Imperial College London

# IDENTIFICATION OF GENOMIC COPY NUMBER VARIATIONS ASSOCIATED WITH OBESITY-RELATED PHENOTYPES

**NIKMAN ADLI BIN NOR HASHIM**

This thesis is submitted for the degree of Doctor of Philosophy (PhD), 2018

Division of Diabetes, Endocrinology & Metabolism, Section of Investigative Medicine

Department of Medicine

Imperial College London

# Declaration of originality

I hereby declare that this thesis is produced by myself, **NIKMAN ADLI BIN NOR HASHIM**, with appropriate citations being referenced. Any work resulted from collaborations and assistance have been appropriately mentioned and acknowledged in the text. A full list of references is provided in the Bibliography section. This thesis is the one upon which I expect to be examined.

**NAME OF AUTHOR: NIKMAN ADLI BIN NOR HASHIM**

**DATE: 12 JUNE 2018**

# 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.

*In the name of Allah The Most Gracious.*
*And by virtue of the teachings of Muhammad s.a.w.*
*May this knowledge be beneficial to humankind.*
*To my religion.*
*To my Malaysians*
*To my family.*
*This PhD is for you, Abah & Mummy.*

*In memory of Teh Adilla bt. Mustaza.*
*May God bless your soul.*
*May you be placed among the highest ranks in heaven*

# Acknowledgements

First and foremost, my utmost gratitude to my supervisor, my 'mother', Professor Alex Blakemore for giving me the opportunity to work for such an extraordinary research group. For all the guidance, encouragements and unending patience to keep me motivated, I am really blessed to have you as my supervisor, thank you. I am very thankful to Dr Robin Walters for all his help in realising this project, in fact, it was through his kind assistance that I managed to do all the major work for this PhD project. I also would like to thank Professor Marjo-Riitta Jarvelin for giving me the opportunity to participate in the NFBC projects. In addition, throughout my 4 years in this PhD journey, I have met countless of brilliant people all around the world, again, it is all thanks to Professor Alex for giving me this wonderful opportunity to work with the ever-vibrant NutriTech people. I would like to extend my gratitude to my colleagues, Ms Hanis Ramzi (and Nas), Mr Andrianos Yiorkas, Ms Olivia Szepietowski, Dr Sanne Alsters, Ms Sumaya Almansoori, Mr Erdal Ozdemir and Ms Mitra Ameri for all the assistance, support and knowledge that have been shared from many discussions since the beginning of this project. My respect and gratitude also go to the Brunel team, Dr Terry Dovey and Dr Fotios Drenos for all the help and guidance given to me from our weekly meetings, particularly in the statistical part of the project. Not to forget, I would like to thank my PhD stage assessors, Professor Gary Frost and Dr Anna Need for their cooperation and valuable ideas on how to improve this project. A special mention goes to my financial sponsor, Majlis Amanah Rakyat, for funding this 4-year PhD journey. Without their support and assistance, I would not be where I am today. Thank you MARA. Also, thank you Vellacotts and team Sri Rahayu.

*Abah & Mummy, I owe you a lifetime debt for all the everlasting love and support, I love you.*

# Abstract

Obesity is a chronic disease with increasing prevalence worldwide. Recent advances in human genetics have led to the discovery of genetic variations underlying obesity, including copy number variants (CNVs). The overall aim of this thesis was to explore the contribution of CNVs to human obesity by (i) producing high quality CNV catalogues from various SNP genotyping datasets; (ii) identifying specific CNVs previously reported to be associated with obesity-related phenotypes (two rare deletion variants in the 16p11.2 region, and multi-allelic CNVs in the amylase gene region on chromosome 1) in 5 study cohorts in order to perform replication studies and extend understanding of their phenotypic effects; and (iii) to identify rare, novel CNVs in morbid obesity. The CNVs were predicted and quality-checked based on the signal intensity data from the SNP genotyping data, which yielded CNVs with consistent characteristics across the study cohorts, and comparable with other studies. Salivary amylase (*AMY1*) CNVs were genotyped by a collaborator using Paralogue Ratio Tests (PRTs) in a small intensively-phenotyped cohort of overweight-to-obese individuals. Statistical analyses indicated no significant associations for *AMY1* CNVs with body mass index, waist-to-hip ratio, body fats and starch intake, further increasing the evidence that *AMY1* CNVs are not associated with obesity, as reported in previous studies. Both 16p11.2 deletion variants (~600-kb and ~200-kb) were confirmed as associated with BMI and WHR in the UK Biobank and Northern Finland Birth Cohort 1966 participants: 600-kb deletion carriers had 12-fold increased risk of obesity in the UK Biobank. Finally, 3 potentially causative CNVs were discovered in 3 different morbidly obese people from the Northern Finland general population; duplications in 17p13 and 22q11.2 regions and a 5p14.3 deletion. The 3 CNVs

spanned regions involving brain function and implicated in obesity. The findings are consistent with previous studies, highlighting the role of rare CNVs in the aetiology of obesity.

.

# Table of Contents

**Declaration of originality** ..... **2**

**Copyright Declaration** ..... **3**

**Acknowledgements** ..... **5**

**Abstract** ..... **6**

**Table of Contents** ..... **9**

**List of Figures** ..... **16**

**List of Tables** ..... **19**

**Abbreviations** ..... **21**

**CHAPTER 1. Introduction** ..... **25**

**1.1. An overview of this thesis** ..... **26**

1.1.1. Overall aim of the thesis ..... 26

1.1.2. Scope of the thesis ..... 26

1.1.3. Structure of the thesis ..... 27

**1.2. A brief introduction to human genetic variation** ..... **31**

1.2.1. Single nucleotide polymorphisms ..... 31

1.2.2. Structural variations ..... 32

**1.3. Copy number variations** ..... **35**

1.3.1. Emergence of research into genomic copy number variations ..... 35

1.3.2. The origin of CNVs in the human genome ..... 36

1.3.3. The importance of CNVs in human health ..... 41

**1.4. Obesity as a chronic disease** ..... **44**

1.4.1. The definition of obesity ..... 44

1.4.2. Phenotyping of obesity ..... 45

1.4.3. Implications of obesity for economy and health ..... 47

1.4.4. Increasing prevalence in obesity ..... 48

1.4.5. Morbidity and mortality of obesity ..... 49

**1.5. The genetics of obesity** ..... **50**

1.5.1. Heritability of obesity ..... 50

1.5.2. The genetic causes of obesity ..... 52

1.5.3. Leptin deficiency in monogenic obesity ..... 54

1.5.4. More genes implicated in monogenic obesity ..... 55

1.5.5. Chromosomal abnormalities in syndromic obesity ..... 57
1.5.6. Linkage analysis in common obesity ..... 59
1.5.7. GWAS in common obesity ..... 60
1.5.8. Missing heritability ..... 62
**CHAPTER 2. Study Cohorts ..... 64**
**2.1. Northern Finland Birth Cohort 1966 ..... 65**
2.1.1. Overview ..... 65
2.1.2. Data and sample collections ..... 68
**2.2. Northern Finland Birth Cohort 1986 ..... 71**
2.2.1. Overview ..... 71
2.2.2. Data and sample collections ..... 73
**2.3. NutriTech cohort ..... 76**
2.3.1. Overview ..... 76
2.3.2. Data and sample collections ..... 78
**2.4. Personalised Medicine of Morbid Obesity cohort ..... 80**
2.4.1. Overview ..... 80
2.4.2. Data and sample collections ..... 81
**2.5. United Kingdom Biobank ..... 84**
2.5.1. Overview ..... 84
2.5.2. Data and sample collections ..... 85
**CHAPTER 3. General Methods ..... 87**
**3.1. Introduction ..... 88**
**3.2. Biological sample collection ..... 89**
3.2.1. NFBC1966 and NFBC1986 cohorts ..... 89
3.2.2. NutriTech cohort ..... 90
3.2.3. PMMO cohort ..... 90
3.2.4. UK Biobank cohort ..... 91
**3.3. Genomic DNA extraction ..... 92**
3.3.1. NFBC19866 and NFBC1986 cohorts ..... 92
3.3.2. NutriTech cohort ..... 92
3.3.3. PMMO cohort ..... 93
3.3.4. UK Biobank cohort ..... 93
**3.4. Obesity-related phenotypic measurements ..... 94**
3.4.1. Body mass index and waist-to-hip ratio ..... 94

3.4.2. Adiposity ..... 97
**3.5. SNP genotyping ..... 98**
3.5.1. Arrays used in different cohorts ..... 98
3.5.2. Processing of genotyped data ..... 101
3.5.3. Access to the genotyped data of UK Biobank cohort ..... 105
**3.6. Quality controls ..... 105**
3.6.1. Quality control 1: Sex discordance ..... 106
3.6.2. Quality control 2: Missing genotype data, high autosomal heterozygosity rates ..... 108
3.6.3. Quality control 3: Sample call rates ..... 109
3.6.4. Sample exclusions ..... 109
3.6.5. Rare variant calling ..... 111
3.6.6. Quality controls in the UK Biobank ..... 114
**3.7. Population structure adjustments ..... 116**
3.7.1. A, T, C, G allele conversion ..... 116
3.7.2. Adjustment 1: Inbreeding ..... 117
3.7.3. Adjustment 2: Pair-wise Identity-by-descent (IBD) Estimation ..... 119
3.7.4. Adjustment 3: Multidimensional Scaling (MDS) ..... 121
3.7.5. Sample exclusions ..... 124
3.7.6. Adjustments in the UK Biobank ..... 125
**3.8. Principal component analysis (PCA) ..... 126**
**3.9. CNV calling ..... 128**
3.9.1. Split signal information data ..... 129
3.9.2. Create cohort-specific population frequency of B allele ..... 130
3.9.3. Calling of CNVs ..... 130
**3.10. CNV quality controls ..... 131**
3.10.1. Filter 1: QC from SNP genotyping ..... 131
3.10.2. Filter 2: QC from CNV calling ..... 131
3.10.3. Filter 3: low-quality and low-confidence CNVs ..... 133
3.10.4. Filter 4: spurious CNV regions ..... 133
3.10.5. Filter 5: merge CNVs ..... 134
**3.11. Other methods ..... 134**
**3.12. Methods overview in each Result Chapter ..... 135**
**CHAPTER 4. Genetic Data Processing ..... 136**
**4.1. Introduction ..... 137**

**4.2. Aim of this chapter ..... 139**
**4.3. General methods used in this chapter ..... 141**
**4.4. Results: Sample Quality Controls..... 142**
4.4.1. NFBC1966 cohort ..... 142
4.4.2. NFBC1986 cohort ..... 147
4.4.3. NutriTech cohort ..... 152
4.4.4. PMMO cohort ..... 156
4.4.5. UK Biobank cohort ..... 160
**4.5. Results: Population structure adjustments ..... 161**
4.5.1. NFBC1966 cohort ..... 161
4.5.2. NFBC1986 cohort ..... 166
4.5.3. Merged NFBC1986 cohort ..... 171
4.5.4. NutriTech cohort ..... 175
4.5.5. PMMO cohort ..... 179
4.5.6. Merged PMMO cohort ..... 183
4.5.7. UK Biobank cohort ..... 187
**4.6. Discussions ..... 188**
4.6.1. Results overview ..... 188
4.6.2. The importance of doing quality controls on SNP genotyping data ..... 189
4.6.3. Resolving admixture for population homogeneity ..... 191
**CHAPTER 5. CNV Calling ..... 193**
**5.1. Introduction ..... 194**
**5.2. Aim of this chapter ..... 197**
**5.3. General methods used in this chapter ..... 199**
**5.4. Specific methods..... 200**
5.4.1. Statistical analyses ..... 200
**5.5. Results..... 201**
5.5.1. NFBC1966 cohort ..... 201
5.5.2. NFBC1986 cohort ..... 206
5.5.3. NutriTech cohort ..... 215
5.5.4. PMMO cohort ..... 220
5.5.5. General characteristics of CNVs across study cohorts ..... 224
**5.6. Discussions ..... 228**
5.6.1. Results overview ..... 228

5.6.2. Comparison of CNV characteristics with other studies ..... 229
5.6.3. PennCNV tool as a preferred CNV calling tool ..... 231
5.6.4. Hidden Markov Model to predict copy number changes ..... 232
**CHAPTER 6. Multi-Allelic CNVs in Human Amylase Gene Region ..... 234**
**6.1. Introduction ..... 235**
**6.2. Aim of this chapter ..... 241**
**6.3. General methods used in this chapter ..... 243**
**6.4. Specific methods ..... 244**
6.4.1. Signal intensity correlation with BMI Using SNPs in Human Amylase Region ..... 244
6.4.2. CNV genotyping using Multiplex Paralogue Ratio Tests ..... 245
6.4.3. Statistical analyses ..... 246
**6.5. Results ..... 248**
6.5.1. Quality controls & population structure adjustments ..... 248
6.5.2. Distributions of phenotypes and amylase copy numbers ..... 249
6.5.3. Sex, age and group effects ..... 253
6.5.4. Association analysis of AMY1 copy numbers with obesity-related phenotypes, adiposity contents and starch intakes ..... 255
6.5.5. Assessing the accuracy of PRT-based *AMY1* copy numbers ..... 262
6.5.6. Association analysis between signal intensity data and BMI ..... 262
6.5.7. *AMY1* and *AMY2A* copy number correlations ..... 264
**6.6. Discussions ..... 267**
6.6.1. Attempted replication of previously reported association of BMI ..... 267
6.6.2. CNV Genotyping of amylase gene region ..... 270
6.6.3. Looking at starch consumption ..... 272
**CHAPTER 7. Rare Deletions in The Chromosome 16p11.2 Region ..... 274**
**7.1. Introduction ..... 275**
**7.2. Aim of this chapter ..... 281**
**7.3. General methods used in this chapter ..... 283**
**7.4. Specific methods ..... 284**
7.4.1. Identification of 16p11.2 deletions from the NFBC CNV catalogues ..... 284
7.4.2. Run of homozygosity (ROH) scans on putative 16p11.2 copy number regions ..... 284
7.4.3. Selecting potential 16p11.2 deletion carriers from ROH scans ..... 285
7.4.4. Signal intensity plots for the 16p11.2 region ..... 285
7.4.5. Visual inspections of signal intensity plots of potential 16p11.2 deletion carriers ..... 286

7.4.6. Statistical analyses ..... 287
**7.5. Results ..... 290**
7.5.1. Distributions of obesity-related traits ..... 290
7.5.2. Sex, age and batch effects ..... 293
7.5.3. Obesity-related phenotype correlation analysis ..... 297
7.5.4. Confirmation of proximal 16p11.2 deletion carriers (~600 kb) ..... 298
7.5.5. Confirmation of distal 16p11.2 deletion carriers (~200 kb) ..... 302
7.5.6. Association analyses of proximal 16p11.2 deletion on obesity-related traits ..... 307
7.5.7. Association analyses of distal 16p11.2 deletion on obesity-related phenotypes ..... 312
7.5.8. Risk of obesity in the general population ..... 315
7.5.9. Risk of obesity in a pooled case-control association analysis of PMMO and UKBB ..... 317
**7.6. Discussions ..... 322**
7.6.1. Results overview ..... 322
7.6.2. The frequency of 16p11.2 deletion variants ..... 324
7.6.3. Cognitive defects as a comorbidity of obesity ..... 326
7.6.4. Structural instability in 16p11.2 region and obesity ..... 328
**CHAPTER 8. Discovery of Rare CNVs in Morbid Obesity ..... 331**
**8.1. Introduction ..... 332**
**8.2. Aims of this chapter ..... 334**
**8.3. General methods used in this chapter ..... 336**
**8.4. Specific methods ..... 337**
8.4.1. Exclusion of small CNV calls from the CNV catalogue ..... 337
8.4.2. CNV list and file conversions ..... 337
8.4.3. Permutation-based CNV association tests ..... 338
8.4.4. Visual inspections of signal intensity plots ..... 339
8.4.5. Statistical analyses ..... 340
8.4.6. Public online database search ..... 340
**8.5. Results ..... 341**
8.5.1. Study cohort overview ..... 341
8.5.2. Distributions of obesity-related traits ..... 341
8.5.3. Sex and batch effects ..... 343
8.5.4. Principal components as covariates ..... 344
8.5.5. Obesity-related phenotype correlation analysis ..... 344
8.5.6. Discovery and validation of suggestive top hits ..... 347
8.5.7. Public online database search ..... 353

