLJ31306,TBPL2,TOMM20L,AKR1B1P5,LOC390483,HEATR4,FLJ22447,C14orf181,FLJ44817,LOC401777,LOC440180,LOC440181,C14orf53, LOC441687,ACOT1,ACOT6,LOC645431,LOC645683,C14orf34,LOC645798,FLJ43390,LOC646279,FTHL13,LOC646548,TMEM90A,LOC646701,RPS28P1, LOC653695,RPL37P5,LOC729451,LOC729637,RPL31P4,C14orf38,LOC729850,LOC729868,LOC729941,LOC730011,LOC731223,LOC100036574,LOC100128233, LOC100128793,C14orf33,RPL9P5,LOC100129185,LOC100129782,LOC100129913,LOC100129928,LOC100130048,LOC100130174,LOC100130399,LOC100130473, LOC100130542,LOC100130614,LOC100130953,LOC100130971,LOC100131379,LOC100131497,LOC100131580,LOC100131708,LOC100131950,LOC100132365, LOC100132603.*

chr	Linkage Trait	Linkage Region Start	Linkage Region End	Overlapping ROH pools	Genes within linkage region
18	FFM %	4,468,352	33,015,239	S35(3), S14(1)	*AQP4,C18orf1,CDH2,CIDEA,DSC1,DSC2,DSC3,DSG1,DSG2,DSG3,DTNA,GALNT1,GATA6,GNAL,IMPA2,LAMA3,MC2R,MC5R,MEP1B,NDUFV2,NPC1,PPIAP14,PTPN2, PTPRM,RBBP8,ROCK1,SNRPD1,SS18,TAF4B,TTR,ZFP161,ZNF24,NOL4,RNMT,NAPG,RIOK3,VAPA,B4GALT6,PPP4R1,SLMO1,ZNF271,RALBP1,AFG3L2,MAPRE2,RAB31, KIAA1012,EPB41L3,ANKRD12,KIAA0802,SLC39A6,ZNF521,C18orf10,CABYR,RNU4P3,C18orf8,RNF138,RNF125,MOCOS,CEP192,RPRD1A,ELP2,IMPACT,SPIRE1,PSMG2,TWSG1,CH MP1B,MIB1,KLHL14,HRH4,FAM38B,CTAGE1,FAM59A,MPPE1,OR4K7P,ARHGAP28,CEP76,KIAA1772,FLJ11996,FHOD3,ASXL3,SEH1L,CHST9,C18orf21, TXNDC2,ZNF397,TUBB6,C18orf45,L3MBTL4,CABLES1,ESCO1,OSBPL1A,C18orf19,KRT18P8,INO80C,TTC39C,PSMA8,C18orf15,MCART2,DSG4,C18orf16,ANKRD29, WBP2P1,APCDD1,C18orf18,ZNF519,ABHD3,RAB12,ZNF396,LAMA1,C18orf58,LOC284230,KCTD1,LOC284269,LOC339290,LRRC30,LOC342705,DKFZp779B1634, ANKRD30B,C18orf34,FGF7P1,RAC1P1,HNRPA1L3,A26B2,LOC390831,LOC390834,OR4K8P,LOC390843,ARIH2P,hCG_1642947,LOC390848,FLJ38028,FLJ45994, CXADRP3,MGC26718,LOC440487,LOC440489,LOC440491,LOC441811,WBP11P1,DHFRP1,PA2G4P3,LOC642597,LOC644669,LOC645355,TTMA,LOC645387, LOC645423,LOC645573,LOC645859,AMAC1L1,LOC646044,LOC646053,LOC646065,LOC646171,LOC646203,LOC646359,LOC646442,LOC646583,LOC646595, LOC646629,LOC646778,LOC646842,LOC647983,LOC648532,LOC728211,LOC728473,LOC728592,CIAPIN1P,LOC728606,LOC729309,LOC729391,LOC729589, LOC729602,LOC729774,LOC729804,LOC729863,LOC729950,LOC730045,LOC732448,LOC791126,ZNF397OS,LOC100128219,LOC100128238,LOC100128324, LOC100128776,LOC100128893,LOC100128925,LOC100129355,LOC100129556,LOC100129769,LOC100129773,LOC100130055,LOC100130329,LOC100130404, LOC100130468,LOC100130480,LOC100130487,LOC100131036,LOC100131282,LOC100131500,LOC100131915,LOC100131966,LOC100132255,LOC100132664, LOC100132679.*
18	Muscle mass %	4,468,352	33,015,239	S35(3), S14(1)	*AQP4,C18orf1,CDH2,CIDEA,DSC1,DSC2,DSC3,DSG1,DSG2,DSG3,DTNA,GALNT1,GATA6,GNAL,IMPA2,LAMA3,MC2R,MC5R,MEP1B,NDUFV2,NPC1,PPIAP14,PTPN2, PTPRM,RBBP8,ROCK1,SNRPD1,SS18,TAF4B,TTR,ZFP161,ZNF24,NOL4,RNMT,NAPG,RIOK3,VAPA,B4GALT6,PPP4R1,SLMO1,ZNF271,RALBP1,AFG3L2,MAPRE2,RAB31, KIAA1012,EPB41L3,ANKRD12,KIAA0802,SLC39A6,ZNF521,C18orf10,CABYR,RNU4P3,C18orf8,RNF138,RNF125,MOCOS,CEP192,RPRD1A,ELP2,IMPACT,SPIRE1,PSMG2,TWSG1,CH MP1B,MIB1,KLHL14,HRH4,FAM38B,CTAGE1,FAM59A,MPPE1,OR4K7P,ARHGAP28,CEP76,KIAA1772,FLJ11996,FHOD3,ASXL3,SEH1L,CHST9,C18orf21, TXNDC2,ZNF397,TUBB6,C18orf45,L3MBTL4,CABLES1,ESCO1,OSBPL1A,C18orf19,KRT18P8,INO80C,TTC39C,PSMA8,C18orf15,MCART2,DSG4,C18orf16,ANKRD29, WBP2P1,APCDD1,C18orf18,ZNF519,ABHD3,RAB12,ZNF396,LAMA1,C18orf58,LOC284230,KCTD1,LOC284269,LOC339290,LRRC30,LOC342705,DKFZp779B1634, ANKRD30B,C18orf34,FGF7P1,RAC1P1,HNRPA1L3,A26B2,LOC390831,LOC390834,OR4K8P,LOC390843,ARIH2P,hCG_1642947,LOC390848,FLJ38028,FLJ45994, CXADRP3,MGC26718,LOC440487,LOC440489,LOC440491,LOC441811,WBP11P1,DHFRP1,PA2G4P3,LOC642597,LOC644669,LOC645355,TTMA,LOC645387, LOC645423,LOC645573,LOC645859,AMAC1L1,LOC646044,LOC646053,LOC646065,LOC646171,LOC646203,LOC646359,LOC646442,LOC646583,LOC646595, LOC646629,LOC646778,LOC646842,LOC647983,LOC648532,LOC728211,LOC728473,LOC728592,CIAPIN1P,LOC728606,LOC729309,LOC729391,LOC729589, LOC729602,LOC729774,LOC729804,LOC729863,LOC729950,LOC730045,LOC732448,LOC791126,ZNF397OS,LOC100128219,LOC100128238,LOC100128324, LOC100128776,LOC100128893,LOC100128925,LOC100129355,LOC100129556,LOC100129769,LOC100129773,LOC100130055,LOC100130329,LOC100130404, LOC100130468,LOC100130480,LOC100130487,LOC100131036,LOC100131282,LOC100131500,LOC100131915,LOC100131966,LOC100132255,LOC100132664, LOC100132679.*
20	FFM%	53,647,394	58,795,536	S10(1), S11(3), S24(2)	*ATP5E,BMP7,CSTF1,CTSZ,EDN3,GNAS,MC3R,PCK1,PPP1R3D,AURKA,TFAP2C,RAE1,STX16,VAPB,HMG1L1,SYCP2,SPO11,SLMO2,TH1L,C20orf43,PPP4R1L,RBM38, PMEPA1,CASS4,RAB22A,CDH26,C20orf177,MGC4294,NPEPL1,TUBB1,ZBP1,C20orf108,PHACTR3,RPS4L,C20orf85,MRPS16P,ZNF831,GCNT7,CBLN4,CTCFL,RPL39P, ANKRD60,MTCO2L,LOC149773,GNASAS,APCDD1L,PTMAP6,C20orf106,C20orf197,LOC284757,C20orf66,C20orf107,LOC645605,LOC728902,LOC729296, LOC100128291,LOC100128608,LOC100128748,LOC100129576,LOC100129869,LOC100130156,LOC100131175,LOC100131365,LOC100131710,LOC100131806, LOC100133271.*
20	Bone mass	217,281	56,511,689	S1(4), S2(4), S3(4), S4(4), S7(3), S8(3), S9(2), S10(1), S11(2), S24(2)	*ACTGP3,ADA,ADRA1D,JAG1,AHCY,ASIP,AVP,BCL2L1,BCL2L7P1,BFSP1,BMP2,BMP7,BPI,ENTPD6,CD40,CDC25B,CEBPB,CENPB,CHGB,CSE1L,CSNK2A1,CST1,CST2,CST3,CST4,CST5, CSTF1,CSTP1,CSTP2,CYB5P4,CYP24A1,DNMT3B,E2F1,EIF4EP1,EPB41L1,EYA2,FDXP1,FKBP1A,FOXS1,GGT7,GHRH,GNRH2,GSS,HCK,FOXA2,HNF4A,ID1, IDH3B,INSM1,EIF6,ITPA,KCNB1,KCNG1,KCNS1,LBP,MC3R,MMP9,MYBL2,NFATC2,NKX2.*