**8.6. Discussions ..... 358**

8.6.1. Results overview ..... 358

8.6.2. The 17p13.3 CNVs, cognitive health and obesity ..... 359

8.6.3. The 22q11.2 CNVs, cognitive health and obesity ..... 361

8.6.4. The 5p14.3 CNVs, cognitive health and obesity ..... 364

**CHAPTER 9. Conclusions and Future Work ..... 367**

**9.1. Summary of findings ..... 368**

9.1.1. Quality SNP genotyping data is prerequisite for CNV analysis ..... 368

9.1.2. Cataloguing high-quality CNV calls ..... 370

9.1.3. No association of *AMY1* copy number with obesity-related phenotypes ..... 372

9.1.4. Proximal and distal 16p11.2 deletions are associated with obesity-related phenotypes 374

9.1.5. Discovery of rare CNVs in the 5p14.3, 17p13 and 22q11.2 regions ..... 376

9.1.6. Neurobehavioural connection to obesity ..... 379

**9.2. Limitations & Future Work ..... 381**

9.2.1. Investigation of salivary and pancreatic amylase CNVs ..... 381

9.2.2. Investigation of the proximal and distal 16p11.2 duplications ..... 383

9.2.3. Interrogation of the rare, novel CNVs in other populations ..... 385

**Bibliography ..... 387**

**Appendices ..... 410**

# List of Figures


| Figure number | Page number |
|---|---|
| **Chapter 1** | |
| Figure 1- 1 | 34 |
| Figure 1- 2 | 39 |
| Figure 1- 3 | 40 |
| Figure 1- 4 | 53 |
| **Chapter 2** | |
| Figure 2- 1 | 66 |
| Figure 2- 2 | 82 |
| **Chapter 3** | |
| Figure 3- 1 | 102 |
| **Chapter 4** | |
| Figure 4 -1 | 144 |
| Figure 4- 2 | 146 |
| Figure 4- 3 | 148 |
| Figure 4- 4 | 150 |
| Figure 4- 5 | 153 |
| Figure 4- 6 | 155 |
| Figure 4- 7 | 157 |
| Figure 4- 8 | 159 |
| Figure 4- 9 | 163 |
| Figure 4- 10 | 165 |
| Figure 4- 11 | 167 |
| Figure 4- 12 | 169 |
| Figure 4- 13 | 170 |
| Figure 4- 14 | 172 |
| Figure 4- 15 | 174 |

| Figure number | Page number |
|---|---|
| Figure 4- 16 | 176 |
| Figure 4- 17 | 178 |
| Figure 4- 18 | 180 |
| Figure 4- 19 | 182 |
| Figure 4- 20 | 184 |
| Figure 4- 21 | 186 |
| **Chapter 5** | |
| Figure 5- 1 | 202 |
| Figure 5- 2 | 205 |
| Figure 5- 3 | 208 |
| Figure 5- 4 | 209 |
| Figure 5- 5 | 212 |
| Figure 5- 6 | 213 |
| Figure 5- 7 | 214 |
| Figure 5- 8 | 216 |
| Figure 5- 9 | 219 |
| Figure 5- 10 | 221 |
| Figure 5- 11 | 223 |
| **Chapter 6** | |
| Figure 6- 1 | 238 |
| Figure 6- 2 | 250 |
| Figure 6- 3 | 251 |
| Figure 6- 4 | 252 |
| Figure 6- 5 | 260 |
| Figure 6- 6 | 261 |
| Figure 6- 7 | 263 |
| Figure 6- 8 | 265 |
| Figure 6- 9 | 266 |
| **Chapter 7** | |
| Figure 7- 1 | 277 |

| Figure number | Page number |
|---|---|
| Figure 7- 2 | 291 |
| Figure 7- 3 | 292 |
| Figure 7- 4 | 295 |
| Figure 7- 5 | 299 |
| Figure 7- 6 | 300 |
| Figure 7- 7 | 301 |
| Figure 7- 8 | 303 |
| Figure 7- 9 | 305 |
| Figure 7- 10 | 306 |
| **Chapter 8** | |
| Figure 8- 1 | 342 |
| Figure 8- 2 | 345 |
| Figure 8- 3 | 346 |
| Figure 8- 4 | 349 |
| Figure 8- 5 | 350 |
| Figure 8- 6 | 352 |
| Figure 8- 7 | 353 |
| Figure 8- 8 | 355 |
| Figure 8- 9 | 357 |
| **Chapter 9** | |
| | |

# List of Tables


| Table number | Page number |
|---|---|
| **Chapter 1** | |
| Table 1- 1 | 45 |
| **Chapter 2** | |
| Table 2- 1 | 67 |
| Table 2- 2 | 69 |
| Table 2- 3 | 70 |
| Table 2- 4 | 72 |
| Table 2- 5 | 74 |
| Table 2- 6 | 75 |
| Table 2- 7 | 77 |
| Table 2- 8 | 79 |
| Table 2- 9 | 83 |
| Table 2- 10 | 86 |
| **Chapter 3** | |
| | |
| **Chapter 4** | |
| Table 4- 1 | 151 |
| **Chapter 5** | |
| Table 5- 1 | 203 |
| Table 5- 2 | 210 |
| Table 5- 3 | 217 |
| Table 5- 4 | 222 |
| Table 5- 5 | 226 |
| Table 5- 6 | 227 |
| **Chapter 6** | |
| Table 6- 2 | 256 |

| Table number | Page number |
|---|---|
| Table 6- 3 | 257 |
| **Chapter 7** | |
| Table 7- 1 | 308 |
| Table 7- 2 | 311 |
| Table 7- 3 | 313 |
| Table 7- 4 | 316 |
| Table 7- 5 | 318-319 |
| Chapter 8 | |
| Table 8- 1 | 348 |
| Chapter 9 | |
| | |

# Abbreviations

| | |
|---|---|
| 1000G | The 1000 Genomes Project |
| aCGH | array comparative genomic hybridisation |
| ADHD | attention deficit-hyperactivity disorder |
| ANOVA | Analysis of Variance |
| ASD | autism spectrum disorder |
| ASW | African ancestry in Southwest USA |
| AT | adipose tissue |
| BAF | B Allele Frequency |
| BMI | body mass index |
| CEO | Chief Executive Officer |
| CEU | Utah residents with Northern and Western European ancestry from the CEPH collection |
| CHB | Han Chinese in Beijing, China |
| CHD | congenital heart defects |
| CHD | Chinese in Metropolitan Denver, Colorado |
| CHIAG | Community Health Index Advisory Group |
| CI | confidence interval |
| CMT1A | Charcot-Marie-Tooth syndrome type 1A |
| CNS | central nervous system |
| CNV | copy number variant |
| CNVR | copy number variable region |
| DALY | disability-adjusted life year |
| DCR | Down's syndrome critical region |
| DD | developmental delay |
| DECIPHER | DatabasE of genomiC varIation and Phenotype in Humans using Ensembl Resources |

| | |
|---|---|
| DGV | TCAG Database of Genomic Variants |
| DNA | deoxyribonucleic acid |
| dPCR | digital PCR |
| DS | Down's syndrome |
| DSB | double-stranded breaks |
| DSD | disorders of sexual developments |
| DZ | dizygotic twins |
| EDTA | Ethylenediaminetetraacetic acid |
| EMA | European Medicines Agency |
| ERV | endogeneous retrovirus |
| FISH | fluorescence in-situ hybridisation |
| FoSTeS | forkhead stalling and template switching |
| gDNA | genomic DNA |
| GIANT | The Genetic Investigation of Anthropometric Traits |
| GIH | Gujarati Indians in Houston, Texas |
| GRS | genetic risk scores |
| GWAS | genome-wide association study |
| HapMap | Haplotype map |
| hg19 | Genome Reference Consortium Human Reference 37 genome build |
| HMM | Hidden Markov Model |
| HWE | Hardy-Weinberg Equilibrium |
| IBD | identity-by-descent |
| ID | intellectual disability |
| indels | insertions/deletions polymorphisms |
| JPT | Japanese in Tokyo, Japan |
| KS | Klinefelter syndrome |
| LA | linkage analysis |
| LCR | low copy repeats |
| LD | linkage disequilibrium |
| LINE | long interspersed nuclear elements |

| | |
|---|---|
| LRR | Log R Ratio |
| LWK | Luhya in Webuye, Kenya |
| MAF | minor allele frequency |
| MAGIC | The Meta-Analyses of Glucose and Insulin-related traits Consortium |
| MDS | Multidimensional Scaling |
| MEX | Mexican ancestry in Los Angeles, California |
| MKK | Maasai in Kinyawa, Kenya |
| MMBIR | microhomology-mediated break-induced replication |
| MRC | Medical Research Council |
| MREC | North West Multi-centre Research Ethics Committee |
| MRI | magnetic resonance imaging |
| MZ | monozygotic twins |
| NAHR | non-allelic homologous recombination |
| NFBC1966 | Northern Finland Birth Cohort 1966 |
| NFBC1986 | Northern Finland Birth Cohort 1986 |
| NHEJ | non-homologous end joining |
| NHS | National Health Service |
| NutriTech | The Application of New Technologies and Tools to Nutrition Research project |
| OMIM | Online Mendelian Inheritence in Man |
| OR | odds ratio |
| PC | principal component |
| PCA | Principal Component Analysis |
| PHE | Public Health England |
| PLS | Potocki-Lupski syndrome |
| PMMO | Personalised Medicine for Morbid Obesity |
| PRS | polygenic risk scores |
| PRT | paralogue ratio test |
| PWS | Prader-Willi syndrome |
| QC | quality control |

| | |
|---|---|
| qPCR | quantitative Polymerase Chain Reaction |
| QTL | quantitative trait loci |
| RA | rheumatoid arthritis |
| RNA | ribonucleic acid |
| ROH | runs of homozygosity |
| RR | R Ratio |
| SD | standard deviation |
| SE | standard error |
| SES | socio-economic status |
| SHFLD3 | split-hand/foot malformation and the long bone deficiency-3 |
| SINE | short interspersed nuclear elements |
| SMS | Smith-Magenis syndrome |
| snoRNA | small nucleolar RNA |
| SNP | single nucleotide polymorphism |
| SV | structural variation |
| SZ | schizophrenia |
| T2DM | type 2 diabetes mellitus |
| TE | Tris-EDTA |
| TS | Turner's syndrome |
| TSI | Toscani in Italy |
| UK BiLEVE | UK Biobank Lung Exome. Variant Evaluation |
| UKBB | United Kingdom Biobank |
| WC | waist circumference |
| WHO | World Health Organisation |
| WHR | waist-to-hip ratio |
| YRI | Yoruba in Ibadan, Nigeria |

# CHAPTER 1. Introduction

# 1.1. An overview of this thesis

## 1.1.1. Overall aim of the thesis

The overall aim of the work described in this thesis was to explore the role of genomic copy number variants (CNVs) in obesity and obesity-related phenotypes, such as body mass index (BMI), waist-to-hip ratio (WHR), and various measures of body fat content.

## 1.1.2. Scope of the thesis

In order to realise the above aim, I took advantage of using the signal intensity information from single nucleotide polymorphism (SNP) genotyping data of several study cohorts that were provided by my supervisor, Professor Alex Blakemore, *via* multiple collaborative efforts. The research was carried out by implementing a series of bioinformatics methods based on the signal intensity information, (with an exception for the investigation of the multi-allelic CNVs in **Chapter 6** in which laboratory-based methods were involved) to identify CNVs. This was followed by statistical analyses to explore their phenotypic implications.

The scope of this thesis includes, first, the processing of raw SNP genotyping data involving a series of quality control (QC) checks to generate a clean genetic data from each study cohort. This was followed by the prediction of CNVs by using a commonly used CNV calling algorithm, the PennCNV tool, and the validation of CNVs by visually inspecting individual signal intensity plots generated using perl scripts, developed and kindly provided by Dr Robin Walters (University of Oxford). This thesis also considers the use of a Paralogue Ratio Test (PRT) method to genotype a specific CNVR harbouring a highly multi-allelic CNV. The laboratory analysis was carried out by John Armour's research group (University of

Nottingham), so the development and application of the specific PRT method was not within the scope of this thesis. Upon validation of the presence of CNVs under study, a number of statistical tests were applied to assess the associative relationships of the CNVs with obesity-related quantitative phenotypes (i.e. BMI, WHR and adiposity) as well as their implications for risk of obesity (considered as a binary trait). In addition, a small part of the work involved the use of public databases to search for relevant information with regards to the CNVs under study. However, functional and molecular analyses do not fall into the scope of this thesis. Here, I provide statistical evidence to support the previously-reported known CNV relationships with obesity-related phenotypes and obesity. Additionally, this thesis also identifies large, rare CNVs likely contributing to obesity in individuals from the general population (using two Northern Finnish Birth Cohorts).

### 1.1.3. Structure of the thesis

**Chapter 1** provides the overview of the thesis, an introduction to genetic variation as applied to the investigation of human diseases, background information on obesity as a global health issue, the development of understanding of the genetics of obesity, recognition of the increasing importance of copy number variations in the human genome in that context, as well as the relevance of this thesis in addressing the contribution of CNVs to obesity and related phenotypes.

**Chapter 2** offers a detailed description of each study cohort that was involved in this thesis. This includes two birth cohorts representing the Northern Finland adult and adolescent populations, a small dietary intervention cohort (NutriTech) consisting of overweight and obese individuals with European ancestry, a pilot bariatric surgery cohort (PMMO) consisting

of obese and morbidly obese patients with European ancestry, and a large general population cohort containing older adults who currently reside in the UK (the UK Biobank).

**Chapter 3** describes all the general methods that were involved in this thesis. Some methods that were specific for a particular study, were not included in **Chapter 3**. Those specific methods were described in detail at the beginning of the relevant result chapters. This chapter also includes the command lines used in the computational work.

**Chapter 4** is the first results chapter of this thesis. In this chapter, I provide the results of generating a clean genetic data for each study cohort, so that it can be used in downstream analyses. All study cohorts were involved in this chapter, except for the UK Biobank (UKBB) cohort (quality control of the UKBB cohort was performed by other UKBB collaborators, as I describe briefly in **Chapter 3**). The work described encompasses the genotyping QC checks to remove samples with poor genotyping quality, suspected DNA contamination, sex discordance and extreme heterozygosity rates. After removing these samples, each study cohort underwent a process of adjusting for the population structure, which involved removing the samples with evidence of inbreeding, samples with cryptic relatedness and population outliers. At the end of the work described in this chapter, I had provided a cleaned genetic dataset for each study cohort, ready for detection of CNVs by analysis of signal intensity.

**Chapter 5** is the second results chapter of this thesis. In it, I describe the results of the CNV calling process for each study cohort (except for the UKBB cohort). In this chapter, I used one algorithm, the PennCNV tool, to call CNVs in every individual within a study cohort. A series of QC procedures, based on the CNV quality parameters were then applied, generating a clean CNV catalogue for each study cohort involved in this chapter. These CNV catalogues were used as the first step to identify the presence of any CNVs of interest in subsequent result

chapters. In addition, the CNV catalogues were also used as proxy-CNV lists to identify further potential CNVs that are associated with BMI, culminating in the work in **Chapter 8**.

**Chapter 6** is the third results chapter of this thesis, which describes an analysis of CNVs in the salivary amylase (*AMY1*) gene region with obesity-related phenotypes, such as BMI, waist circumference (WC), waist-to-hip ratio (WHR) and various measures of fat content. The study cohort involved in this chapter was a dietary intervention cohort (NutriTech), which although small, had undergone extremely well-controlled and detailed phenotyping. I was interested in investigating the relationship between the *AMY1* CNVs and BMI because there have been conflicting reports regarding the association findings from previous studies(Carpenter *et al.*, 2015, Falchi *et al.*, 2014, Mejia-Benitez *et al.*, 2015, Perry *et al.*, 2007, Usher *et al.*, 2015, Viljakainen *et al.*, 2015). Due to the availability of data and the relevance of starch intake in the context of salivary amylase function(Rukh *et al.*, 2017), the analysis was extended to assess the relationship between *AMY1* CNVs and dietary starch intake. This chapter also explores the possibility of the SNPs spanning the human amylase gene region to be associated with BMI. In addition, this chapter describes assessment of the differences in the even-to-odd number of *AMY1* gene copies. The findings were then compared to the findings reported by previous researchers. This is the only chapter in this thesis that involved direct laboratory-based methods (i.e. use of PRT to genotype the highly polymorphic *AMY1* region CNVs).

**Chapter 7 i**s the fourth results chapter of this thesis. In this chapter, I describe the identification of proximal (~600 kb) and distal (~200 kb) deletion variants of the chr16p11.2 region in an adult birth cohort (NFBC1966), a pilot obesity patient cohort (PMMO) and a very large general population cohort (UKBB). Both deletion variants were of interest in this thesis because they have been implicated in causing Mendelian obesity in previous studies(Bochukova *et al.*, 2010, Walters *et al.*, 2010). The preliminary identification of each

Chr16p11.2 deletion was confirmed by visual inspections of the signal intensity plots, and then by analyses of each 16p11.2 deletion variant in relation to obesity-related phenotypes (i.e. BMI and WHR). In addition, this chapter also presents the implications of these deletions for the risk of obesity and morbid obesity, within each study cohort as well as in an obesity case/control design.

**Chapter 8** is the final results chapter of this thesis. This chapter describes the use of the CNV catalogues generated in **Chapter 5** as proxy-CNV lists for permutation-based BMI-CNV association tests. The study cohorts involved in this chapter were the two Northern Finland birth cohorts; NFBC1966 (adults) and NFBC1986 (adolescents). The intention was to discover any potential CNVs (genome-wide) with large effects on BMI and on morbid obesity – rare CNVs and singletons were included to increase the possibility of discovering large causative CNVs in isolated cases. As before, the discovery stage was followed by a validation stage using signal intensity plots to confirm the presence of the CNVs. Depending on the number of CNV carriers, statistical association analyses were subsequently carried out on the obesity-related phenotypes (i.e. BMI and WHR). This chapter also explores public online databases to investigate whether the CNVs found in this chapter had been reported previously, including their reported effects on BMI and obesity.

**Chapter 9** summarises the main findings from each result chapter in this thesis. On top of that, it also describes the limitations of the result chapters, and proposes several ways that this research might be taken forward to extend the work presented in this thesis.

## 1.2. A brief introduction to human genetic variation

### 1.2.1. Single nucleotide polymorphisms

The human genome contains various types of variation that contribute to inter-individual differences in phenotype. By investigating variations in the genetics of different individuals, researchers can identify which result in the observed phenotypic variability, influence individual responses to various exposures, or confer susceptibility to diseases, such as obesity or cancer. Studies of human genetic variation are important in identifying susceptibility loci and candidate genes causing diseases. Thirty-seven years after the discovery of the DNA double-helix structure (Watson and Crick, 1953), a transglobal collaborative project was undertaken to sequence the whole human genome, called The Human Genome Project (International Human Genome Sequencing Consortium 2004). Completion of this worldwide-scale project in 2003, with publication of the draft sequence and mapping of the human genome (International Human Genome Sequencing, 2004), paved the way for comprehensive cataloguing of "normal" human genetic variation, enabling sequence variants in disease-affected individuals to be compared to those of 'healthy' individuals.