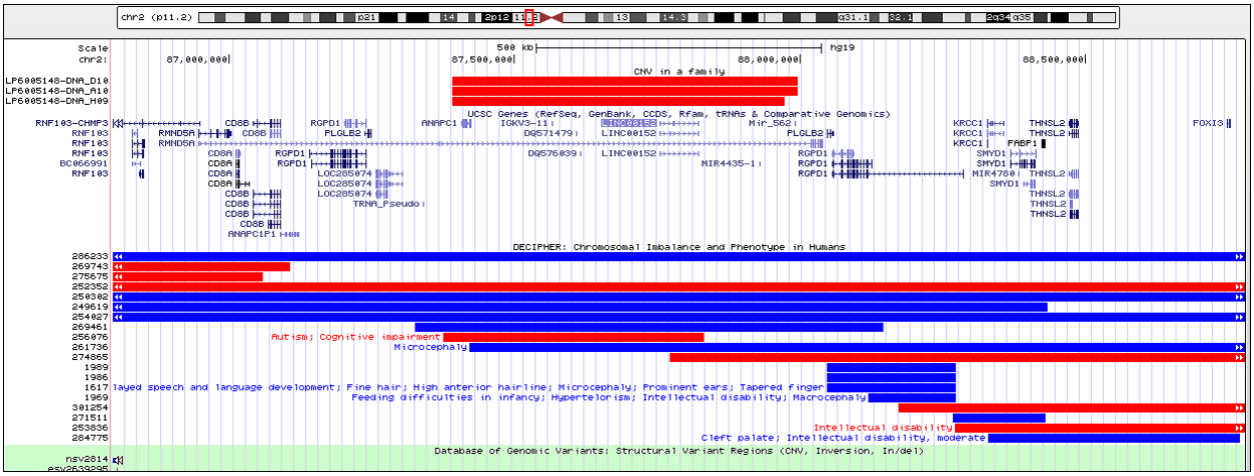
Appendix 1.7Large CNVs (>500kb) detected in the sample set.

Chr	Start (hg19)	End (hg19)	Size (kb)	# of SNPs	CNV State	Sample ID	Family ID	Age (y)	BMI (kg/m2)	Phenotype	Overlapping genes
2	87392157	87995393	603.2	36	Het Deletion	DMF5S2	4	56	32.9	Obese	*LINC00152,MIR4435-1,MIR4435-2*
2	87392157	87995393	603.2	36	Het Deletion	DMF5S5	4	53	28.1	Overweight	
2	87392157	87972393	580.2	35	Het Deletion	DMF5S1	4	41	33.6	Obese	
5	46324183	49548900	3224.7	92	Het Deletion	QO_F5_05	5	20	39.9	Obese	*NA*
5	46324183	49548900	3224.7	92	Het Deletion	QO_F9_08	9	23	30.2	Obese	
6	110883633	111591037	707.4	401	Het Duplication	DMF3S1	3	51	35.4	Obese	*AMD1,CDK19,GSTM2P1,GTF3C,KIAA1919,RPF2,SLC16A10*
6	110887169	111591037	703.9	396	Het Duplication	DMF3S2	3	54	28.7	Overweight	
8	42942128	43789824	847.7	400	Het Duplication	QO_F8_02	8	45	36.7	Obese	*FNTA,SGK196,HGSNAT,A26A1,LOC347028,hCG_1795560,LOC643654,LOC100128173,LOC100130474,LOC100130767,LOC100131789*
8	42937273	43788229	851.0	402	Het Duplication	QO_F8_03	8	19	42.6	Obese	
9	68996222	69992206	996.0	35	Het Deletion	QO_F5_04	5	18	19	Normal	*ANKRD20A4,BX255923.3,CBWD5,CBWD6,FOXD4L6,LOC100133920,PGM5P2*
9	68996222	69881429	885.2	33	Het Deletion	QO_F5_06	5	14	30.1	Obese	
10	46949421	47703946	754.5	264	Het Duplication	QO_F7_11	7	18	39.8	Obese	*ANXA8L2,FAM21B,ANTXRL,LOC340844,SLC9A3P4,CTGLF11P,ASAH2C,LOC728449*
10	46949421	47703946	754.5	264	Het Duplication	QO_F7_03	7	26	20.1	Normal	
10	46949421	47703946	754.5	264	Het Duplication	QO_F7_02	7	52	36.1	Obese	
11	50677929	51373124	695.2	80	Het Deletion	QO_F5_05	5	20	39.9	Obese	*OR4C7P,OR4R2P,OR4A8P,OR4A7P,OR4A5,OR4A3P,OR4C46,OR4A6P,OR4A2P,OR4A4P*
11	50751699	51373124	621.4	57	Het Deletion	QO_F4_01	1	62	26.6	Overweight	
11	50691976	51373124	681.1	74	Het Deletion	QO_F8_04	8	14	21.29	Normal	
11	50691976	51373124	681.1	74	Het Deletion	QO_F9_08	9	23	30.2	Obese	
14	25099980	25612838	512.9	469	Het Duplication	QO_F5_05	5	20	39.9	Obese	*GZMB,STXBP6*

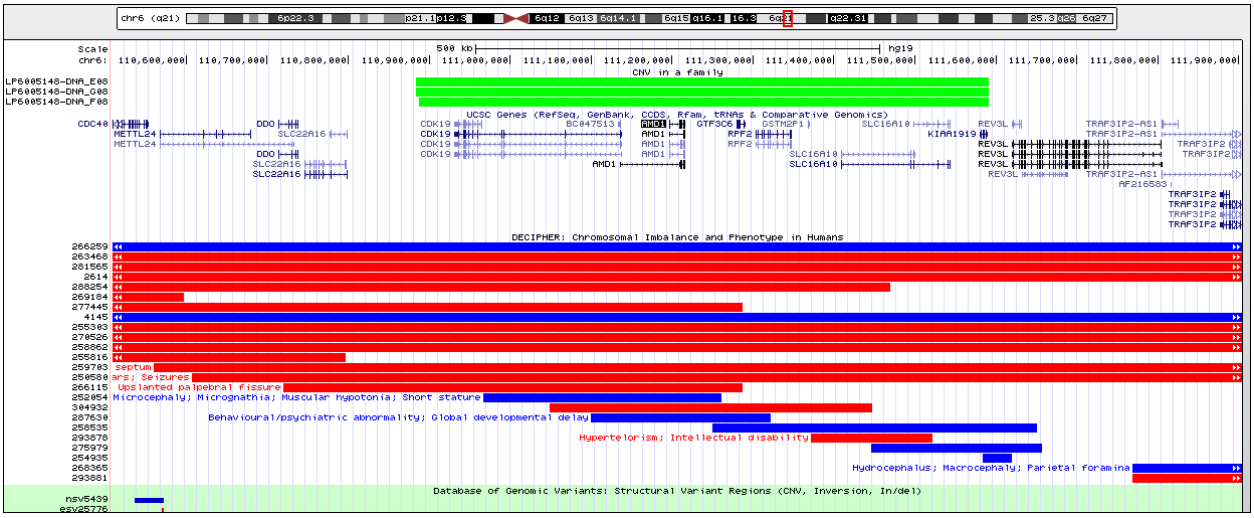
Chr	Start (hg19)	End (hg19)	Size (kb)	# of SNPs	CNV State	Sample ID	Family ID	Age (y)	BMI (kg/m2)	Phenotype	Overlapping genes
14	106,103,358	106,917,103	813.7	85	Het Duplication	QO_F6_09	6	55	24.7	Normal	*IGHVIII-82,IGHVIII-76-1,IGHVIII-67-4,IGHVIII-67-3,IGHVIII-67-2,IGHVIII-51-1,IGHVII-78-1,IGHVII-74-1,IGHVII-67-1,IGHVII-65-1,IGHVII-62-1,IGHVII-60-1,IGHVII-53-1,IGHVII-51-2,IGHV7-81,IGHV756,IGHV578,IGHV5-51,IGHV4-80,IGHV4-61,IGHV4-59,IGHV4-55,IGHV3-79,IGHV376,IGHV3-75,IGHV3-74,IGHV3-73,IGHV3-72,IGHV3-71,IGHV3-66,IGHV365,IGHV3-64,IGHV3-63,IGHV3-62,IGHV3-60,IGHV3-57,IGHV3-54,IGHV353,IGHV3-52,IGHV2-70,IGHV1-69,IGHV1-68,IGHV1-67,IGHV158,LOC192127,LOC192130,C14orf99.*
14	19,457,976	20,402,122	944.1	43	Het Duplication	QO_F7_11	7	18	39.8	Obese	*ANG,APEX1,NP,RNASE1,RNASE4,RNASE6,TEP1,PARP2,FAM12A,OSGEP,CCNB1IP1,FAM12B,OR4K5,OR4K1,OR4K16P,OR4K15,RPPH1,TMEM55B,TTC5,RNASE11,OR4K14,OR4L1,OR11H6,KLHL33,LOC254028,OR11P1P,RANBP20P,RNASE10,OR6S1,OR4Q2,OR4K13,OR4U1P,OR4T1P,OR4K17,OR4N5,OR11G1P,OR11G2,OR11H5P,OR11H7,OR11H4,RNASE9,RNASE12,SNORD126,LOC100128220,GAFA1,LOC100131302,LOC100132444,LOC100133314*
16	29,581,764	30,199,579	617.8	300	Het Deletion	QO_F4_02	1	32	51.8	Obese	*ALDOA,ASPHD1,C16orf54,C16orf92,CDIPT,CDIPTAS1,CORO1A,DOC2A,FAM57B,GDPD3,HIRIP3,INO80E, KCTD13,KIF22,MAPK3,MAZ,MVP,PAGR1,PPP4C,PRRT2,QPRT,SEZ6L2,SLC7A5P1,SPN,TAOK2,TBX6,TMEM219,YPEL3*
19	797,032	1,498,203	701.2	526	Het Duplication	QO_F4_02	1	32	51.8	Obese	*ATP5D,CIRBP,CNN2,CFD,ARID3A,EFNA2,ELA2,GAMT,GPX4,POLR2E,RPS15,STK11,MED16,APC2,ABCA7,SBNO2,HMHA1,DAZAP1,PCSK4,C19orf24,WDR18,KISS1R,MUM1,MIDN,C19orf22,C19orf6,REEP6,GRIN3B,PLK5P,C19orf23,C19orf25,C19orf26,ADAMTSL5,NDUFS7,LOC401895,LOC729119*
19	1,788,247	2,294,552	506.3	368	Het Duplication	QO_F4_02	1	32	51.8	Obese	*AMH,CSNK1G2,MKNK2,OAZ1,SF3A2,AP3D1,LSM7,PLEKHJ1,BTBD2,FAM108A1,KLF16,DOT1L,C19orf36,SCAMP4,ADAT3,JSRP1,MOBKL2A,C19orf34,C19orf35,LINGO3,LOC728187,UNQ6493*