Common single-nucleotide changes in the human genome are called single nucleotide polymorphisms (SNPs). This type of variation was first reported by Kan and Dozy (1978) when they described DNA sequence polymorphisms near to the human beta-globin gene (Kan and Dozy, 1978). Since then, many more SNPS have been reported across different populations. In parallel with The Human Genome Project, another global-scale project called The HapMap Project was initiated to map the common patterns of genetic variations in the human genome, leading to the dramatic increase in SNP discoveries (International HapMap, 2003)

and mapping of patterns of linkage disequilibrium between them. The 1000 Genomes Project (1000G) is a natural extension of The HapMap project, whereby 1,000 individuals were sequenced to provide a more comprehensive, global reference of common variations in the human genome (Genomes Project *et al.*, 2015). To date, approximately 84 million SNPs have been described by the 1000G in 2,504 individuals from 26 different populations (Genomes Project *et al.*, 2015). The culmination of these projects has been to enable genome-wide association studies (GWAS) to extensively utilise the SNPs in the search for genomic loci that are associated with diseases, disorders and other health-related phenotypes.

### 1.2.2. Structural variations

Other than SNPs, the 1000G project also reported 60,000 structural variants and 3.6 million short insertions/deletions polymorphisms (indels) in the human genome (Genomes Project *et al.*, 2015). The initial assumption was that SNPs were responsible for the greatest part of the variation observed in the genomic compositions of different individuals, and any two unrelated healthy individuals were thought to be genetically ~99% identical (Przeworski *et al.*, 2000). However, the discoveries of other types of genetic variations by the 1000G have highlighted the contribution of structural variations and indels to inter-individual diversity in the human genome (Genomes Project *et al.*, 2015). Structural variations (SVs) are genomic alterations involving a stretch of DNA sequence that affect the structural composition of the genome relative to the human reference genome – including copy number variants (CNVs), indels, translocations, inversions (Feuk *et al.*, 2006). **Figure 1- 1** shows examples of the different types of SVs. Indels are usually defined as insertions or deletions of a DNA sequence with sizes < 1kb, while CNVs are generally defined as gains and losses of a DNA sequence with sizes > 1kb. Indels and CNVs affect the structural integrity of the genome by changing the copy

number state of a particular locus. In contrast, translocations and inversions are SVs that change the position and orientation of the affected locus instead of changing the copy number.

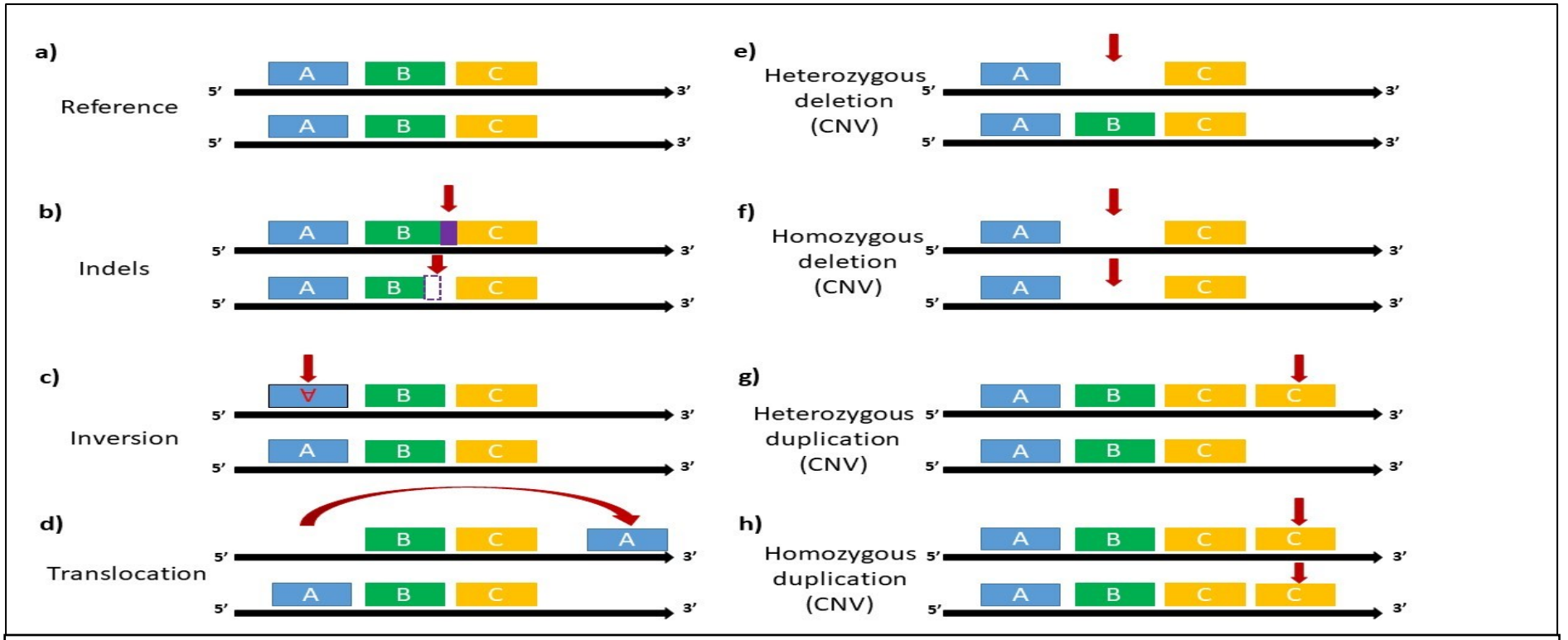


**Figure 1- 1:** A visualisation of the different types of structural variations (SVs) in the human genome; **a)** reference 'normal' **b)** short indels **c)** inversion **d)** translocation **e)** CNV – heterozygous deletion (1-copy loss) **f)** CNV – homozygous deletion (2-copy loss) **g)** CNV – 1-copy duplication **h)** CNV – 2-copy duplication. The blue, green and orange boxes represent different gene on two homologous chromosomes in the 5' to 3' direction. Red arrows show the genes affected by the SVs. The full-purple box represents a short insertion, while the dotted purple box represents a short deletion. CNV = copy number variant.

# 1.3. Copy number variations

## 1.3.1. Emergence of research into genomic copy number variations

Recent rapid advances in CNV research provide a platform to assess the potential implications of CNVs on human phenotypes. It had previously been assumed that every healthy human had two copies (paternally- and maternally-inherited) of every autosomal gene. Large genomic variants outside of heterochromatic regions (regions of tightly-packed DNA) were assumed to be pathogenic and the extent of common structural variation involving functional genes was unsuspected. However, with the arrival of high-throughput genotyping technology, advanced array comparative genomic hybridisation (aCGH) platforms and most recently, new generation sequencing methods, more CNVs have been discovered. These began to be systematically catalogued and listed in the TCAG Database of Genomic Variants (DGV) in 2007. To date, over 500,000 CNV regions have been compiled in DGV from ~72 studies (updated May 2016). These CNV regions mainly consist of heterozygous and homozygous deletions (copy number 1 and 0), as well as 1-copy and 2-copy duplications (copy number 3 and 4). In addition, there are also highly polymorphic CNV regions with more than 4 genomic copies, such as the human amylase gene cluster region on chromosome 1 (**Chapter 6**). As with sequence variation, including SNPs, the international effort to identify and catalogue benign structural variants in "healthy" controls allows more evidence-based assessment of the potential pathogenicity of CNVs found in individuals with a range of disorders, including severe or complex obesity (de Smith *et al.*, 2007).

### 1.3.2. The origin of CNVs in the human genome

The underlying mechanisms of CNV formation are DNA repair, replication or recombination, and the process is highly dependent on the presence of segmental duplications or low copy repeats (LCRs). These are blocks of repetitive DNA sequences, 1 kb to ~400 kb in length, that have high sequence homology with another genomic region (> 95% sequence identity) (Bailey *et al.*, 2002). LCRs which cover ~5% of the human genome (Bailey *et al.*, 2002), are thought to form by the combination of tandem DNA repetition events and random dispersed rearrangements to other genomic regions (Eichler, 2001). These repeats are correlated with CNVs across the whole genome (Cooper *et al.*, 2007, Redon *et al.*, 2006), while they have been associated with regions with high genomic instability (Bailey *et al.*, 2004). In addition, small deletions have also been reported to occur in between stable breakpoints that were enriched with *Alu* repeat elements instead of LCRs (de Smith *et al.*, 2007), providing evidence that CNVs' breakpoints are not limited to LCRs. The relationship of these repeat elements to CNVs formation is highlighted by the higher LCR density in regions enriched with CNVs (~25%), as compared to other CNV-poor genomic regions (~5%) (Cooper *et al.*, 2007).

The formation of CNVs may be result from a range of different molecular mechanisms including non-allelic homologous recombination (NAHR), non-homologous end joining (NHEJ) or forkhead stalling and template switching (FoSTeS) (Zhang *et al.*, 2009c, Korbel *et al.*, 2007, Zhang *et al.*, 2009a). Out of these three proposed mechanisms, only FoSTeS can cause the formation of complex CNVs that consist of duplications, deletions, and inversions at once (Zhang *et al.*, 2013a).

NAHR is a homologous recombination process that involves the exchange of highly homologous genetic material (i.e. repeat elements), at different chromosomal positions (non-

allelic) during cell division. **Figure 1- 2** shows the events during NAHR that can cause the formation of CNVs. The mis-alignment of homologous stretches in the same orientation followed by the crossing over event can cause the formation of deletions or duplications in between the homologous stretches, while inversions can be formed if the alignment of the homologous stretches is on the opposite orientation (Zhang *et al.*, 2013a). Other than LCRs, short interspersed nuclear elements (SINEs) and long interspersed nuclear elements (LINEs) have also been known to mediate NAHR (Singer, 1982). Both of these elements are highly homologous sequences capable of amplifying themselves (retrotransposons) that are scattered throughout the genome of eukaryotic organisms (Singer, 1982). It has been suggested that NAHR plays an important role in causing a number of recurrent rearrangements underlying genomic disorders (Stankiewicz and Lupski, 2002), and it is very likely to be responsible for causing the formation of most of the large and recurrent CNVs in the human genome (Hastings *et al.*, 2009).

The second mechanism proposed to give rise to CNVs is NHEJ, which is one of the DNA repair mechanisms. The NHEJ usually occurs when there is a double-stranded break (DSB) caused by external exposures, such as radiation. **Figure 1- 3** shows the processes involved in the DNA repair by NHEJ mechanism that can lead to the formation of CNVs. The NHEJ-mediated DNA repair is initiated by the binding of Ku proteins to the DSB site, and the recruitment of DNA PKcs. DNA PKcs are DNA kinases responsible for initiating a signalling cascade that subsequently recruits ligases (XRCC4 and Ligase IV) to the DSB site. Upon recruitment, the ligases re-ligate the double-stranded DNA to complete the NHEJ-mediated DNA repair. However, the NHEJ is an error-prone process, whereby inaccuracy of the ligation and rearrangement errors are likely to cause the formation of deletions and duplications, as shown in **Figure 1- 3**. The NHEJ events are believed to be driven by microhomology (de Smith

*et al.*, 2007) and its breakpoints are concentrated in highly repetitive sequences such as LCRs, SINEs, LINEs and most recently, *Alu* elements (de Smith *et al.*, 2007).

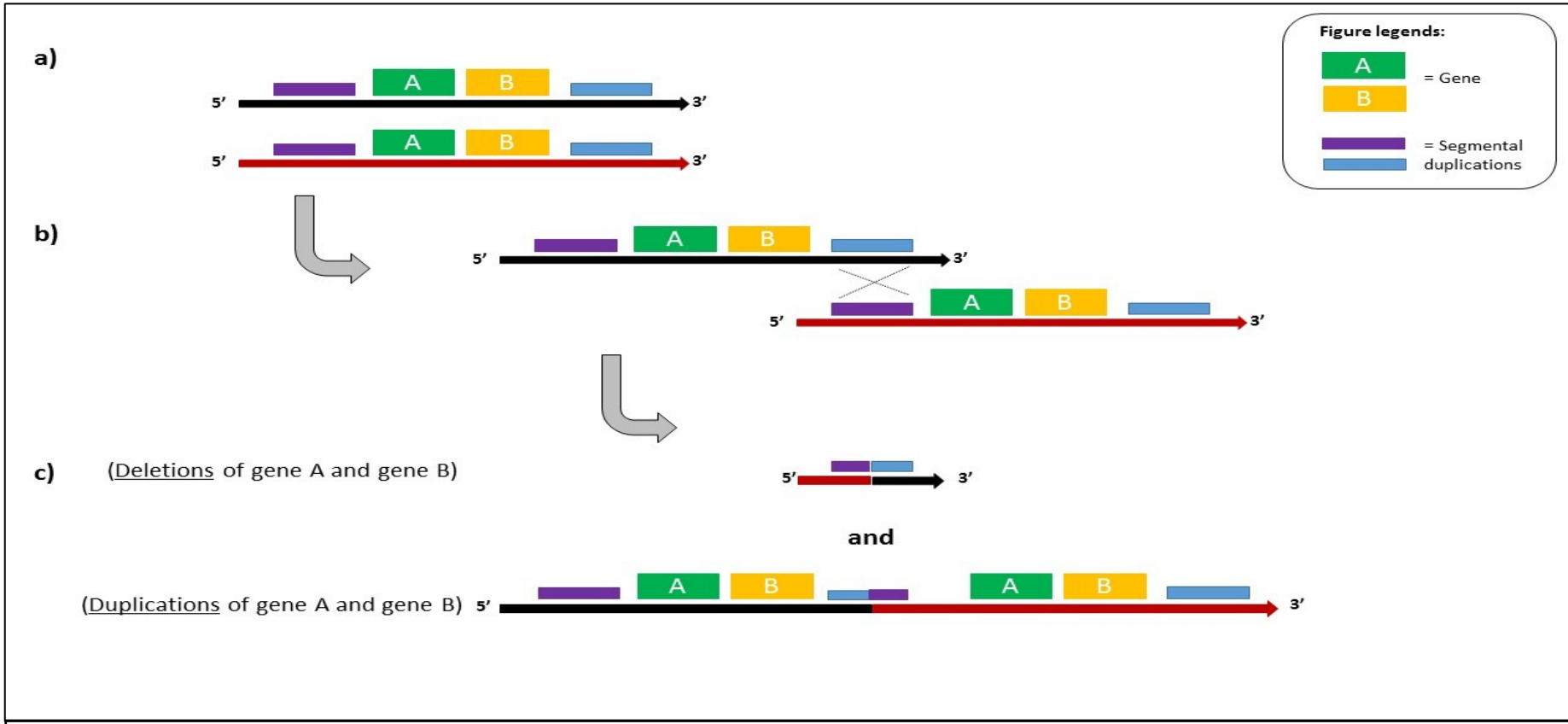


**Figure 1- 2:** The processes involved during the non-allelic homologous recombination (NAHR) mechanism that can cause the formation of genomic deletions and duplications. a) Gene A and gene B are located between two segmental duplications on a pair of homologous chromosomes. b) The segmental duplications flanking both genes mis-align in the same orientation during crossing over in meiosis. c) The recombination results in a chromosome with both genes deleted, and another chromosome with both genes being duplicated.

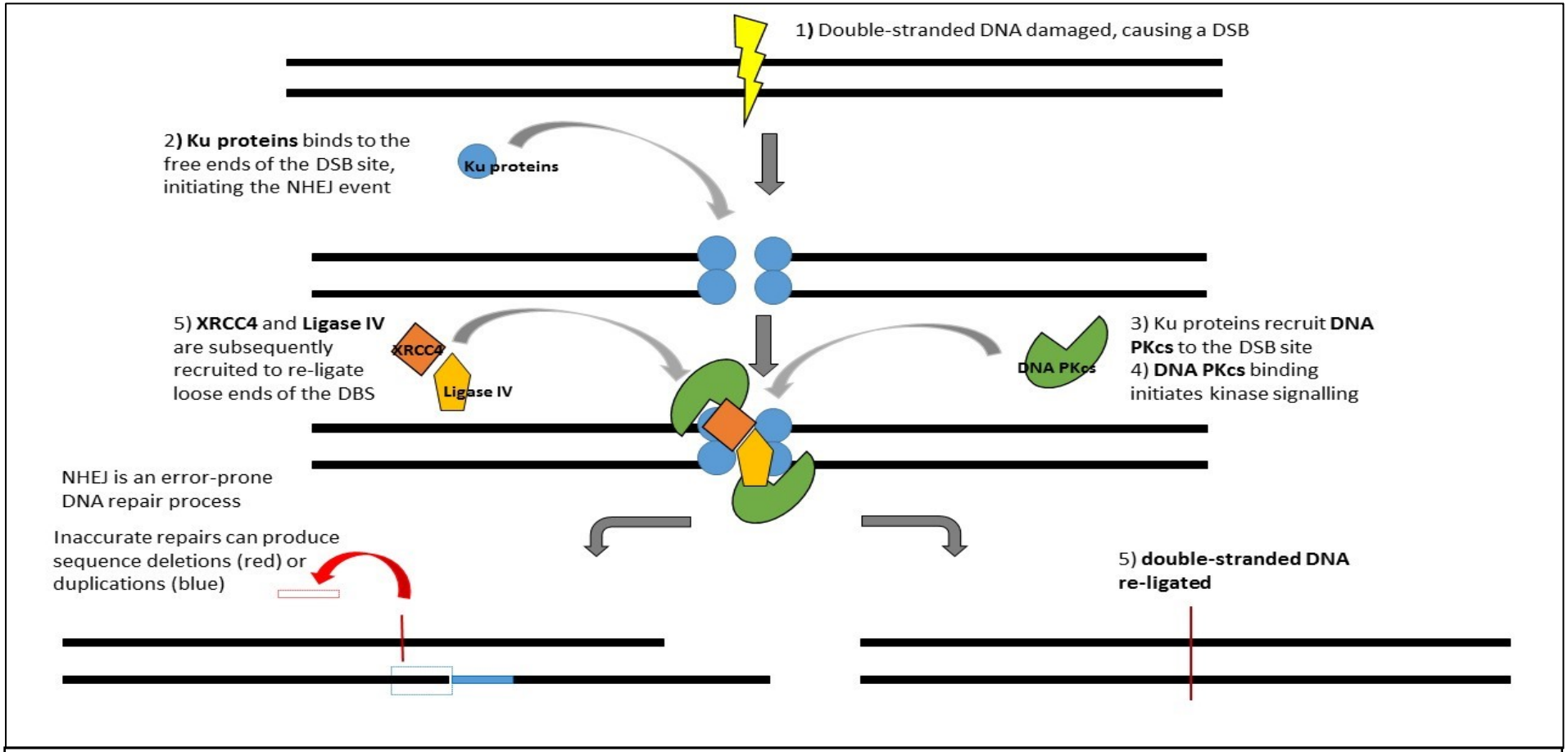


**Figure 1- 3:** The processes involved in the non-homologous end-joining (NHEJ) DNA repair mechanism that can lead to the formation of genomic deletions and duplications. The DNA repair mediated by NHEJ is prone to errors during ligation and rearrangements at double-stranded DNA break (DSB) sites, which can give rise to CNVs.