Appendix 1.8 UCSC browser tracks for the seven large non-centromeric CNVs detected in the study

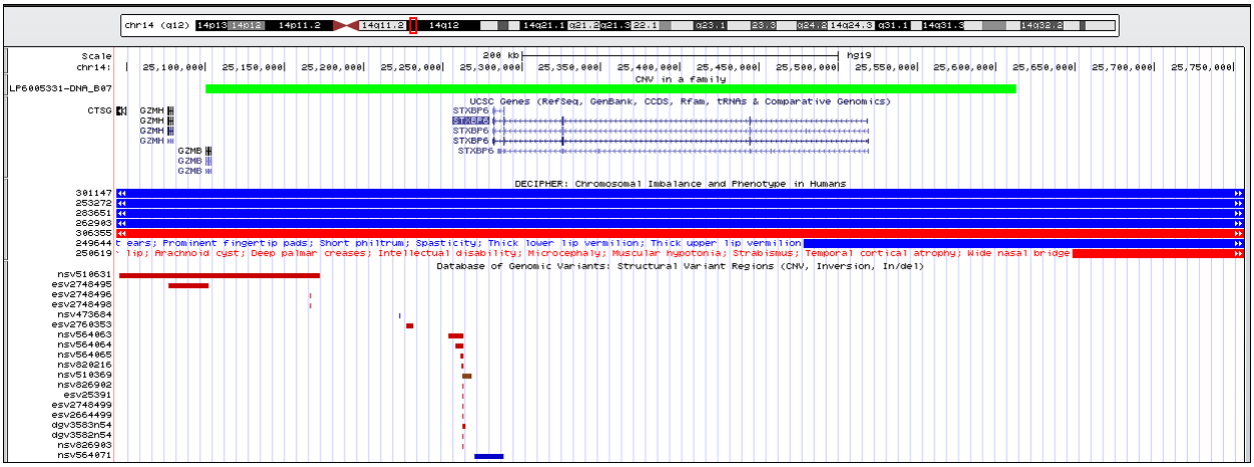
CNV on chromosome 2:



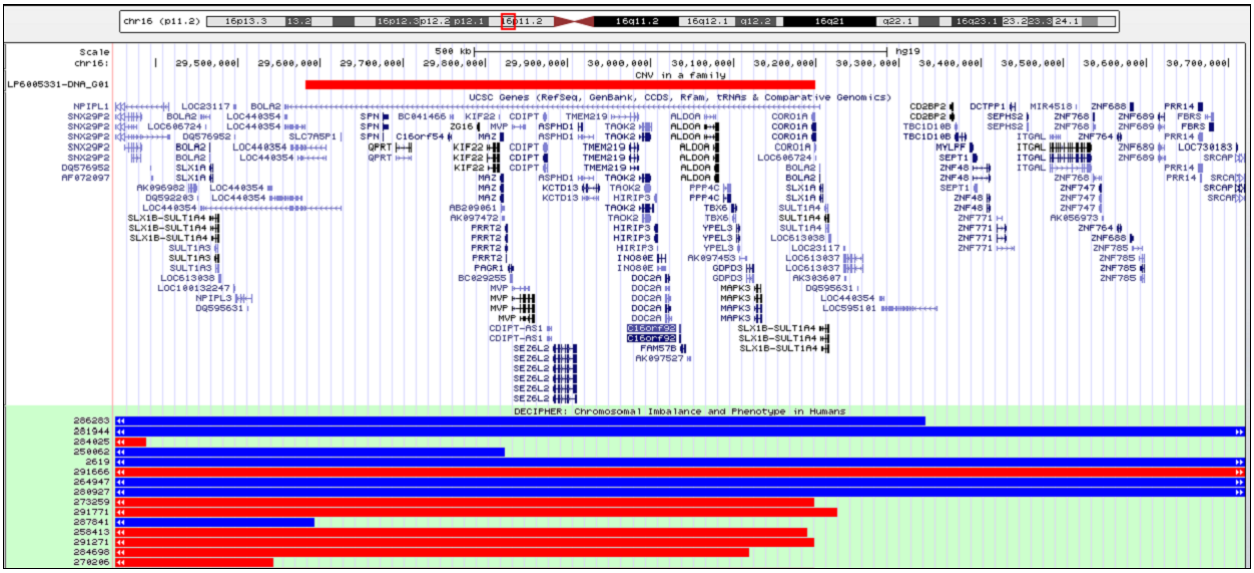
CNV on chromosome 6:



CNV on chromosome 14:



CNV on chromosome 16:



Appendix 1.9 The top 100 CpG hits in association with BMI

Marker	Chr	Position	P-value	Gene symbol	Gene name
cg17501210	6	166970252	8.23 X10^{-8}	*RPS6KA2*	ribosomal protein S6 kinase, 90kDa, polypeptide 2
cg18181703	17	76354621	1.72 X10^{-6}	*SOCS3*	suppressor of cytokine signaling 3
cg23943013	5	79551904	2.88 X10^{-6}	*SERINC5*	serine incorporator 5
cg05864010	11	85376948	4.60 X10^{-6}	*CREBZF*	CREB/ATF bZIP transcription factor
cg06183831	1	9883466	7.42 X10^{-6}	*CLSTN1*	calsyntenin 1
cg14551938	8	49173274	8.68 X10^{-6}	*NA*	NA
cg07146912	3	73671377	8.90 X10^{-6}	*PDZRN3*	PDZ domain containing ring finger 3
cg13057167	14	103750565	1.06X10^{-5}	*NA*	NA
cg03717755	6	16136539	1.12X10^{-5}	*MYLIP*	myosin regulatory light chain interacting protein
cg25499461	19	46389574	1.12X10^{-5}	*IRF2BP1*	interferon regulatory factor 2 binding protein 1
cg11519238	1	183605567	1.13X10^{-5}	*NA*	NA
cg19570406	X	152684911	1.18X10^{-5}	*ZFP92*	ZFP92 zinc finger protein
cg22192069	6	13305324	1.20X10^{-5}	*TBC1D7*	TBC1 domain family, member 7
cg15924102	1	234852892	1.28X10^{-5}	*NA*	NA
cg13808918	2	220041352	1.37X10^{-5}	*CNPPD1*	cyclin Pas1/PHO80 domain containing 1
cg14276242	17	5095449	1.44X10^{-5}	*ZNF594*	zinc finger protein 594
cg27563121	15	52405647	1.82X10^{-5}	*BCL2L10*	BCL2-like 10 (apoptosis facilitator)
cg15453943	5	104434827	1.86X10^{-5}	*RAB9BP1*	RAB9B, member RAS oncogene family pseudogene 1
cg23898951	10	4511903	1.88X10^{-5}	*NA*	NA
cg20158159	10	75683345	2.00X10^{-5}	*C10orf55*	chromosome 10 open reading frame 55
cg03786327	13	100644473	2.02X10^{-5}	*NA*	NA
cg18306516	10	81103258	2.05X10^{-5}	*NA*	NA
cg04763450	17	79455576	2.07X10^{-5}	*NA*	NA
cg04705499	19	46796459	2.12X10^{-5}	*NA*	NA
cg10661002	7	25018503	2.27X10^{-5}	*OSBPL3*	oxysterol binding protein-like 3
cg16264888	12	82752007	2.41X10^{-5}	*NA*	NA
cg02450666	1	236845763	2.54X10^{-5}	*NA*	NA
cg00491523	16	6913336	2.58X10^{-5}	*RBFOX1*	RNA binding protein, fox-1 homolog (C. elegans) 1
cg15930703	2	225775142	2.79X10^{-5}	*DOCK10*	dedicator of cytokinesis 10
cg19748509	6	29910778	2.79X10^{-5}	*HLA-A*	major histocompatibility complex, class I, A
cg13223561	2	132433923	2.85X10^{-5}	*NA*	NA
cg18369680	20	41984014	3.10X10^{-5}	*NA*	NA
cg24179288	6	29867285	3.11X10^{-5}	*NA*	NA
cg09260951	22	39096513	3.16X10^{-5}	*JOSD1*	Josephin domain containing 1