The third molecular mechanism that has been recently proposed to give rise to CNVs is FoSTeS, which is a mechanism based on the DNA replication process. FoSTeS is thought to be responsible for the formation of complex CNVs via intricate rearrangements during DNA replication. It was initially proposed to elucidate the formation of complex duplications identified in Pelizaeus-Merzbacher disease, in which the complex genomic rearrangements could not be explained by NAHR or NHEJ (Lee *et al.*, 2007). During DNA replication, DNA lesions or nicks can stall the active replication fork upon reaching the damaged site (hence the term forkhead stalling), causing the lagging strand to disengage from the active fork and switch to another template where another replication fork occurs for replication, hence the term template-switching. This process of stalling and switching of templates can generate duplications, deletions, and if there are many FoSTeS occurring at the same time within a genomic region, complex CNVs can be formed in the region. It has been suggested that FoSTeS is mediated by the presence of microhomology, and because of that, it is also called microhomology-mediated break-induced replication (MMBIR).

### 1.3.3. The importance of CNVs in human health

The number of studies to investigate structural variations, particularly CNVs has grown rapidly in recent years. In the last decade, evidence for a significant contribution of CNVs to the inter-individual phenotypic diversity in the general population is growing (Almal and Padh, 2012, Wain *et al.*, 2009, Zhang *et al.*, 2009b). Many phenotypic traits were believed to be driven by dosage-sensitive genes or through their regulatory regions. If deletions or duplications occur involving these dosage-sensitive genes (or to important regulatory elements), these events could be detrimental to the expression of the genes, since changes

in the gene copy numbers were significantly correlated with gene expression (de Smith *et al.*, 2010, Henrichsen *et al.*, 2009).

Common CNVs have been reported to be associated with risk of diseases and some show evidence of balancing selection. For example, deletions in a copy number variable region containing genes for low-affinity, cell-surface receptors (Fc receptors) have been previously associated with various inflammatory diseases, such as rheumatoid arthritis (RA) and systemic lupus erythematosus in different populations (Rahbari *et al.*, 2017). These receptors are crucial in mediating the inflammatory response and deletions in one of the Fc receptor genes, the *FCGR3B* gene, have been shown to increase the risk of RA (Rahbari *et al.*, 2017). CNVs at highly polymorphic genomic regions such as in the 8p23.1 chromosomal region have also been implicated in psoriasis (Hollox *et al.*, 2008, Stuart *et al.*, 2012). This region which contains a cluster of beta-defensin genes, has copy numbers ranging from 2 to 7 (Hollox *et al.*, 2008). It has been found that increasing beta-defensin copy numbers significantly increases the risk of psoriasis (Hollox *et al.*, 2008, Stuart *et al.*, 2012).

Moreover, CNVs in individuals with chromosomal abnormalities (i.e. genomic syndromes) may explain the differences in severity observed between them. For example, Down's syndrome (DS) is a genetic disorder caused by either the whole, or part of chromosome 21 being duplicated once, which resulted in the presence of a third chromosome 21 (also known as trisomy 21). Patients with DS have a wide spectrum of severity in their clinical features, while their prognosis, competencies and disabilities differ greatly from one another. With the right support in their clinical and social needs, they can have an active lifestyle. However, some DS patients have increased risks for mortality and morbidity during their early stages of development (de Smith *et al.*, 2010). Yang, *et. al.* (2002) reported that ~40% of the DS patients had congenital heart defects (CHD), while half of them had to undergo surgical procedures

for CHD in the first year of their lives (Yang *et al.*, 2002). Crucially, the DS patients with CHD had ~76% of survival rate compared to the DS patients without CHD (~91%) (Yang *et al.*, 2002). Therefore, investigations to elucidate the mechanisms and pathways that drive the variations in each phenotype are significantly important for these DS patients. By identifying susceptibility loci across the genome, researchers can assist the clinicians to predict the development of certain chronic phenotypes, such as CHD among the DS patients (de Smith *et al.*, 2010).

In the same case of DS, the differential effects of CNVs with different breakpoints on chromosome 21 have been previously described (de Smith *et al.*, 2010). The authors mentioned that having a full extra copy of chromosome 21 did not result in the total phenotypic expression associated with DS (de Smith *et al.*, 2010). Although mental retardation and distinct facial characteristics were common in patients with DS, the severity of these clinical features remains varied (de Smith *et al.*, 2010, Kallen *et al.*, 1996). Other than CHD occurrences in some of the DS patients (as mentioned previously), a 21q21 duplication (sizes ranged from ~0.5 Mb to ~6 Mb) could cause familial Alzheimer's disease, but not DS (Cabrejo *et al.*, 2006). The discovery of another duplication in the 21q22 region (~4.3 Mb) was significant in classifying the DS critical region (DCR), since all affected individuals from the same family exhibited DS's facial features and mild cognitive dysfunction (Ronan *et al.*, 2007). Therefore, these findings have suggested that genomic syndromes with variable phenotypic expressions and severity can be caused by different CNV events in the same genomic region, further highlighting the importance of CNVs in human health.

## 1.4. Obesity as a chronic disease

### 1.4.1. The definition of obesity

Obesity is a multi-factorial health condition that is characterised by an excessive amount of fat stored as adipose tissue in the body. It has become a global health issue due to its contribution to various chronic diseases and the many detrimental effects of obesity on overall well-being. Adipose tissue is a necessary component in maintaining the overall well-being of a human body. This type of tissue provides storage for food and water, insulation for controlling the body temperature, while giving support and protection to the internal organs. Other than that, the adipose tissue excretes important hormones that regulate the reproductive system. In addition, the adipose tissue plays an important role in the regulation of the immune system due to its functional property of generating new immune cells. Insufficient adiposity could lead to detrimental complications such as infertility, miscarriage, inability to fight off infections, and osteoporosis.

Consistent with its important role in health and reproductive fitness, the regulation and maintenance of adiposity in a human body is tightly controlled by a complex system with various signalling pathways. However, a combination of low physical activity, a calorie-dense diet, deteriorating psychological conditions (i.e. sleep deprivation, stress) and genetic variations could have adverse effects on the complex signalling pathways regulating the adiposity of the body, eventually leading to the development of obesity.

### 1.4.2. Phenotyping of obesity

Obesity is commonly classified based on the body mass index (BMI), a value calculated by dividing the body weight, in kilograms (kg), with the square of height, in meters (m). The World Health Organisation (WHO) have determined the obesity classifications according to increasing health risks posed by obesity (WHO, 2000), as shown in **Table 1- 1**. In adults, individuals with BMI ≥ 30 kg/m$^2$ are considered to be obese. Infants (aged 2 to 5 years), children (aged 6 to 11 years) and adolescents (aged 12 to 19 years) who have a BMI ≥ 95$^{th}$ percentile according to the clinical growth charts based on age and sex (WHO, 2000, Chinn, 2006), are defined as obese. In extreme cases of obesity, adults with BMI ≥ 40 kg/m$^2$ and adolescents with BMI ≥ 99$^{th}$ percentile in their clinical growth chart are classified as morbidly obese(WHO, 2000). These obesity classifications are regardless of sex.

| Weight category for adults | BMI, kg/m$^2$ |
|---|---|
| Underweight | < 18.5 |
| Normal weight | 18.5 to 24.9 |
| Overweight | ≥ 25.0 |
| Obese | ≥ 30.0 |
| Class I | 30.0 to 30.9 |
| Class II | 35.0 to 39.9 |
| Class III (morbidly obese) | ≥ 40.0 |

**Table 1- 1:** The categories of body mass index (BMI) according to World Health Organisation (WHO) based on increasing health risks (WHO, 2000).

The BMI has been extensively used as an indicator of obesity due to its associative relationships with adiposity (Gallagher *et al.*, 2000), disease burden (Must *et al.*, 1999) and mortality (Calle *et al.*, 1999). Other than that, it is widely accepted as a proxy-measure for obesity because of its calculation simplicity and data availability, since weight and height information are usually collected as part of any clinical examinations. The ease of collection

of these measures allows direct comparisons between different populations across various nations worldwide. The major drawback of using BMI to classify obesity is its inability to differentiate between weights of fat mass and muscle mass. This problem could lead to misidentifying a highly-muscular person as being overweight or obese (Muller *et al.*, 2010). Similarly, since BMI does not imply the exact health condition of an individual, older people who may have lost muscle mass replaced by excessive fat content could be mis-classified as healthy (Muller *et al.*, 2010).

Other phenotypic measurements commonly used to measure general obesity are waist circumference (WC) and waist-to-hip ratio (WHR) (Janssen *et al.*, 2002). These two measures are often available from clinical examinations, albeit to a lesser extent than BMI. They are clinically feasible to obtain, and they can be used together or independently, with BMI in assessing obesity. Unlike BMI, the WC and WHR for men and women are different due to the anatomical differences between the two sexes. These two phenotypic measures are more informative than BMI when assessing central obesity (also known as abdominal obesity). Central obesity is a type of obesity that occurs when an excessive amount of fat is localised around the abdomen, which is associated with detrimental effects on health. Both WC and WHR are significantly correlated with BMI and are related to increased risk of mortality (due to all causes) (Janssen *et al.*, 2002). They have stronger predictive powers to estimate abdominal fat in young adults compared to old people (Seidell, 2010) and may not be easily applicable to people who are very obese, when waist definition may be difficult. Another difficulty is that there is a lack of evidence to suggest an appropriate cut-off for each of these abdominal measurements in classifying obesity in general.

The use of BMI, WC and WHR data for assessing general and abdominal obesity are useful in clinical settings, but more detailed information on the anatomical distribution of abdominal

fat would be preferable for assessing risk of metabolic syndromes, such as obesity. Although BMI has a strong correlation with fat mass (Keys *et al.*, 2014), a recent study showed that the significant of association was only reserved for obese individuals, with no association detected between BMI and fat mass in non-obese people (Muller *et al.*, 2010), underlining the importance of more detailed fat measurements. Furthermore, sex and age differences have significant effects on the associative relationship between BMI and body fat mass (Muller *et al.*, 2010). Thus, the use of adiposity measurements, such as content percentages in total abdominal adiposity, subcutaneous adiposity, visceral adiposity and pancreatic adiposity, could give a better indication of risk to metabolic health. However, the cost of phenotyping the body fats is high; not all clinical or research settings have the necessary equipment to do so. In addition, body fat scanning is time-consuming and technologically demanding. These issues have hindered the use of body fat measurements in comparisons between different populations, but a large-scale MRI-based survey is currently being undertaken in the UK biobank (West *et al.*, 2016).

In this thesis, I used BMI as an indicator of obesity as a binary trait, while BMI and WHR were used as quantitative obesity-related phenotypes in all results chapters. In the NutriTech dataset, magnetic resonance imaging (MRI) and DEXA-based measures of fat content of various depots, were used in addition to BMI and WHR in **Chapter 6**.

### 1.4.3. Implications of obesity for economy and health

Obesity has become a major epidemic health issue (James *et al.*, 2001), with its consequences include increased risks of various cancers (Slattery *et al.*, 2007, Stoll, 1999, Wolff, 1987), cardiovascular diseases (Hubert *et al.*, 1983, Manson *et al.*, 1990), type 2 diabetes mellitus (T2DM) (Hu *et al.*, 2001), hypertension (Racette *et al.*, 2003), asthma (Chen

*et al.*, 2002) and depression (Friedman *et al.*, 2002, Roberts *et al.*, 2000). Furthermore, this epidemic pose significant burdens to economic development of the nations (Hamilton *et al.*, 2018, James and McPherson, 2017). In England, it was reported that the National Health Service (NHS) had spent an estimated £6.1 billion to address health issues related to obesity between 2014 and 2015, an amount which surpassed the combined total of spending for the fire service, judicial system and police (PHE, 2017). In the wider society, the overall cost of obesity is currently estimated to be £27 billion (PHE, 2017). Crucially, the UK's economic cost attributed to obesity has been projected to be £9.7 billion by the year 2050 (PHE, 2017), at which ~20% of the world's population are expected to be obese (PHE, 2017).

Besides the economic implications, obesity is harmful to the health and quality of life of the people affected. In England, these individuals have been reported to have difficulties in performing physical activities due to their feelings of being discriminated against because of their appearances and excessive weights (Jackson and Steptoe, 2017). On top of that, the aerobic capacities of the obese people were hindered by obesity, which resulted in low levels of physical activities (Mattsson *et al.*, 1997, Tsuritani *et al.*, 2002).

### 1.4.4. Increasing prevalence in obesity

Obesity is a health problem of a global scale, in which the health consequences caused by obesity have affected approximately 2 billion people globally (Friedrich, 2017). From 1980 to 2015, the prevalence of obesity has been increasing steadily every year (Friedrich, 2017). The prevalence rates for obesity worldwide were reported to have nearly doubled between 1980 and 2008, from 5% to 10% in men and from 8% to 14% in women (WHO - Global Health Observatory data. Accessed via http://www.who.int/gho/ncd/risk_factors/obesity_text/en/). A recent study showed that nearly 10% of the world's population, ~107 million children and

~604 million adults, were reported to be obese in 2015 (Friedrich, 2017). In the same year, ~63% of adults were either overweight or obese in the UK, and ~27% of men and women were obese (PHE, 2017). These findings showed rapidly increasing trends in both sexes when compared to the obesity prevalence in 1993 for the UK population (13.2% of men and 16.4% of women) (PHE, 2017).

The prevalence rates of obesity among children are increasing faster than among the adults in most countries, especially in India and China (Friedrich, 2017). Between the year 2015 and 2016, nearly 20% of children aged 10 and 11 years olds were obese in the UK, while children aged 4 and 5 years old had a nearly 10% of prevalence rate of obesity (PHE, 2017). This is alarming because obesity has become more prevalent during the early years of development in young people, and they are highly likely to have obesity going on into adolescence and adulthood (PHE, 2017).

### 1.4.5. Morbidity and mortality of obesity

A disability-adjusted life year (DALY) is defined by the WHO as one year of 'healthy' life lost (WHO – Health statistics and information systems. Accessed via http://www.who.int/healthinfo/global_burden_disease/metrics_daly/en/).The accumulated DALYs across a population can be considered as the burden of disease for the population. Globally, it was estimated that approximately 4 million deaths and 120 million DALYs were attributed to obesity in 2015 (Friedrich, 2017). Between the year 1990 and 2015, the global mortality and DALY rates attributed to obesity increased by 28.3% and 35.8% respectively (Friedrich, 2017). The leading cause of these deaths was cardiovascular disease (Friedrich, 2017). In the UK, Public Health England (PHE) has reported that an estimated of 30,000 deaths can be attributed to obesity each year (PHE, 2017). In addition, obesity is reported to take

away an average of 9 years of lifespan from an individual (PHE, 2017). These statistics have shown that obesity is a critical health risk in the UK and worldwide; it could become the number one cause of preventable deaths in the UK surpassing tobacco smoking (PHE, 2017), if it is not prevented from now on.

It is crucial, therefore, to understand that obesity is a chronic disease responsible for millions of deaths worldwide, not just the consequence of a state of an uncontrolled eating behaviour due to character flaws or lack of self-control. The compassion for the obese individuals from the societal point of view is of paramount importance for obesity to be recognised as a chronic and progressive health condition, for which the only treatment that is effective long-term is currently bariatric surgery. Naturally, prevention is highly desirable, but public health strategies for this appear to have been ineffective so far and many people are already severely affected. By understanding the mechanisms underlying obesity's aetiology, researchers can assist clinicians to lay the foundations for the development of new therapeutic agents and/or personalised care pathways for those affected.

## 1.5. The genetics of obesity

### 1.5.1. Heritability of obesity

As discussed in the previous section, the prevalence of obesity is increasing rapidly on a global scale. This rapid increase in obesity is certainly contributed to by an unhealthy environment that tends to cause obesity (obesogenic) – calorie-enriched diets, low levels of physical activity, stressful lifestyle and sleep deprivation due to work and other commitments. However, not everyone living in the same unhealthy environment is affected and develops

obesity. With their unique genetic make-ups, different individuals respond to the same obesogenic environment in different ways.

Heritability is defined as an estimated proportion of phenotypic variance that is attributable to genetic variations between individuals in a population (Hindorff *et al.*, 2009). Twin studies are commonly conducted to unravel the genetic and the environmental cues contributing to the expression of specific phenotypes. This method focuses on comparing the concordance rates between monozygotic twins (MZ) and dizygotic twins (DZ) of the same population.