Marker	Chr	Position	P-value	Gene symbol	Gene name
cg23975840	12	117042895	$3.52X10^{-5}$	*NA*	NA
cg13882285	9	125111986	$3.98X10^{-5}$	*NA*	NA
cg23345151	18	77292947	$4.13X10^{-5}$	*NA*	NA
cg14258335	X	2843196	$4.29X10^{-5}$	*ARSD*	arylsulfatase D
cg06711873	19	22804548	$4.44X10^{-5}$	*NA*	NA
cg05012044	19	12171668	$4.46X10^{-5}$	*NA*	NA
cg22697239	11	44626708	$4.47X10^{-5}$	*CD82*	CD82 molecule
cg10270055	21	43221559	$4.52X10^{-5}$	*PRDM15*	PR domain containing 15
cg25810999	17	39156115	$4.68X10^{-5}$	*KRTAP3-2*	keratin associated protein 3-2
cg09103591	22	50699708	$4.70X10^{-5}$	*MAPK12*	mitogen-activated protein kinase 12
cg26113962	1	23546304	$4.71X10^{-5}$	*NA*	NA
cg05660681	19	53233083	$4.76X10^{-5}$	*ZNF611*	zinc finger protein 611
cg19169586	4	82157420	$4.81X10^{-5}$	*NA*	NA
cg03238359	11	60623567	$4.83X10^{-5}$	*PTGDR2*	prostaglandin D2 receptor 2
cg07977178	11	63997715	$4.83X10^{-5}$	*DNAJC4*	DnaJ (Hsp40) homolog, subfamily C, member 4
cg22141781	12	95397704	$4.84X10^{-5}$	*NDUFA12*	NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 12
cg23877906	8	67898826	$4.90X10^{-5}$	*NA*	NA
cg04150276	14	53256600	$4.99X10^{-5}$	*GNPNAT1*	glucosamine-phosphate N-acetyltransferase 1
cg19750657	13	38935967	$5.15X10^{-5}$	*UFM1*	ubiquitin-fold modifier 1
cg09966847	14	77943678	$5.16X10^{-5}$	*ISM2*	isthmin 2
cg05298224	6	32016360	$5.23X10^{-5}$	*TNXB*	tenascin XB
cg03889700	4	2869818	$5.24X10^{-5}$	*ADD1*	adducin 1 (alpha)
cg07573872	19	1126342	$5.30X10^{-5}$	*SBNO2*	strawberry notch homolog 2 (Drosophila)
cg03455746	6	29802584	$5.34X10^{-5}$	*NA*	NA
cg23866916	19	1155738	$5.40X10^{-5}$	*SBNO2*	strawberry notch homolog 2 (Drosophila)
cg24833065	19	55694352	$5.53X10^{-5}$	*PTPRH*	protein tyrosine phosphatase, receptor type, H
cg17115737	7	75580813	$5.60X10^{-5}$	*POR*	P450 (cytochrome) oxidoreductase
cg10504000	11	2402391	$5.63X10^{-5}$	*CD81*	CD81 molecule
cg14819942	15	35414228	$5.81X10^{-5}$	*NA*	NA
cg11389960	15	31277661	$5.94X10^{-5}$	*MTMR10*	myotubularin related protein 10
cg12046183	6	29859829	$6.19X10^{-5}$	*NA*	NA
cg27313201	1	154165904	$6.27X10^{-5}$	*TPM3*	tropomyosin 3
cg11772086	5	112535690	$6.35X10^{-5}$	*MCC*	mutated in colorectal cancers
cg27337752	7	65955680	$6.64X10^{-5}$	*NA*	NA
cg24748448	17	3493352	$6.65X10^{-5}$	*TRPV1*	transient receptor potential cation channel, subfamily V, member 1
cg05933510	1	9257092	$6.73X10^{-5}$	*NA*	NA
cg26250400	19	15921959	$6.76X10^{-5}$	*NA*	NA
cg06988236	6	30293962	$6.85X10^{-5}$	*NA*	NA
cg20850195	17	62223337	$6.96X10^{-5}$	*NA*	NA

Marker	Chr	Position	P-value	Gene symbol	Gene name
cg02716826	9	33447032	$7.10X10^{-5}$	*NA*	NA
cg05546241	6	29921779	$7.27X10^{-5}$	*NA*	NA
cg16266227	1	202129204	$7.32X10^{-5}$	*PTPN7*	protein tyrosine phosphatase, non-receptor type 7
cg23344992	4	3681759	$7.47X10^{-5}$	*NA*	NA
cg11070191	1	39582205	$7.56X10^{-5}$	*MACF1*	microtubule-actin crosslinking factor 1
cg16620797	3	46741803	$7.58X10^{-5}$	*TMIE*	transmembrane inner ear
cg20275928	17	79283959	$7.66X10^{-5}$	*LINC00482*	long intergenic non-protein coding RNA 482
cg01251710	6	887156	$7.70X10^{-5}$	*NA*	NA
cg05101437	7	92262970	$7.71X10^{-5}$	*CDK6*	cyclin-dependent kinase 6
cg04519918	16	57278807	$7.75X10^{-5}$	*ARL2BP*	ADP-ribosylation factor-like 2 binding protein
cg22393213	2	237478526	$7.87X10^{-5}$	*ACKR3*	atypical chemokine receptor 3
cg23989336	17	79283915	$7.96X10^{-5}$	*LINC00482*	long intergenic non-protein coding RNA 482
cg14824301	11	61515894	$8.17X10^{-5}$	*NA*	NA
cg16853770	12	123355562	$8.41X10^{-5}$	*VPS37B*	vacuolar protein sorting 37 homolog B (S. cerevisiae)
cg04389058	3	57041402	$8.43X10^{-5}$	*ARHGEF3*	Rho guanine nucleotide exchange factor (GEF) 3
cg15585264	19	41770827	$8.48X10^{-5}$	*HNRNPUL1*	heterogeneous nuclear ribonucleoprotein U-like 1
cg11494773	7	48128242	$8.56X10^{-5}$	*UPP1*	uridine phosphorylase 1
cg22111074	3	49690653	$8.61X10^{-5}$	*BSN*	bassoon presynaptic cytomatrix protein
cg00574958	11	68607622	$8.74X10^{-5}$	*CPT1A*	carnitine palmitoyltransferase 1A (liver)
cg10080732	17	1626575	$8.85X10^{-5}$	*WDR81*	WD repeat domain 81
cg02470521	2	176971304	$8.87X10^{-5}$	*HOXD11*	homeobox D11
cg06791867	11	44880944	$9.09X10^{-5}$	*TSPAN18*	tetraspanin 18
cg00611535	12	56367227	$9.14X10^{-5}$	*RAB5B*	RAB5B, member RAS oncogene family
cg17058475	11	68607737	$9.15X10^{-5}$	*CPT1A*	carnitine palmitoyltransferase 1A (liver)
cg02342053	9	37592791	$9.26X10^{-5}$	*TOMM5*	translocase of outer mitochondrial membrane 5 homolog (yeast)
cg16030261	14	104669590	$9.32X10^{-5}$	*NA*	NA
cg02958600	2	181738051	$9.48X10^{-5}$	*NA*	NA

Appendix 2

Copyright Permissions

Appendix 2.1 Permission for reproducing Figure 1

NATURE PUBLISHING GROUP LICENSE TERMS AND CONDITIONS

Mar 25, 2015

This is a License Agreement between Mashael Al-Shafai ("You") and Nature Publishing Group ("Nature Publishing Group") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Nature Publishing Group, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

License Number

3596161453793

License date

Mar 25, 2015

Licensed content publisher

Nature Publishing Group

Licensed content publication

Nature Reviews Genetics

Licensed content title

The genetics of human obesity

Licensed content author

Christopher G. Bell, Andrew J. Walley and Philippe Froguel

Licensed content date

Feb 10, 2005

Volume number

6

Issue number

3

Type of Use

reuse in a dissertation / thesis

Requestor type

academic/educational

Format

print and electronic

Portion

figures/tables/illustrations

Number of figures/tables/illustrations

1

High-res required

no

Figures

FIGURE 1 | Physiological regulation of energy balance.

Author of this NPG article

no

Your reference number

None

Title of your thesis / dissertation

Family-based Investigation of the Genetics and Epigenetics of Obesity in Qatar

Expected completion date

Sep 2015

Estimated size (number of pages)

150

Total

0.00 USD

Terms and Conditions

Terms and Conditions for Permissions

Nature Publishing Group hereby grants you a non-exclusive license to reproduce this material for this purpose, and for no other use,subject to the conditions below:

1. NPG warrants that it has, to the best of its knowledge, the rights to license reuse of this material. However, you should ensure that the material you are requesting is original to Nature Publishing Group and does not carry the copyright of another entity (as credited in the published version). If the credit line on any part of the material you have requested indicates that it was reprinted or adapted by NPG with permission from another source, then you should also seek permission from that source to reuse the material.

2. Permission granted free of charge for material in print is also usually granted for any electronic version of that work, provided that the material is incidental to the work as a whole and that the electronic version is essentially equivalent to, or substitutes for, the print version.Where print permission has been granted for a fee, separate permission must be obtained for any additional, electronic re-use (unless, as in the case of a full paper, this has already been accounted for during your initial request in the calculation of a print run).NB: In all cases, web-based use of full-text articles must be authorized separately through the 'Use on a Web Site' option when requesting permission.

3. Permission granted for a first edition does not apply to second and subsequent editions and for editions in other languages (except for signatories to the STM Permissions Guidelines, or where the first edition permission was granted for free).

4. Nature Publishing Group's permission must be acknowledged next to the figure, table or abstract in print. In electronic form, this acknowledgement must be visible at the same time as the figure/table/abstract, and must be hyperlinked to the journal's homepage.

5. The credit line should read:
   Reprinted by permission from Macmillan Publishers Ltd: [JOURNAL NAME] (reference citation), copyright (year of publication)
   For AOP papers, the credit line should read:
   Reprinted by permission from Macmillan Publishers Ltd: [JOURNAL NAME], advance online publication, day month year (doi: 10.1038/sj.[JOURNAL ACRONYM].XXXXX)

   Note: For republication from the *British Journal of Cancer,* the following credit lines apply.
   Reprinted by permission from Macmillan Publishers Ltd on behalf of Cancer Research UK: [JOURNAL NAME] (reference citation), copyright (year of publication)For AOP papers, the credit line should read:
   Reprinted by permission from Macmillan Publishers Ltd on behalf of Cancer Research UK: [JOURNAL NAME], advance online publication, day month year (doi: 10.1038/sj.[JOURNAL ACRONYM].XXXXX)

6. Adaptations of single figures do not require NPG approval. However, the adaptation should be credited as follows:

Adapted by permission from Macmillan Publishers Ltd: [JOURNAL NAME] (reference citation), copyright (year of publication)

Note: For adaptation from the *British Journal of Cancer,* the following credit line applies.
Adapted by permission from Macmillan Publishers Ltd on behalf of Cancer Research UK: [JOURNAL NAME] (reference citation), copyright (year of publication)

7. Translations of 401 words up to a whole article require NPG approval. Please visithttp://www.macmillanmedicalcommunications.com for more information.Translations of up to a 400 words do not require NPG approval. The translation should be credited as follows:

   Translated by permission from Macmillan Publishers Ltd: [JOURNAL NAME] (reference citation), copyright (year of publication).

   Note: For translation from the *British Journal of Cancer,* the following credit line applies.
   Translated by permission from Macmillan Publishers Ltd on behalf of Cancer Research UK: [JOURNAL NAME] (reference citation), copyright (year of publication)

We are certain that all parties will benefit from this agreement and wish you the best in the use of this material. Thank you.

Special Terms:

v1.1

Questions? customercare@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777.

Gratis licenses (referencing $0 in the Total field) are free. Please retain this printable license for your reference. No payment is required.

Appendix 2.2 Permission request for reproducing Figures 3 and 5

Dear Mashal,

Thank you for your e-mail and apologies that this was not picked up before.
I have requested our inside sales department in a separate e-mail to follow up or re-direct this to the correct person.
I hope you will have an answer in time.

Kind regards,

SihamBasri
Customer Service Representative

Illumina Netherlands B.V.
Freddy van Riemsdijkweg 15
5657 EE Eindhoven

Click here for global contact information.

Sign up for an online account on MyIllumina

-----<Email Received>-----
From: res-te4@qatar-med.cornell.edu
Sent: 5/28/2015 04:26 AM PDT
To: customerservice@illumina.com
CC:
Subject: FW: Permissions for reproducing figures

Dear Siham
I hope this email finds you well.
I have not heard back yet regarding my request in the email below.
I will highly appreciate if you can follow up in this regard especially that my thesis submission date is soon (8th of June), and I need to provide permissions for the figures I reproduced.