In 1986, a twin study conducted on approximately 2,000 MZ and DZ each, revealed a high heritability (>75%) for height, weight and BMI at 20 years of age (Stunkard *et al.*, 1986a). The heritability for each of the phenotype was increased to ~80% after 5 years. Based on these findings, the author proposed that human fatness is substantially controlled by genetics (Stunkard *et al.*, 1986a). After a few years, a larger twin study conducted on more than 7,000 Finnish twins (MZ and DZ) of the same sex had estimated the heritability of BMI for men and women to be 72% and 66.4% respectively (Turula *et al.*, 1990). In addition, an epidemiological study on approximately 5,000 young twin pairs (aged 8-11 years) in the UK has shown significant evidence of a strong genetic contribution on childhood obesity, despite growing up in an obesogenic environment and during the obesity epidemic (Wardle *et al.*, 2008b). The researchers reported high heritability (>75%) for BMI and WC, as well as a substantial proportion of genetic contribution shared between BMI and WC (60%) (Wardle *et al.*, 2008b). These studies came to a similar conclusion that the concordance rate between the MZ pairs (~0.7) was nearly 3 times that of the DZ pairs (~0.3). Since then, many twin studies have shown similarly strong evidence of high heritability in body weight, with the heritability estimates for BMI ranged from ~30% to 90% (Min *et al.*, 2013), for WC at ~75% (Rose *et al.*, 1998, Wardle

*et al.*, 2008b) and for WHR at ~45% (Rose *et al.*, 1998). These findings highlight the influence of the genetic component in increasing the risk of obesity. Furthermore, a meta-analysis based on 50 years of twin studies consisting of ~14 million twin pairs in total and various weight measurements revealed the heritability estimate for the 'weight maintenance' group to be ~72% (Polderman *et al.*, 2015). Furthermore, in a study conducted on approximately 500 Danish adoptees (adults), the adoptee's BMI was strongly correlated with the BMIs of the biological parents, but not the adopted parents (Stunkard *et al.*, 1986b). This further indicates the role of the genetic component over the shared environmental component in families.

### 1.5.2. The genetic causes of obesity

The strong evidence from twin studies has highlighted the importance of the genetic contribution to body weight regulation and obesity development. However, it is believed that exposure to obesogenic environment is also necessity for obesity to manifest and develop into a severe condition. Since genetics influences different individual responses to the same obesogenic environment, an interaction between genes and the environment is thought to play an important role in the aetiology of obesity, and there is emerging evidence that this is the case.

Obesity is a highly heterogeneous condition, with different forms of obesity resulting from different proportions of genetic and environment factors. **Figure 1- 4** illustrates the monogenic, chromosomal or syndromic, and polygenic obesity classifications according to the relative contributions between the genetics and environment. Monogenic obesity is a rare, severe obesity primarily caused by a single gene defect. Syndromic obesity is another rare form of obesity that is clinically diagnosed together with various cognitive defects (i.e. intellectual disability, developmental delay), dysmorphic features, and/or developmental

abnormalities (i.e. large head, or particular facial features). The main causes of syndromic obesity are deletions or duplications in specific chromosomal regions, such the loss of the paternal 15q11.2 chromosomal region in Prader-Willi syndrome (PWS; OMIM #176270).

On the other hand, polygenic (or common) obesity is the most common form of obesity in the general population. Common obesity is likely to be caused by a combination of environmental factors and common genetic variations. It should be noted however, that the endpoint of obesity is the same, and it is generally not possible to distinguish between aetiologies without molecular investigation.

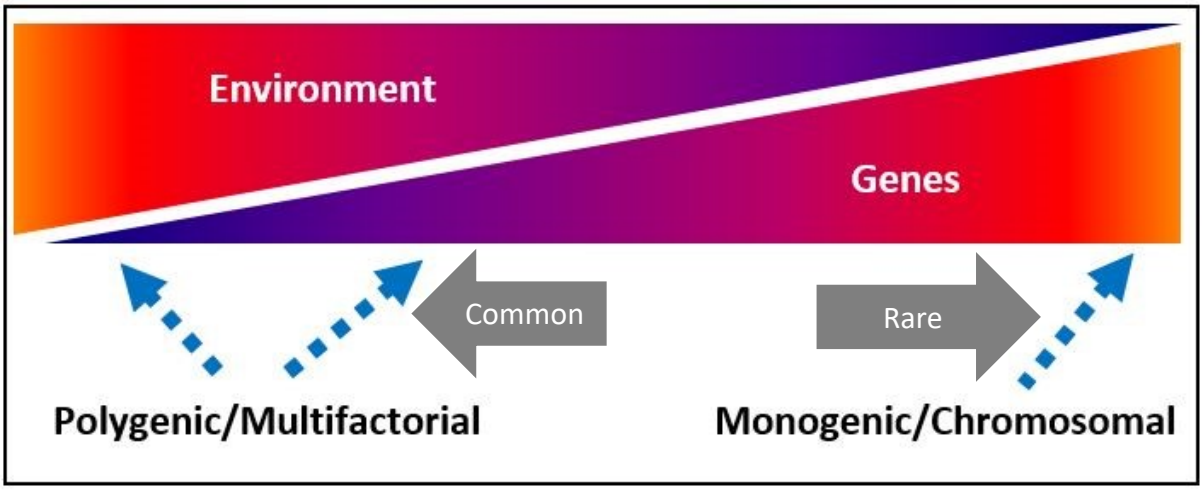


**Figure 1- 4:** The genetics and environment interaction in different sub-forms of obesity. Monogenic obesity is caused by the dysfunction of a single gene and minimal environmental factors. Chromosomal or syndromic obesity is caused by abnormalities in specific chromosomal regions and minimal environmental factors. Polygenic obesity (common obesity) is caused by both obesogenic environment and common genetic variations.

### 1.5.3. Leptin deficiency in monogenic obesity

Mouse-model studies were crucial in understanding the genetic basis of body weight, diabetes and obesity. The pioneering studies were described by Coleman and Hummel in the late 1960s and early 1970s, with the introduction of *obese (ob)* and *diabetic (db)* mice models (Coleman and Hummel, 1967, Coleman and Hummel, 1973). In these studies, both mouse models were reported to develop early-onset morbid obesity, resistance to insulin, hyperphagia (excessive eating behaviour) and hyperglycaemia (excessive blood glucose) (Coleman and Hummel, 1967, Coleman and Hummel, 1973). Further investigations have led to the cloning and sequencing of causative genes from both mouse models, encoding the leptin hormone (in *ob/ob* mouse) and the leptin receptor (in *db/db* mouse). In human, leptin which is produced primarily by adipocytes, is a hormone known to inhibit hunger and regulate energy balance (Farooqi and O'Rahilly, 2009). It acts through the leptin receptor (a cytokine receptor family) and plays a significant role in a signalling pathway involved in energy homeostasis, which is called the leptin-melanocortin pathway (Farooqi and O'Rahilly, 2009). In this pathway, leptin is secreted to reduce appetite while inducing satiety in response to the increasing level of adiposity. In 1997, a detailed examination of two severely obese children from the same consanguineous family revealed the first evidence of leptin deficiency in humans, hence providing the first evidence of the key role of leptin in human obesity (Montague *et al.*, 1997). In the study, Montague, *et. al.* (1997) sequenced the leptin gene (*LEP*) in two severely obese children whose serum leptin levels were extremely low despite their fat mass being strikingly high, each of them had a homozygous deletion of a single nucleotide that caused a frame-shift mutation in the *LEP* gene (Montague *et al.*, 1997). Since then, more leptin deficiency cases caused by mutations in the *LEP* gene have identified in different studies, raising more evidence that leptin deficiency causes severe early-onset

obesity as one of its clinical features, along with disturbances of reproductive and immune functions (Farooqi *et al.*, 2002, Strobel *et al.*, 1998).

One year after the discovery of the first human leptin deficiency, the first case of leptin receptor deficiency in human was described in a morbidly obese individual from another consanguineous family (Clement *et al.*, 1998). The affected person had a homozygous mutation in the leptin receptor gene (*LEPR*), which deprived the leptin receptor of its intracellular and transmembrane domains (Clement *et al.*, 1998). An increasing number of leptin receptor deficiency cases have been reported ever since, with common clinical features of obesity, hyperphagia and delayed puberty being described in affected individuals (Farooqi *et al.*, 2007).

The discoveries of the *LEP* and *LEPR* genes has increased our understanding of the genetic basis of obesity. Both genes play important roles in the leptin-melanocortin pathway that regulates appetite and eating behaviour in the central nervous system. This led researchers to investigate other genes that are known to be involved in the leptin-melanocortin pathway.

### 1.5.4. More genes implicated in monogenic obesity

The first case of monogenic obesity described in human was leptin deficiency (Montague *et al.*, 1997), while leptin receptor deficiency was identified shortly after that (Clement *et al.*, 1998). This discovery has paved the way to bringing obesity research into the modern era. Since then, severely obese individuals have been examined in detail to provide evidence for more candidate genes in obesity. Successive discoveries were followed through by the identification of various mutations in other known genes involved in the leptin-melanocortin pathway, such as pro-hormone convertase subtilisin/kexin type 1 (*PCSK1*) (Jackson *et al.*,

1997), proopiomelanocortin (*POMC*) (Krude *et al.*, 1998), and melanocortin-4 receptor (*MC4R*) (Vaisse *et al.*, 1998, Yeo *et al.*, 1998).

The *POMC* gene (OMIM *176830) encodes pro-opiomelanocortin, which is subsequently cleaved into smaller signalling peptides,I ncluding melanocyte stimulating hormones (α, β, γ -MSH). The β-MSH peptide exerts a key role role in weight regulation by binding to the melanocortin-4 receptor (encoded by the *MC4R* gene; OMIM *155541) in the brain to cause a signalling cascade responsible in maintaining energy balance regulating food intake. The *PCSK1* gene (OMIM *162150) encodes pro-hormone convertase, an enzyme involved in the hormone processing by cleaving protein precursors – its substrates include pro-opiomelanocortin.

Other than genes known to be directly involved in energy balance and food intake, several mutations in genes related to neural development have also been identified as causative of monogenic obesity. A balanced translocation in single-minded homolog 1 (*SIM1*) gene, which encodes SIM bHLH transcription factor 1, was associated with obesity (Holder *et al.*, 2000). This transcription factor is believed to be involved in the nuclei development of the hypothalamus, where the leptin-melanocortin signalling takes place. A disrupted *SIM1* gene (OMIM *603128) has been proposed to be responsible for the dysmorphic characteristics, abnormal brain development and cognitive impairments of the Down's syndrome (DS) patients (Fan *et al.*, 1996). Mutation screening of the brain derived neurotrophic factor, *BDNF* gene (OMIM *113505) revealed several mutations associated with obesity, eating disorder and attention deficit-hyperactivity disorder (ADHD) (Friedel *et al.*, 2005). The *BDNF*-encoded protein is commonly found in brain regions that control eating behaviour and body weight; thus, it is highly likely to be involved in the regulation and management of these functions. Brain derived neurotrophic factor also plays an important role in the survival, growth and

maturation of neurons. In addition, a *de novo* mutation found in neurotrophic tyrosine kinase receptor type 2 gene (*NTRK2*; OMIM *600456) has been implicated in severe obesity and developmental delay (DD) (Yeo *et al.*, 2004), possibly via the same pathway as the *BDNF* gene because the *BDNF*-encoded protein is a substrate for this membrane-bound receptor. Since these three genes have important involvements in managing the hypothalamus functions, disruptions in these genes cause hyperphagia in response to energy imbalance (Walley *et al.*, 2009). The advent of exome sequencing has allowed molecular characterisation of greater numbers of extremely obese individuals, and several new genes have been revealed in the last five years, including, from the Blakemore group, *CPE* deficiency (Alsters *et al.*, 2015).

### 1.5.5. Chromosomal abnormalities in syndromic obesity

Syndromic obesity is caused by specific genomic anomalies or chromosomal abnormalities spanning several genes. Our current knowledge in this field has recently been reviewed by Kaur, *et. al*. (2017), who state that 79 forms of syndromic obesity have been reported, but most remain to be fully elucidated (Kaur *et al.*, 2017). The obesity phenotype in this form may not be the most dominant feature; it is usually displayed together with other clinical features such as mental retardation, DD, intellectual disabilities (ID), dysmorphic characteristics, or cognitive defects (Farooqi, 2006). One of the most commonly known form of syndromic obesity is PWS, which is usually caused by a heterozygous paternal deletion in the 15q11-q1 region (OMIM #176270). PWS is an imprinted autosomal dominant disorder, characterised by hyperphagia which causes severe early-onset obesity, ID, various dysmorphic features and central nervous system (CNS) dysfunctions, including autism. The characteristics of PWS are due to loss of activity of the paternal copy of the region, usually by deletion, or, more rarely by maternal uniparental disomy (the presence of two copies from one parent), or the

disruption of the imprinting centre (Ohta *et al.*, 1999). The deleted region typically spans several genes(de Smith *et al.*, 2009), including the imprinted small nuclear ribonucleoprotein polypeptide N gene (*SNRPN*; OMIM *182279), HBII-85 snoRNA genes (*SNORD116-1*; OMIM *605436) and the necdin gene (*NDN*; OMIM *602117). Nonetheless, the exact mechanism underlying the phenotypic expressions of PWS is still unclear and warrants further investigation (de Smith *et al.*, 2009).

There are other clinical syndromes that have been reported in isolated cases. Due to the extremely rare nature of these clinical syndromes, most of them are still under-reported across different populations. For example, the 17p13.3 telomeric region has been implicated in two separate clinical syndromes; the 17p13.3 deletion syndrome, also known as Miller-Dieker syndrome (OMIM #247200) and the 17p13.3 duplication syndrome (OMIM #612576). Both syndromes have been reported to have various deletion and duplication breakpoints spanning different parts of the 17p13.3 chromosomal region. The 17p13.3 deletion syndrome has been associated with ID and abnormal brain structure (Bruno *et al.*, 2010), while the 17p13.3 duplication syndrome has been associated with ID and obesity (Belligni *et al.*, 2012, Kuroda *et al.*, 2014).

Another example is the chromosomal abnormalities in the 16p11.2 region, in which there are two non-overlapping copy number variable regions that have been associated with syndromic obesity and various psychiatric disorders (Bachmann-Gagescu *et al.*, 2010, Bochukova *et al.*, 2010, Maillard *et al.*, 2015, Walters *et al.*, 2010): a the distal ~200-kb 16p11.2 region and a proximal ~600-kb 16p11.2 region. Rare heterozygous deletions of the distal region (also known as the 16p11.2 Deletion syndrome, 220-kb; OMIM #611344) have been implicated in severe early-onset obesity, insulin resistance and developmental delay (DD) (Bachmann-Gagescu *et al.*, 2010, Bochukova *et al.*, 2010), while rare heterozygous

deletion of the proximal region (also known as the 16p11.2 Deletion syndrome, 593-kb; OMIM #611913) is associated with obesity, autism and mental retardation (Maillard *et al.*, 2015, Walters *et al.*, 2010). Interestingly, duplications of the same proximal region (also known as the 16p11.2 Duplication syndrome; OMIM #614671) is implicated in underweight, mirroring the effect of deletion on the risk of obesity (Jacquemont *et al.*, 2011). This highlights the dosage effects of the proximal 16p11.2 region in modulating the risk of syndromic obesity.

From these examples, it is clear that syndromic obesity can be caused by chromosomal structural abnormalities, including both losses and gains of copy number. While the neuro-psychiatric defects are common features in many clinical syndromes, impaired neurological control for food intake and energy balance could be the reason for the fat mass increase in patients with syndromic obesity.

### 1.5.6. Linkage analysis in common obesity

The monogenic and syndromic forms of obesity have the tendency to be highly penetrant, and, thus, identifying causal genetic variants has been successful to date. Common obesity can be caused by the combined effect of several genetic variants in a number of genes, in addition to the various factors from an obesogenic environment, so different methodologies have been adopted for its investigation. One way to search for genetic susceptibility loci in common obesity is to assess the segregation of the genetic markers across generations in affected individuals within the same family. For this reason, genome-wide linkage analysis (LA) was designed and by the year 2006, genome-wide LA studies had been successful in identifying approximately 250 quantitative trait loci (QTL) associated with common obesity (Rankinen *et al.*, 2006). Based on the linkage results, researchers have identified candidate genes for common obesity through the use of positional cloning of the reported QTLs,

including three genes that were believed to have direct involvements in processes related to obesity (Boutin *et al.*, 2003, Durand *et al.*, 2004, Meyre *et al.*, 2005, Suviolahti *et al.*, 2003). Glutamate decarboxylase 2 (*GAD2*; OMIM *138275) (Boutin *et al.*, 2003) is involved in the upregulation of food intake, while the ectonucleotide pyrophosphatase/phosphodiesterase 1 gene (*ENPP1*; OMIM *173335) is involved in the inhibition of insulin receptor activity, thereby affecting cellular energy intake (Meyre *et al.*, 2005). In addition, the solute carrier family 6 member 14 (*SLC6A14*; OMIM *300444) encodes the tryptophan transporter. Tryptophan is a precursor of serotonin, which has an important role in the regulation of appetite (Durand *et al.*, 2004, Suviolahti *et al.*, 2003).

Despite these promising discoveries, none have been successfully replicated in other studies. Moreover, a meta-analysis consisted of ~31,000 individuals from 37 studies could not produce a strong evidence of linkage association between any of the reported QTLs and BMI or common obesity (defined by BMI). This demonstrates the limitation of using the linkage approach when there are no strong effects exerted by common genetic variants in a gene. In particular, failure of replication increased the belief that common genetic variants have small effects in complex diseases such as common obesity.

### 1.5.7. GWAS in common obesity

After years of development, the emergence of SNP genotyping microarray technology has allowed a more robust, high-throughput scanning of genotypes on a genome-wide scale. This approach, which is called genome-wide association study (GWAS), is based on the idea that nearby common genetic variants (minor allele frequency, MAF > 5%) are inherited together in haplotype blocks due to them being in linkage disequilibrium (LD) (International HapMap, 2005). LD is the non-random association of two or more genetic variants in a general

population generally due to close physical proximity limiting recombination events between them. This realisation has provided the researchers with the capability to indirectly capture information on more than 14 million common genetic variants (~80% of the common variations) using only half a million SNPs (Barrett and Cardon, 2006, Herrera and Lindgren, 2010). GWAS is a hypothesis-free approach that typically compares allele frequencies of each polymorphic variant between cases and controls. This array-based SNP genotyping technology is reliable and cost-effective and has been applied to a large number of cohorts worldwide, and researchers has formed into transglobal research consortia such as The Genetic Investigation of Anthropometric Traits (GIANT; https://portals.broadinstitute.org/collaboration/giant/index.php/Main_Page) and The Meta-Analyses of Glucose and Insulin-related traits Consortium (MAGIC; https://www.magicinvestigators.org/), pooling data and thereby, providing the statistical power that had been lacking in linkage and candidate gene studies.

The first successful GWAS in identifying the susceptibility loci for common obesity was reported in 1997, in which Frayling T. M*., et. al.* identified a locus called the fat mass- and obesity-associated gene (*FTO*; OMIM *610966) as being robustly associated with BMI and predisposition to obesity (Frayling *et al.*, 2007). Since then, the evidence of the *FTO* gene as the obesity-associated loci has been replicated in other studies, while many other loci have also been found to be significantly associated with BMI and other obesity-related phenotypes (Goodarzi, 2018, Loos, 2018). Large scale analysis has been fruitful, a key study including more than 300,000 individuals revealed 91 loci associated with BMI at genome-wide significance level, of which 56 were novel findings (Locke *et al.*, 2015) and more than 500 obesity-related loci have been identified to date (Loos, 2018).

It is important to know that, while the genes responsible for causing common obesity are still unknown in many cases, it is possible to identify genes within or nearby the associated loci that are physiologically relevant to the aetiological basis of obesity. Thus far, analysis of general adiposity or BMI has highlighted central appetite control pathways in the brain, whereas investigation of genetic factors influencing where fat is stored in the body implicate adipocyte biology as a key pathway (Loos, 2018). Typically, each GWAS-identified marker confers a very small additional risk: Thus, associated loci may be highly significant at the level of population analysis, but not predictive of obesity in individual people. Genotyping data from GWAS-identified loci can be combined into genetic risk scores (GRS) or polygenic risk scores (PRS) to increase predictive power, but this still does not reach the level of clinical utility (Loos and Janssens, 2017).

### 1.5.8. Missing heritability

As discussed previously, the heritability of BMI is estimated to be ~70% (ranging from 30% to 90%) (Min *et al.*, 2013). However, 97 BMI-associated loci taken together only explained less than 3% of the variance observed in BMI, while it was estimated that the common genetic variations in the whole genome only accounts for less than half proportion of BMI variation (Goodarzi, 2018, Loos, 2018). Furthermore, most of the associated SNPs confer modest effects on the risk of common obesity; the strongest BMI-associated SNP, rs9939609 in the *FTO* gene accounted only ∼1% of the variance observed in BMI (Frayling *et al.*, 2007). This shows that common genetic variations could not sufficiently account for the full heritability observed in BMI, which means that there is a considerable portion in BMI's heritability still unaccounted for ("missing heritability"). It is also important to note that in most common diseases, associated SNPs only account for ~2% to ~15% heritability (McCarroll, 2008). The

"missing heritability" is a prominent issue in the genetics of common diseases. It has been suggested that rare and low frequency variants, along with structural variants such as CNVs, may explain the "missing heritability" of common diseases, such as common obesity (Herrera and Lindgren, 2010, Manolio *et al.*, 2009). This is partly due to the tendency of rare and low frequency variants to have highly penetrant effects (Herrera and Lindgren, 2010).

Recently, many technological advancements have emerged to identify novel rare variants in various diseases, such as exome sequencing. More relevant to this thesis, the extensions of improved SNP genotyping arrays for CNV genotyping and algorithmic CNV calling methods have provided researchers with a mechanism to re-use the extensive pre-existing SNP data generated by global GWAS efforts, using genotype coupled with signal intensity data to investigate the contribution of rare and common CNVs in the heritability of complex diseases and related phenotypes. This provides the background and impetus for the work described in this thesis, which is aimed at exploration of the role of specific common and rare CNVs to obesity and related traits.

# CHAPTER 2. Study Cohorts

# 2.1. Northern Finland Birth Cohort 1966

## 2.1.1. Overview

The Northern Finland Birth Cohort 1966 (henceforth referred to as the NFBC1966 cohort) is a longitudinal, prospective birth cohort study including collection of a variety of data on anthropometric, clinical, biochemical, and mental health of pregnant mothers who were expected to give birth in the year 1966, and their offspring who were born in that year (Rantakallio, 1988). All pregnant mothers and new-borns from the northernmost Finnish provinces of Oulu and Lapland were invited to be included in this birth cohort project (Rantakallio, 1988). The ethics for the NFBC1966 project was approved by The Ethics Committee of the Northern Ostrobothnia Hospital District (Rantakallio, 1988). This project was initiated by Professor Paula Rantakallio (The University of Oulu, Finland), and is led by Professor Marjo-Riitta Jarvelin (Imperial College London, UK) at present. Initially, the objective of the NFBC1966 project was to examine the risk factors for childhood mortality and morbidity focusing on premature births (Rantakallio, 1988), before establishing itself as a full epidemiological study throughout the life course of the participants (Welling *et al.*, 2017). The overall aim of this project is to promote health and well-being of the Northern Finland population (Welling *et al.*, 2017). A comprehensive information on the NFBC1966 project is provided here: http://www.oulu.fi/nfbc/.

The NFBC1966 cohort consisted of 12,068 mothers, and 12,231 children born in 1966. A total of 12,058 live-born offspring have had regular follow-ups since their mother's pregnancy. Various clinical examinations and health-related questionnaires were conducted to monitor their growth parameters and overall well-being (**Figure 2- 1**). All participating mothers and offspring (when they reached adulthood) gave written informed consent for their involvement in the NFBC1966 project. **Table 2- 1** shows the number of births in Oulu and Lapland provinces, recorded for the NFBC1966 cohort.

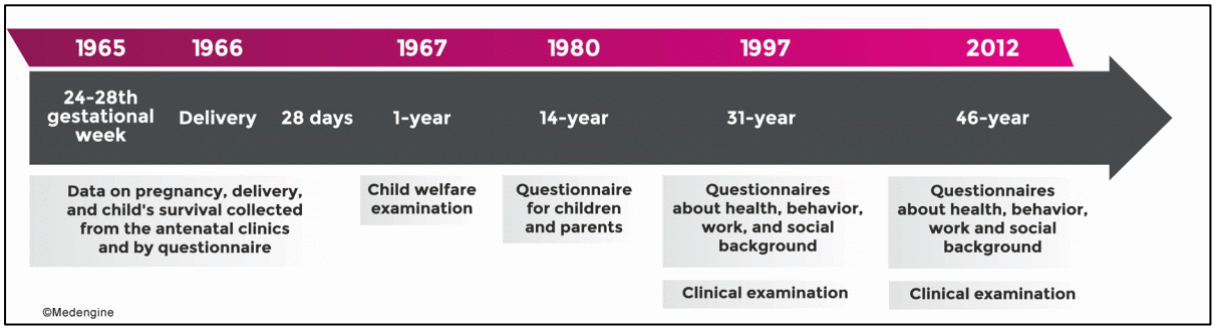


**Figure 2- 1**: Clinical examinations and health-related postal questionnaires during clinical follow-ups on the live-born children of the NFBC1966 cohort. This figure was taken from the official web page of the NFBC1966 project, via http://www.oulu.fi/nfbc/node/19663. NFBC1966 = Northern Finland Birth Cohort 1966.

| Sex/Birth type | Total (n) | Live-born (n) | Still-born (n) |
|---|---|---|---|
| **Sex** | 12,231* | 12,058 | 173 |
| Male | 6,265 | 6,169 | 96 |
| Female | 5964 | 5,889 | 75 |
| **Birth type** | | | |
| Singletons | 11,905 | 11,744 | 161 |
| Twins | 326 | 314 | 12 |
| Triples | 0 | 0 | 0 |

**Table 2- 1:** The number of birth records in the NFBC1966 cohort. *included two still-born offspring whose sex could not be determined. n = number. This table was adapted from the information given on the official web page of the NFBC1966 project, via http://www.oulu.fi/nfbc/node/44315. NFBC1966 = Northern Finland Birth Cohort 1966.

### 2.1.2. Data and sample collections

During all clinical examinations and questionnaires, various anthropometric (i.e. weight, height, anatomical measurements) and clinical data (i.e. sex, medication and clinical records) were collected regularly from the live-born children of the NFBC1966 cohort. In addition, biological sources and tissues, such as blood, faecal and urine samples were also collected during the clinical visits. To date, the data and samples of these live-born children at 31 years and 46 years old, are now available. For this thesis, the NFBC1966 cohort dataset was based on the biological sources collected, clinical examinations and questionnaires performed at the 31-year time point. Briefly, postal questionnaires were sent to the participants with known addresses in Finland (n=11,322) with a response rate of 77% (n=8,767) (Welling *et al.*, 2017). A total of 8,467 participants were invited for clinical examinations at 31 years old. Approximately 71% of them attended the clinical visits (n=6,033) (Welling *et al.*, 2017). All clinical examinations were carried out in a manner that is in accordance to the principles of the Declaration of Helsinki.

**Table 2- 2** shows the number of participations that were recorded for the NFBC1966 cohort. Approximately 92% of the blood samples collected (46% of the total live-births) in this project were sent for genomic DNA extraction and SNP genotyping (n=5,557). Thus, the NFBC1966 cohort dataset used in this thesis included only the participants at 31 years of age, whose DNA was sent for SNP genotyping (n=5,557). The characteristics of these participants are summarised in **Table 2- 3**.

| Participants | n (% of total) |
|---|---|
| Total live-births | 12,058 (100.00) |
| Postal questionnaires | 8,767 (72.71) |
| Clinical examinations | 6,033 (50.03) |
| Genomic DNA collection | 6,033 (50.03) |
| Sent for SNP genotyping | 5,557 (46.04) |

**Table 2- 2:** The participation summary of the live-births in the NFBC1966 cohort. n = number.

NFBC1966 = Northern Finland Birth Cohort 1966.

| Characteristics | Both |
|---|---|
| Number of participants (n) | 5,557 |
| Mean Age (years) | 31.0 |
| Male | 31.1 |
| Female | 31.0 |
| Sex (n) | |
| Male | 2,718 |
| Female | 2,837 |
| Missing data | 2 |
| Mean BMI at 31 years old (kg/m$^2$) | 24.65 |
| Male | 25.20 |
| Female | 24.14 |
| Mean WHR at 31 years old (ratio unit) | 0.86 |
| Male | 0.91 |
| Female | 0.81 |
| Obesity/Morbid obesity status (n) | |
| Obese (30.0 ≤ BMI < 40.0 kg/m$^2$) | 391 |
| Morbidly obese (BMI ≥ 40.0 kg/m$^2$) | 37 |
| Missing data | 70 |

**Table 2- 3**: The characteristics of the adult participants (at 31-year time point) who had their DNA sent for SNP genotyping in the NFBC1966 cohort. Adults with BMI between 30 kg/m$^2$ and 40 kg/m$^2$, were classified as obese. Adults with BMI equal to or more than 40 kg/m$^2$, were classified as morbidly obese. n = number; BMI = Body mass index, WHR = waist-to-hip ratio, NFBC1966 = Northern Finland Birth Cohort 1966.

# 2.2. Northern Finland Birth Cohort 1986

## 2.2.1. Overview

The Northern Finland Birth Cohort 1986 (referred to hereafter as the NFBC1986 cohort) is a prospective epidemiological birth cohort study, consisting of all live-births that took place between July 1985 and June 1986 in the two northern provinces of Finland (Oulu and Lapland) (Jarvelin *et al.*, 1997). This longitudinal birth cohort was based on the pregnant women who were expected to give birth within the 1-year time frame (Jarvelin *et al.*, 1997). The ethical approval for the NFBC1986 project was given by the Ethics Committee of the Northern Ostrobothnia Hospital District (Jarvelin *et al.*, 1997). The NFBC1986 cohort was also initiated by Professor Paula Rantakallio and is currently led by Professor Marjo-Riitta Jarvelin. The main objective of this project is to promote overall healthy well-being of the population under study, with specific interests in social environment, cardiovascular and other health issues (i.e. obesity, diabetes) (Jarvelin *et al.*, 1997). Detailed information on the NFBC1986 project can be found via http://www.oulu.fi/nfbc/. The NFBC1986 cohort is considered to be a matched, but younger, cohort to the NFBC1966.

The NFBC1986 project team recruited 9,362 mothers with 9,479 births taking place within the expected time frame (between July 1985 and June 1986) (Hartikainen-Sorri *et al.*, 1990, Jarvelin *et al.*, 1997, Rantakallio and Oja, 1990). Written informed consent was obtained from the parents, as well as their offspring (born within between July 1985 and June 1986) at the mean age of 15.5 years of age (Hartikainen-Sorri *et al.*, 1990, Jarvelin *et al.*, 1997, Rantakallio and Oja, 1990). **Table 2- 4** shows the number of births that occurred between July 1985 and June 1986 in Oulu and Lapland (North Finland), stratified by sex and birth types.

| Sex/Birth type | Total (n) | Live-born (n) | Still-born (n) |
|---|---|---|---|
| **Sex** | 9,479 | 9,432 | 47 |
| Male | 4,891 | 4,865 | 26 |
| Female | 4,587 | 4,567 | 20 |
| **Birth type** | 9,479 | 9,432 | 47 |
| Singletons | 9,247 | 9,203 | 44 |
| Twins | 226 | 223 | 3 |
| Triples | 6 | 6 | 0 |

**Table 2- 4:** The number of births in the NFBC1986 cohort, according to sex and birth types. n = number. This table was adapted from the information given on http://www.oulu.fi/nfbc/node/44315. n = number, NFBC1986 = Northern Finland Birth Cohort 1986.

### 2.2.2. Data and sample collections

Biological sources (including blood samples), anthropometric (i.e. birth, current) and clinical data (i.e. medication, clinical and history records), as well as various health-related questionnaires have been collected from the parents and the offspring (when appropriate) since the prenatal period. This information was updated in subsequent follow-ups, in particular when the offspring reached adolescence at the age of 16 years old (between 2001 and 2002). During this period, the parents received postal questionnaires that included behavioural, mental, social and developmental questions. Approximately 80% of them responded (n=7,547) (Pinola *et al.*, 2012). Those with known addresses in Finland were invited to attend clinical examinations, with a participation rate of 74% (n=6,798). Various data were obtained from the clinical examinations, including blood pressure, BMI, WHR, insulin levels, glucose levels, and other biochemical and anthropometric measurements. At the same time, whole blood samples were also collected (n=6,427) (Paternoster *et al.*, 2011). All clinical examinations and assessments were conducted according to the principles of the Declaration of Helsinki.

**Table 2- 5** shows the summary of participation in the NFBC1986 cohort. Approximately 63% of the blood samples collected (42.9% from the total live-births) were sent for SNP genotyping (n=4,043). The genotyping was conducted in two batches; 2,016 samples in the first batch (will be referred to as NFBC1986A) and 2,027 samples in the second batch (will be referred to as NFBC1986B). For this thesis, only the adolescents with their DNA samples sent for SNP genotyping were included for analysis (n=4,043). The characteristics of these genotyped adolescents at 16 years of age, are shown in **Table 2- 6**.

| Participants | n (% of total) |
|---|---|
| Total live-births | 9,432 (100.00) |
| Postal questionnaires | 7,547 (80.01) |
| Clinical examinations | 6,798 (72.07) |
| Genomic DNA collection | 6,427 (68.14) |
| Sent for SNP genotyping | 4,043 (42.9) |

**Table 2- 5:** The participation summary of the live-births in the NFBC1986 cohort. Only the live-births with their DNA samples sent for SNP genotyping were included for analysis (n=4,043). n = number, NFBC1986 = Northern Finland Birth Cohort 1986.

| Characteristics | NFBC1986A | NFBC1986B | Both |
|---|---|---|---|
| Number of participants (n) | 2,016 | 2,027 | 4,043 |
| Mean Age (years) | 16.03 | 15.98 | 16.01 |
| Male | 16.03 | 16.01 | 16.02 |
| Female | 16.04 | 15.96 | 16.00 |
| Sex (n) | | | |
| Male | 1,004 | 963 | 1,967 |
| Female | 1,012 | 1,058 | 2,070 |
| Missing data | 0 | 6 | 6 |
| Mean BMI at 16 years old ($kg/m^2$) | 21.33 | 21.31 | 21.32 |
| Male | 21.33 | 21.24 | 21.28 |
| Female | 21.34 | 21.37 | 21.36 |
| Mean WHR at 16 years old (ratio unit) | 0.80 | 0.80 | 0.80 |
| Male | 0.83 | 0.83 | 0.83 |
| Female | 0.77 | 0.78 | 0.78 |
| Obesity/Morbid obesity status (n) | | | |
| Obese ($95^{th}$ percentile ≤ BMI < $99^{th}$ percentile) | 102 | 91 | 193 |
| Morbidly obese (BMI ≥ $99^{th}$ percentile) | 3 | 2 | 5 |
| Missing data | 16 | 58 | 74 |

**Table 2- 6:** The characteristics of the adolescents (at 16-year time point) who had their DNA sent for SNP genotyping in the NFBC1986 cohort. Adolescents with BMI between $95^{th}$ and $99^{th}$ percentile in their clinical growth chart, were classified as obese. Adolescents with BMI in the $99^{th}$ percentile and above in their clinical growth chart, were classified as morbidly obese. n = number; BMI = Body mass index, WHR = waist-to-hip ratio, NFBC1986 = Northern Finland Birth Cohort 1986.

# 2.3. NutriTech cohort

## 2.3.1. Overview

The Application of New Technologies and Tools to Nutrition Research project, also known as the NutriTech project, was an interventional research study that was conducted between April 2012 and January 2016. This study was approved by the UK Research Ethics Committee and registered at ClinicalTrials.gov (identifier: NCT01684917) (Clinicaltrials.gov, 2016a). The primary aim of the NutriTech project was to develop new methods in order to better understand the interplay between diet, health, and disease. The aim was addressed by multiple efforts to develop new biological "markers" that could predict accurately the health status of individuals(Clinicaltrials.gov, 2016a).