Thanks in advance.

Best regards
Mashal

From: Customer Service [customerservice@illumina.com]
Sent: Monday, March 23, 2015 5:20 PM
To: Mashael Al-Shafai; emeainsidesales@illumina.com
Subject: RE: Permissions for reproducing figures

Dear Mashael Al-Shafai,
Thank you for contacting Illumina customer service.
Your query was forwarded to our inside sales department. You will be contacted soon.

Kind regards,

SihamBasri
Customer Service Representative

Illumina Netherlands B.V.
Freddy van Riemsdijkweg 15
5657 EE Eindhoven

Click here for global contact information.

Sign up for an online account on MyIllumina

-----<Email Received>-----
From: res-te4@qatar-med.cornell.edu
Sent: 3/23/2015 07:01 AM PDT
To: customerservice@illumina.com
CC:
Subject: Permissions for reproducing figures

Dear Illumina
I'm a PhD student at Imperial College London. I'm have generated SNP genotyping data, whole genome sequencing and genome wide methylation data using the Illumina service, through a contract with Prof KarstenSuhre from WCMC.
I want to reproduce the tow figures attached in this email for my thesis to Illustrate how the Illumina's SNP genotyping and the methylation profiling technologies work.

I will highly appreciate if you can provide me with written permission for using these two figures if possible.

Best regards

Mashael Al-Shafai

Appendix 2.3 Permission for reproducing Figure 4

Oliver, Brian [Brian.Oliver@ascp.org]

Inbox

13 April 2015 15:43

Mashael,

We can grant permission to you free of any costs as long as you cite us in your thesis. This leads me to my question for you. Do you need our formal permission letter or will this email suffice for your program? Please advise.

Lastly, early congratulations on completing your thesis. If you have to defend it, I hope all goes well.

Best of luck, if I don't hear from you again.

Best,
Brian

~~~~~~

~~~~~~

Brian Oliver, MLIS
Coordinator—Archives, Permissions and Rights
American Society for Clinical Pathology
33 West Monroe, Suite 1600
Chicago, Illinois 60603
Tele: 312-541-4745 Email: brian.oliver@ascp.org

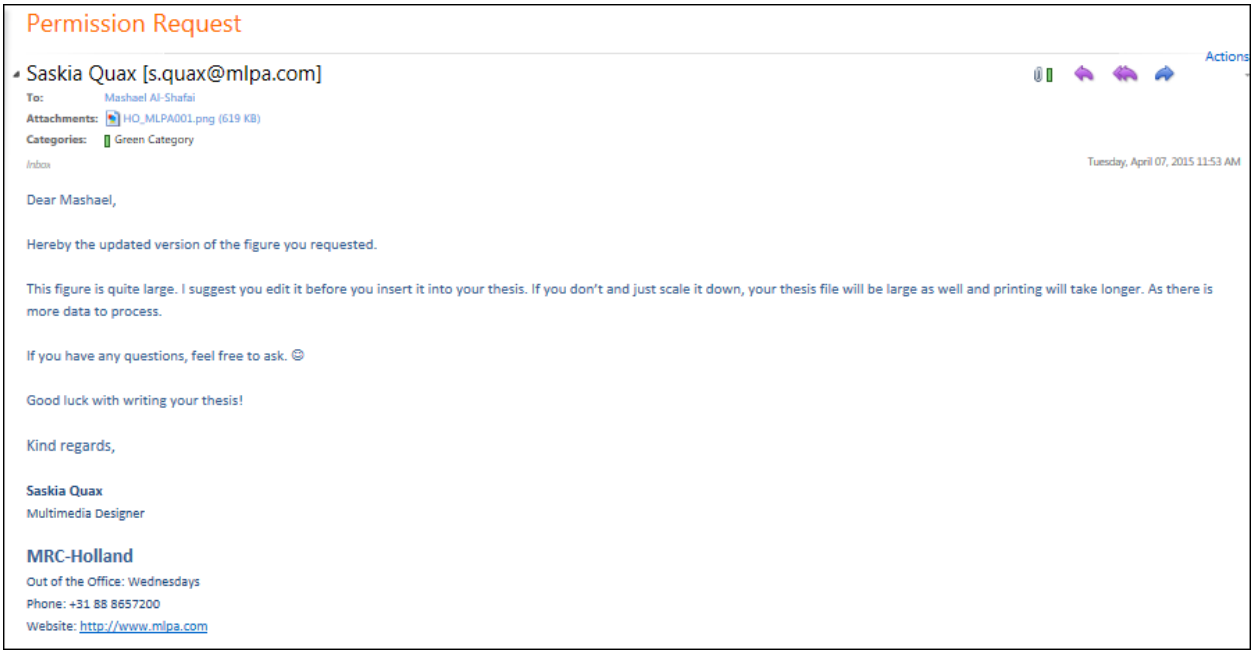
~~~~~~

STRONGERTOGETHER
KNOWLEDGE • ADVANCEMENT • COLLABORATION • GLOBAL COMMUNITY •

~~~~~~

2015 BEYOND THE LAB LONG BEACH 10.28-10.30 ascp.org/ASCP2015

Appendix 2.4 Permission for reproducing Figure 18



Permission Request

Actions

Saskia Quax [s.quax@mlpa.com]

To: Mashael Al-Shafai

Attachments: HO_MLPA001.png (619 KB)

Categories: Green Category

Inbox

Tuesday, April 07, 2015 11:53 AM

Dear Mashael,

Hereby the updated version of the figure you requested.

This figure is quite large. I suggest you edit it before you insert it into your thesis. If you don't and just scale it down, your thesis file will be large as well and printing will take longer. As there is more data to process.

If you have any questions, feel free to ask. ☺

Good luck with writing your thesis!

Kind regards,

Saskia Quax
Multimedia Designer

MRC-Holland
Out of the Office: Wednesdays
Phone: +31 88 8657200
Website: http://www.mlpa.com