The NutriTech project team recruited 74 participants in total. The participants were initially selected based on the criteria shown in **Table 2- 7**. The recruitment and assessment processes were carried out between June 2012 and April 2014, led by Professor Gary S. Frost (Imperial College London, London, UK). The ethics for the NutriTech project was approved by the London Brent Ethics Committee (12/LO/0139). Approximately half of the participants were put into the diet intervention group (Group 1), while the remaining were in Group 2, as controls (48.6%).

| Inclusion Criteria | Exclusion Criteria |
|---|---|
| Healthy overweight (BMI range = 18.5-35 kg/$m^2$, inclusive) | Weight change of ≥ 3kg in the preceding 3 months |
| Age range = 50-65 years, inclusive | Gastrointestinal disease e.g. inflammatory bowel disease or irritable bowel syndrome |
| | Substance abuse |
| | Excess alcohol intake |
| | Pregnancy |
| | Diabetes |
| | Cardiovascular disease |
| | Cancer |
| | Current smokers |
| | Kidney disease |
| | Liver disease |
| | Pancreatitis |
| | Have any metallic or magnetic implants such as pacemakers |
| | Use of medications likely to interfere with energy metabolism, appetite regulation and hormonal balance (including anti-inflammatory drugs or steroids, antibiotics, androgens, phenytoin, erythromycin or thyroid hormones) |
| | Claustrophobia |

**Table 2- 7:** Initial participant selection criteria in NutriTech Research. This table was adapted from the ClinicalTrials.gov website on 'Application of New Technologies and Tools to Nutrition Research (NutriTech)'(Clinicaltrials.gov, 2016a).

### 2.3.2. Data and sample collections

In this thesis, the term 'NutriTech cohort' is represented by the 74 individuals who participated in the NutriTech project. The mean age was 59.3 years, and 52.2% were male. About 51% of the participants (n=38) were put under a strict dietary intervention program for 12 weeks, where they received 20% deficit calorie-restricted diet, with optimal nutrition as per recommendation for current healthy eating in Europe (Group 1). During the 4 weeks prior to the intervention, standard diets were given to all NutriTech participants to quantify their food intakes, in order to identify metabolomic biomarkers of food types, and maintain weight.

The standard diets contained similar energy and fibre intakes, but the energy provided by the macronutrients varied between weeks and between participants. Carbohydrate intake decreased from week 1 to week 3, but protein and fat intakes increased from week 1 to week 3. A detailed description of the macronutrient compositions in the food intake design is carefully laid out by Gibbons H., *et. al.* (Gibbons *et al.*, 2017). In this thesis, I have focused on the energy and macronutrient compositions at week 4 as the baseline; 50% energy from carbohydrates, 35% from fat, 15% from protein in Group 1, and 40% from carbohydrates, 45% from fat, 15% from protein in Group 2. Saliva, faecal and blood samples were taken before and after the dietary intervention period. The saliva samples were used to extract the genomic DNA of the participants (n=74). In addition, anthropometric (i.e. BMI, WHR) and clinical data (i.e. medication and clinical records), as well as truncal (neck to knee) magnetic resonance imaging (MRI) scans (i.e. adiposity compositions) were collected before and after the dietary intervention period. Informed consents were obtained from these participants to participate in the NutriTech project (Clinicaltrials.gov, 2016a). **Table 2- 8** shows the baseline characteristics of NutriTech participants in Group 1 (Dietary intervention) and Group 2 (Control) before the dietary intervention period started at the beginning of week 5. All body

fat (adiposity) measurements were indicated as adipose tissue in body weight percentage (%), unless stated otherwise.

| Characteristics | Group 1 | Group 2 |
|---|---|---|
| Number of participants (n) | 38 | 36 |
| Mean age (years) | 59.3 | 59.2 |
| Sex (n) | | |
| Male | 19 | 17 |
| Female | 19 | 15 |
| Missing data | 0 | 4 |
| Mean BMI at baseline (kg/m$^2$) | 29.0 | 29.3 |
| Male | 29.0 | 29.7 |
| Female | 29.1 | 28.8 |
| Mean WHR at baseline (ratio unit) | 0.91 | 0.92 |
| Male | 0.97 | 0.96 |
| Female | 0.85 | 0.87 |
| Obesity/Overweight status (n) | | |
| Obese (BMI ≥ 30.0 kg/m$^2$) | 15 | 20 |
| Overweight (25.0 < BMI < 30.0 kg/m$^2$) | 23 | 12 |
| Missing data | 0 | 4 |
| Mean total fat at baseline (%) | 38.2 | 35.8 |
| Male | 31.6 | 28.2 |
| Female | 44.8 | 44.4 |
| Mean subcutaneous fat at baseline (%) | 29.7 | 27.7 |
| Male | 22.2 | 19.8 |
| Female | 37.3 | 36.6 |
| Mean visceral fat at baseline (%) | 4.8 | 4.5 |
| Male | 5.7 | 4.9 |
| Female | 3.9 | 4.0 |

**Table 2- 8:** Baseline characteristics of the NutriTech participants at baseline (week 4). Group 1 = Diet intervention, Group 2 = Diet Maintenance (Controls), n = number; BMI = Body mass index, WHR = waist-to-hip ratio.

# 2.4. Personalised Medicine of Morbid Obesity cohort

## 2.4.1. Overview

Personalised Medicine of Morbid Obesity (PMMO) cohort is an observational study cohort that was largely implemented by Dr Sanne Alsters, as part of her Imperial College London PhD project entitled 'Genetic Analysis of Extreme Obesity' (Alsters, 2016). The PMMO project is registered in the UK Clinical Research Network (UKCRN) and ClinicalTrials.gov (Identifier: NCT01365416) (Clinicaltrials.gov, 2016b). It is an ongoing research study led by my supervisor, Professor Alexandra Blakemore, who is the principal investigator of the project. The PMMO project was created to address several aims; 1) to explore the genetic architecture of Type II Diabetes Mellitus (T2DM) and obesity in morbidly-obese patients seeking bariatric surgery, 2) to identify factors that could affect the outcomes of bariatric surgery, and 3) to increase our understanding of the mechanism underlying T2DM remission following bariatric surgery.

The PMMO project aims to recruit at least 2,000 obese-to-morbidly obese patients at the Imperial Weight Centre (London, UK), Chelsea and Westminster Hospital (London, UK) and Royal Derby Hospital (Derby, UK) (Clinicaltrials.gov, 2016b). At the Imperial Weight Centre, patients have been recruited by Dr Sanne Alsters, Dr Jennifer Murphy, Ms Olivia Szepietowski, Mr Erdal Ozdemir, and qualified clinical staff. The patient inclusion criteria are as below:

a) BMI > 28.0 kg/$m^2$
b) Age between 18 and 65 years old
c) Both sex

Patients who have donated blood within the last 3 months prior to their first visit were excluded, along with those who were participating in other clinical trials. Additionally,

patients who intended to, or were receiving medications that are not approved by the European Medicines Agency (EMA) were not included in this study (Alsters, 2016). The PMMO project was approved by the NHS National Research Ethics Service (NRES) London - Riverside (REC reference 11\LO\0935) (Alsters, 2016).

### 2.4.2. Data and sample collections

Various biological sources were collected from the selected patients including blood, faecal, saliva and urine samples. Consents were separately required for tissue samples (i.e. muscle, liver, subcutaneous and visceral fat) to be collected at the time of surgery in a subset of participants (Clinicaltrials.gov, 2016b). Biological and tissue samples were collected at several time points, as depicted in **Figure 2- 2**. The sample collections were carried out by Dr Sanne Alsters, Dr Jennifer Murphy, Ms Olivia Szepietowski, Mr Erdal Ozdemir, and qualified clinical personnel at the three assessment centres. This project has been conducted in accordance with the principles of the Declaration of Helsinki (Alsters, 2016).

Briefly, various anthropometric (i.e. anatomical measures such as height, weight, waist, hip, hand, head, BMI, WHR) and clinical data (i.e. sex, birth date, ethnicity, disease, surgery and medication records), as well as behaviour-related questionnaires (i.e. eating behaviour, psychological and mood questionnaires) were collected before any clinical assessments during the first visit (as baseline). When possible, anthropometric data were also collected during the next 6 studies (**Figure 2- 2**). Here, I only used the BMI and WHR at baseline, as the obesity traits under study in this thesis.

I have mentioned previously that the PMMO project is currently in progress. Hence, the total number of patients available for analysis was lesser than 2,000, which is the intended size for the PMMO project. All selected patients have submitted their written informed

consents to participate in the PMMO project(Alsters, 2016). To date, the DNA samples from 117 patients have been sent for SNP genotyping. The genotyping was carried out in two batches, of which the first batch of 21 DNA samples were genotyped with the NutriTech cohort. The remaining 96 DNA samples were genotyped as the second batch.

In this thesis, the term 'PMMO cohort' represents only the subgroup of 117 patients who had had their DNA genotyped at the time of analysis, as described. The baseline characteristics of the genotyped patients included in this thesis, is shown in **Table 2- 9**.

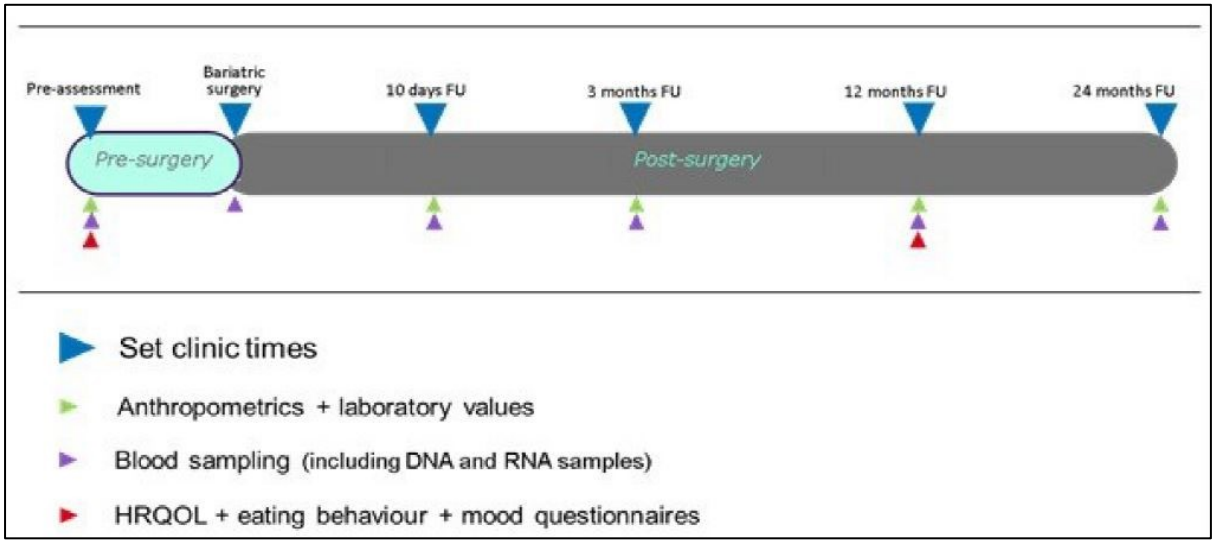


**Figure 2- 2:** The multiple time points of study visits by the patients selected for the PMMO project. New patients could also be recruited during any of these time points. FU = Follow-up, HRQOL = Health-related quality of life. PMMO = Personalised Medicine of Morbid Obesity. This figure has been adapted from Dr. Sanne Alsters's doctoral thesis (Alsters, 2016).

| Characteristics | PMMO cohort |
|---|---|
| Number of participants (n) | 117 |
| Mean Age (years) | 54.88 |
| Male | 56.41 |
| Female | 53.99 |
| Sex (n) | |
| Male | 41 |
| Female | 73 |
| Missing data | 3 |
| Mean BMI at baseline (kg/m$^2$) | 51.48 |
| Male | 51.92 |
| Female | 51.23 |
| Obesity/Morbid obesity status (n) | |
| Obese (30.0 ≤ BMI < 40.0 kg/m$^2$) | 11 |
| Morbidly obese (BMI ≥ 40.0 kg/m$^2$) | 99 |
| Missing data | 7 |

**Table 2- 9:** Baseline characteristics of the genotyped patients of the PMMO cohort at baseline (during the first visit pre-assessment). n = number; BMI = Body mass index, PMMO = Personalised Medicine of Morbid Obesity.

# 2.5. United Kingdom Biobank

## 2.5.1. Overview

The United Kingdom Biobank (UKBB) project, is a prospective study consisting of approximately 500,000 middle-to-old aged people (mean age = 66.45 years) living in the UK (UK Biobank, 2007). It is a resource for health research on a global scale, with numerous international projects using the UKBB data in a wide range of investigations, including in the field of genetics (UK Biobank, 2018). The UKBB project was initiated by the Wellcome Trust medical charity, Department of Health UK, the Government of Scotland, Medical Research Council (MRC), and Northwest Regional Development Agency (UK Biobank, 2018). Currently the UKBB project is led by Professor Rory Collins (the University of Oxford), as the principal investigator and Chief Executive Officer (CEO) of this non-profit making charity organisation. The overarching aim of the project is to improve the treatments, preventive and diagnostic measures of chronic and life-threatening diseases/health complications, such as cancers, stroke, cardiovascular diseases, diabetes, obesity, arthritis, osteoporosis, eye diseases, and mental disorders (UK Biobank, 2018).

From 2006 to 2010, the UK Biobank teams have recruited 502,616 individuals across the UK to participate in this project. The participants' age ranges from 40 to 69 years old. Written informed consents were submitted by the recruited participants. Ethical approvals for this project has been given by the North West Multi-centre Research Ethics Committee (MREC) in the UK, and by the Community Health Index Advisory Group (CHIAG) in Scotland (UK Biobank, 2018).

### 2.5.2. Data and sample collections

The participants of the UKBB project were invited to attend clinical examinations upon recruitment, during which they were asked to provide blood, saliva, and urine samples for future analysis (UK Biobank, 2018). A series of follow-ups were conducted to monitor their health and well-being. At every clinical examination, various data were collected from each participant, including anthropometry (i.e. comprehensive anatomical measures), biochemical data (i.e. insulin level and glucose levels), clinical information (i.e. sex, ethnicity, medication and disease history) and health-related questionnaires (i.e. quality of life, work, physical activities, lifestyle, psychological, and cognitive measures) (Clare Bycroft *et al.*, 2017, UK Biobank, 2007, UK Biobank, 2018). In addition, their diet and exercise habits were also monitored (Clare Bycroft *et al.*, 2017, UK Biobank, 2007, UK Biobank, 2018). Since the UKBB is an ongoing project, these data are being continuously collected and updated in a number of ways (Littlejohns *et al.*, 2017), including the extension of baseline information (Littlejohns *et al.*, 2017, Clare Bycroft *et al.*, 2017). Due to its exceptional sample size, this project has the potential to become a powerful resource for increasing our understanding of diseases from multiple fields of interests.

The latest release of the genotyped data consists of 488,377 participants of the UKBB project (~97% from the total participants recruited to date) (Clare Bycroft *et al.*, 2017). In this thesis, all analyses concerning the UKBB cohort, will only involve the participants who had their DNA sent for SNP genotyping (n=488,377). **Table 2- 10** shows the baseline characteristics of the genotyped participants.

| Characteristics | UKBB cohort |
|---|---|
| Number of participants (n) | 488,377 |
| Mean Age (years) | 66.45 |
| Male | 66.67 |
| Female | 66.28 |
| Sex (n) | |
| Male | 223,507 |
| Female | 264,859 |
| Missing data | 11 |
| Mean BMI at baseline (kg/m$^2$) | 27.42 |
| Male | 27.83 |
| Female | 27.07 |
| Mean WHR at baseline (ratio unit) | 0.87 |
| Male | 0.94 |
| Female | 0.82 |
| Obesity/Morbid obesity status (n) | |
| Obese (30.0 ≤ BMI < 40.0 kg/m$^2$) | 109,229 |
| Morbidly obese (BMI ≥ 40.0 kg/m$^2$) | 9,321 |
| Missing data | 1,992 |

**Table 2- 10:** The baseline characteristics of the participants (at mean age of 66.45 years) who had their DNA sent for SNP genotyping in the UKBB cohort. Adults with BMI between 30 kg/m$^2$ and 40 kg/m$^2$, were classified as obese. Adults with BMI of 40 kg/m$^2$ and above were classified as morbidly obese. n = number; BMI = Body mass index, WHR = waist-to-hip ratio, UKBB = United Kingdom Biobank.

# CHAPTER 3. General Methods

## 3.1. Introduction

This chapter is a compilation of detailed descriptions for all methods that were involved in this thesis.

For data harmonisation purposes, it is important to ensure that all cohorts had been subjected to comparable methodologies. The first section (**Chapter 3.2 Biological sample collection**) describes the methods used by local researchers to collect biological samples (whole blood, saliva) from individuals in each study cohort. The main genetic material used in this thesis was human genomic DNA (gDNA). The second section (**Chapter 3.3 Genomic DNA extraction**) describes the process of isolating the gDNA from the collected whole blood samples for each study cohort. The extracted gDNA samples were processed prior to SNP genotyping, as described in the fourth section (**Chapter 3.5 SNP genotyping**). Additionally, I have described the types and contents of the genotyping arrays used in each case. The genotyping of each study cohort was conducted by a third party, as stated in the text.

I applied QC processes to each SNP genotyping dataset except the UK Biobank (where QC was done by my colleague, Andrianos Yiorkas). SNP genotyping generated raw image data and raw signal intensity data; these data had to be processed and checked for the overall quality of the genetic information per individual. Thus, the fifth and sixth sections (**Chapter 3.6 Quality controls** and **Chapter 3.7 Population structure adjustments**) described the stages of quality control and population structure adjustment for each study cohort. In addition, cohort-specific principal components were estimated for each individual in the seventh section, and used as covariates for any association analyses (**Chapter 3.8 Principal component analysis (PCA)**).

Signal intensity data were used for CNV calling using a CNV calling algorithm tool, and each individual was assessed for overall signal quality in calling for CNVs, as described in the eighth and ninth sections (**Chapter 3.9 CNV calling** and **Chapter 3.10 CNV quality controls**), respectively. After identifying individuals who failed genotyping quality controls, inbreeding, cryptic relatedness and population outliers, these individuals were removed from the CNV catalogue for each study cohort. The individuals that were identified to have poor signal quality in CNV calling were removed as well, which yielded a high confidence and high quality CNV catalogue that is specific to each study cohort.

## 3.2. Biological sample collection

This thesis focuses on analysing the genetic information based on human gDNA. The gDNA of participants were taken either from whole blood or saliva. Below are the methods that had been used to collect the biological samples. I personally purified some DNA samples from the PMMO cohort, but all all samples were processed by the local study teams.

### 3.2.1. NFBC1966 and NFBC1986 cohorts

Upon recruitment, whole blood samples were drawn from each participant after an overnight fasting period. The fasting period started from 10 pm on the day of recruitment, and the blood sampling process was conducted between 8 am and 11 am in the following morning (Pouta *et al.*, 2004, Sabatti *et al.*, 2009, Taponen *et al.*, 2004). All blood samples were kept at -80°C in vacutainer blood collection tubes containing EDTA and anticoagulants, until

further analyses. The blood sampling process was done by qualified clinical staff from various clinical centres in Oulu and Lapland provinces, North Finland.

### 3.2.2. NutriTech cohort

All participants from the NutriTech cohort were given a standard diet for 4 weeks (28 days) to quantify their food intake and maintain their weight (O'Gorman *et al.*, 2017). They were then required to fast for at least 8 hours (overnight), starting from the evening end of the 28$^{th}$ day until the following morning. Saliva samples were collected from each participant on the first morning of the 5$^{th}$ week, using Oragene® DNA OG-500 collection kit (Genotek Inc., Canada), according to manufacturer's standard protocol. At this point, the saliva samples were considered to have been taken at baseline prior to the dietary intervention period. All saliva samples were kept at 21°C until analysed. The saliva sample collection was done by Eleanor Cropp, Milena Trundle, Claire Byrne and Katerina Petropoulou in NIHR/Wellcome Trust Imperial Clinical Research Facility at Hammersmith Hospital, London.

### 3.2.3. PMMO cohort

Participants in this study were seeking bariatric surgery for obesity. In the previous chapter, I have mentioned the different time points of study visits for PMMO cohort (**Figure 2- 2**). Patients had their whole blood samples drawn during pre-assessment (before their respective bariatric surgeries), on the day of the surgery, as well as during each post-surgery visit (10 days after, 3 months after, 12 months after, 24 months after). All participating patients were asked to fast for 1 hour prior to blood sampling. All blood samples were kept at -80°C in BD Vacutainer® $k_2$EDTA (3.6 mg) Blood Collection tubes, until analysed. The blood

sampling process was conducted by Dr Sanne Alsters, Dr Jennifer Murphy, and qualified clinical staff at the Imperial Weight Centre, London, UK.

If blood sampling could not be done, saliva samples were used as the biological samples. Saliva samples were collected from the participants using Oragene DNA OG-500 collection kit (Genotek, inc., Canada), according to the manufacturer's protocol. The saliva sampling process was conducted by Dr Sanne Alsters, Dr Jennifer Murphy, and qualified clinical staff at the Imperial Weight Centre, London, UK.

### 3.2.4. UK Biobank cohort

Briefly, 502,682 individuals living in the UK (recruitment age ranging from 40 to 69 years old; 5.5% response rate) were recruited for the UK Biobank (UKBB) project between 2006 and 2010 (Clare Bycroft *et al.*, 2017). Upon recruitment, the participants were required to attend for biological sample collection and various traits measurements at any of the 22 centres across the UK (Clare Bycroft *et al.*, 2017). Blood samples were collected from each participant using vacutainer blood collection tubes containing EDTA and anticoagulants. The collected blood samples were inverted 10 times to mix the contents. Two 10 ml vacutainer tubes were filled with whole blood from each participant (Welsh, 2017). They were held at 4°C in shipping boxes during transport to a central processing facility (Stockport, UK), and stored at -80°C until further analyses (Elliott *et al.*, 2008). The blood sample collection procedures were conducted by various collaborators involved in the UKBB project nationwide (Welsh, 2017, UK Biobank, 2007). The detailed methods were well described by Elliott, P, *et al.* (2008) and Welsh, S., (2017), and were showcased on the UKBB website (Elliott *et al.*, 2008, Welsh, 2017).

## 3.3. Genomic DNA extraction

Different gDNA extraction procedures were conducted for whole blood and saliva samples. Both procedures are explained as below. After each gDNA extraction procedure, gDNA samples were quantified at 260 nm light absorbance, and checked for quality using A260/280 nm ratio, using Spectrophotometer 120V (ThermoFisher Scientific Inc., USA) according to manufacturer's protocol. A gDNA sample is considered of good quality if its A260/280 nm ratio is between 1.8 and 2.0. Any A260/280 values lower than 1.8 could indicate protein contamination.

### 3.3.1. NFBC19866 and NFBC1986 cohorts

The gDNA for NFBC1966 (n=5,551) and NFBC1986 (n=4,043) participants were extracted from whole blood samples using standard DNA extraction methods (Bennett *et al.*, 2005, Bennett *et al.*, 2004, Sabatti *et al.*, 2009). Briefly, the blood cells were lysed, the proteins and then the gDNA were precipitated prior to the washing steps. Finally, the gDNA were suspended in a liquid form. The gDNA samples were stored at -80°C until analysed. The gDNA extraction procedures were performed by various laboratory personnel from the Department of Medicine, Imperial College London, UK.

### 3.3.2. NutriTech cohort

All saliva samples (n=74) were subjected to gDNA extraction procedures using Oragene® DNA OG-500 collection kit. The procedures were conducted according to manufacturer's standard protocol. Briefly, the saliva samples were incubated with nuclease and purified,

followed by gDNA precipitation, washing steps using ethanol, and hydrating the gDNA. This procedure was performed by Dr Sanne Alsters. After quality checks, good quality gDNA samples were kept at -20°C until analysed.

### 3.3.3. PMMO cohort

Two hundred twenty-four patients (n=244) had their gDNA extracted by LGC Genomics, UK, using standard DNA extraction methods. The rest of the cohort's gDNA samples were extracted by Dr Sanne Alsters, Hanis N. Ramzi, Andrianos Yiorkas (Department of Medicine, Imperial College London, UK), and myself, using the Gentra® Puregene® Blood Kit (Qiagen Inc., Hilden, Germany) according to manufacturer's standard protocol. After gDNA quantification and purity checks, the gDNA samples were stored at -20°C until analysed.

### 3.3.4. UK Biobank cohort

Whole blood samples (n=494,078) were retrieved by an automated sample retrieval system, thawed at room temperature for 2 to 3 hours, and plated into 96-well plates (94 UKBB samples and 2 controls from 1000 Genomes project, per plate) (Clare Bycroft *et al.*, 2017, Welsh, 2017). The robotic sample retrieval process was performed based on a sample selection algorithm to make sure that each plate would have a combination of samples from different centres, sex, ethnicity, age, and other baseline traits (Clare Bycroft *et al.*, 2017). The gDNA was extracted via a customised DNA extraction system (Welsh, 2017) (32 samples per batch, and 16 samples per batch for UK BiLEVE projects) using modified Promega® Maxwell® 16 AS2000-HS DNA Purification instruments and kits(Welsh, 2017). The gDNA was quantified (as absorbance at 260 nm) and checked for purity (as A260/280 ratio) using Trinean DropSense® 96 Multi-channel Spectrophotometer, according to manufacturer's

protocols(Welsh, 2017). The extracted gDNA samples (n=494,078) were kept in bubble wrappers and stored at -80°C prior to sending them for genotyping. The DNA extraction work was conducted by accredited laboratory personnel at the UKBB central processing facility in Stockport, UK (Welsh, 2017). Detailed explanation of the methods, criteria and workflows was provided by Welsh, S. (2017) (Welsh, 2017).

## 3.4. Obesity-related phenotypic measurements

### 3.4.1. Body mass index and waist-to-hip ratio

This section is applicable to all study cohorts described in **Chapter 2**. Height, in centimetres (cm), and body weights, in kilograms (kg), were determined using standard measurement procedures. Body mass index (BMI) was calculated as $kg/m^2$. Waist-to-hip ratio (WHR) is defined as the ratio of the circumference of the waist to that of the hip. Waist and hip circumferences were measured in centimetres (cm) using standard measurement apparatus. The WHR was calculated by dividing the waist circumference by the hip circumference.

**a) NFBC1966 and NFBC1986 cohorts**

Primarily, the anthropometric data (including height and weight) were collected from the parents, and from their offspring born in 1966 and in 1986. The data collected were from the prenatal period, at birth and periodically throughout life. Since both NFBCs are longitudinal cohort studies, the data for the offspring continued to be collected; 6 months (NFBC1966), 1 year (both), 7-8 years (NFBC1986), 14 years (NFBC1966), 16 years (NFBC1986), 31 years (NFBC1966), and 41 years (NFBC1966) after their birth. For the purposes of this PhD, I

analysed the BMI and WHR at 31 years old (NFBC1966) and at 16 years old (NFBC1986) respectively.

**b) NutriTech cohort**

Prior to the start of standard diet consumption for four weeks, participants were initially screened *via* telephone interview. They were then invited to visit the NIHR/Wellcome Trust Imperial Clinical Research Facility at Hammersmith Hospital, London for health screening. Various anthropometric data were collected, including height and weight. In addition, information on lifestyle, medication and medical history of each participant were collected. The data from this health screening visit were considered as baseline. After the 12-week dietary intervention period had been completed, the same traits were measured again to assess the effects of dietary intervention in NutriTech cohort. For this PhD, I looked at the baseline BMI and WHR, and response to the dietary intervention.

**c) PMMO cohort**

Height and weight data were collected for patients from all study visits, when possible; before surgery visit, bariatric surgery day prior to surgery, then 10 days, 3 months, 12 months and 24 months after the surgery. The data from before the surgery was considered as baseline. For this PhD, I analysed the BMI at baseline (i.e. during the first visit) for the PMMO patients.

**d) UK Biobank cohort**

Standard anthropometric data, including heights (cm) and weights (kg) were collected for all participants by accredited personnel during clinical assessment visits at their respective centres across the UK (Clare Bycroft *et al.*, 2017). The data was considered as baseline.

### 3.4.2. Adiposity

This section is applicable to the NutriTech and UKBB cohorts only. Adiposity composition is defined by the excess accumulation of adipose tissue (AT) either in the whole human body, or in specific parts of human body. For this PhD, I focused on AT accumulations in the following parts of human body:

- Abdominal fat (total AT)
- Within the abdominal cavity, wrapping the organs (visceral AT)
- Under the skin (subcutaneous AT)
- Within the pancreas (pancreatic AT)

Professor Jimmy Bell (Westminster) led the teams using imaging methodology to quantitate fat deposition in the abdominal region of subjects from the NutriTech cohort and the UKBB. The same methodology was used in each case. Briefly, magnetic resonance imaging (MRI) was conducted to obtain magnetic resonance images using Philips Achieva 1.5T MRI System scanner (Philips Inc., the Netherlands) and analysed using a medical image analysis software SliceOmatic (Tomovision, Quebec, Canada) (Thomas *et al.*, 2012, Thomas *et al.*, 1998). The detailed procedures of the MRI scanning have been previously described (Thomas *et al.*, 2012, Thomas *et al.*, 1998). Ten mm-thick transverse images were attained all over the body per participant. Total, visceral and subcutaneous AT were measured as volumes in litres (L), and as a percentage of body weight (%). Only the total abdominal fat, visceral and subcutaneous AT % were used as measures of obesity in this thesis. Subsequently, pancreatic AT composition was measured using proton magnetic resonance spectroscopy (1H-MRS) from the same 1.5T scanner (content value calculated as a ratio of CH2/water) (Thomas *et al.*, 2012, Thomas *et al.*, 1998). Another scan for pancreatic AT was conducted using multiple spin echo

magnetic resonance imaging sequences to acquire the ratio of AT content/AT+water (Thomas *et al.*, 2012). For this thesis, the value of AT content/AT+water in each individual was used as the pancreatic fat measurement. All adiposity measurements in NutriTech cohort were conducted by Jimmy Bell, Louise Thomas, and their dedicated team at the University of Westminster, London. For the UKBB, scanning was carried out by local teams and MRI data analysis by a company, AMRA Medical AB (Sweden).

## 3.5. SNP genotyping

### 3.5.1. Arrays used in different cohorts

Here I provided a brief description of the SNP genotyping processes cohort-by-cohort, due to the differences in batch separation, microarray types being used, and the number of samples involved. The genotyping outputs that I received were in the form of raw image data for each genotyped sample.

**a) PMMO**

A total of 117 gDNA samples from the PMMO cohort were genotyped in two batches (Batch 1, n=21 plus 74 NutriTech samples; Batch 2, n=96). Genotyping for Batch 1 was performed on Illumina Infinium® Human OmniExpress-24v1.0 microarrays using Illumina Infinium High Throughput System (Illumina Inc., San Diego, California, USA). The arrays contain ~710,000 optimised tag SNPs as array markers, providing a genome-wide coverage for common variations in humans. Batch 2 was genotyped on Illumina Infinium® OmniExpressExome-8v1.3 microarrays using the same System (Illumina Inc., San Diego,

California, USA). The arrays have an additional >273,000 markers in coding (exonic) regions of the human genome to provide the best capture of putative, functionally important variants. The genotyping of both batches was conducted by High Throughput Genomics, Wellcome Trust Centre for Human Genetics, University of Oxford. The positions of the markers on this array are reported according to the positions of Genome Reference Consortium Human Reference 37 genome build (GRCh37/hg19).

**b) NutriTech**

As described above, all 74 individuals from the NutriTech cohort had their normalised gDNA samples sent for genotyping as part of PMMO's Batch 1 of genotyped samples. In order to have an exclusively NutriTech genetic data, all PMMO samples were excluded prior to post-processing of the raw genetic data.

**c) NFBC1966**

Genotyping was initially performed on the Illumina Infinium® HumanCNV370-Duo v1.0 DNA Analysis microarrays (Illumina Inc., San Diego, California, USA). A total of 5,557 gDNA samples were genotyped by the Department of Medicine, Imperial College London. The arrays contain ~310,000 SNPs tagged as common markers, with an additional <50,000 probes for known copy number regions. On average, >330,000 markers were genotyped for NFBC1966 cohort.

The positions of the markers on this array were reported according to the positions of Genome Reference Consortium Human Reference 36 genome build (GRCh36/hg18). For uniformity, the genotyped dataset for this study cohort was updated to genome build hg19.

**d) NFBC1986**

Genotyping was performed using the Illumina Infinium® HumanOmniExpressExome-8v1.2 arrays (Illumina Inc., San Diego, California, USA) by a collaborative effort from the Department of Medicine, Imperial College London, and High Throughput Genomics, Wellcome Trust Centre for Human Genetics, University of Oxford. A total of 4,043 human gDNA samples (Batch 1, n= 2,016; Batch 2, n=2,027) were genotyped using Illumina Infinium HD Super assay system (Illumina Inc., San Diego, California, USA). The array content comprised optimised SNP markers with genome-wide coverage, including exonic regions to capture putative, functionally-important variants. On average, >950,000 SNPs were genotyped from each batch. The positions of the markers on this array were reported according to the positions of genome build hg19.

**e) UK Biobank cohort**

Briefly, the SNP genotyping of the UKBB cohort was performed on two types of Affymetrix arrays; Affymetrix UK BioBank Axiom® Arrays (n=~443,568) and UK BiLEVE Axiom Arrays (n=50,561), both by the Affymetrix Research Services Laboratory (ARSL) (Affymetrix Inc., Santa Clara, California, USA) (Clare Bycroft *et al.*, 2017). The latter array was initially designed for the UK BiLEVE project which focuses on the genetics of lung health. The ~50,000 participants from the UK BiLEVE project were not randomly selected from the general UK population as they were selected based on their smoking behaviours and lung functions (Wain *et al.*, 2015). In terms of array content, the two arrays have approximately 95% similarity for common markers, although the former is an updated version of the latter. The UK BioBank Axiom® Array content comprises >110,000 rare coding variants, >620,000 genome-wide

SNPs, and >90,000 markers designed to capture variants of specific interests, such as cancer-related markers (UK Biobank, 2014). The detailed genotyping procedures and workflow, array content and genetic data generation were described elsewhere (UK Biobank, 2007, Clare Bycroft *et al.*, 2017). The positions of the markers on both arrays were reported according to the positions of genome build hg19. The data from both arrays were combined, amounting to a total of 805,426 markers being genotyped.

### 3.5.2. Processing of genotyped data

It is important to process the genotyped data in a pipeline where false positives and poorly genotyped samples could be identified and removed. This provides high-quality genetic data for use in downstream analyses. The SNP genotyping outputs were provided to me as a compilation of raw image files per sample. Here I describe the detailed steps of converting the raw genetic outputs from SNP genotyping into a format that can be read and analysed by other commonly used tools for downstream genetic analyses. This section also describes the detailed steps in the quality controls and adjusting for the population structures of the study cohorts involved. Overall, I performed the whole processing of genotyped data for the PMMO, NutriTech, NFBC1966, and NFBC1986 cohorts, according to **Figure 3-1**.

After SNP genotyping had been completed, the raw genotype data were in image data format. In order to convert the raw image data into readable genetic data, software that is compatible with the genotyping array platform has to be downloaded and used. The PMMO, Nutritech, NFBC1966 and NFBC1986 cohorts were genotyped on Illumina arrays, so the GenomeStudio Version 2011.1 (Illumina Inc., San Diego, California, USA) was used for processing of the raw image data. The genotyped samples were processed cohort-by-cohort.

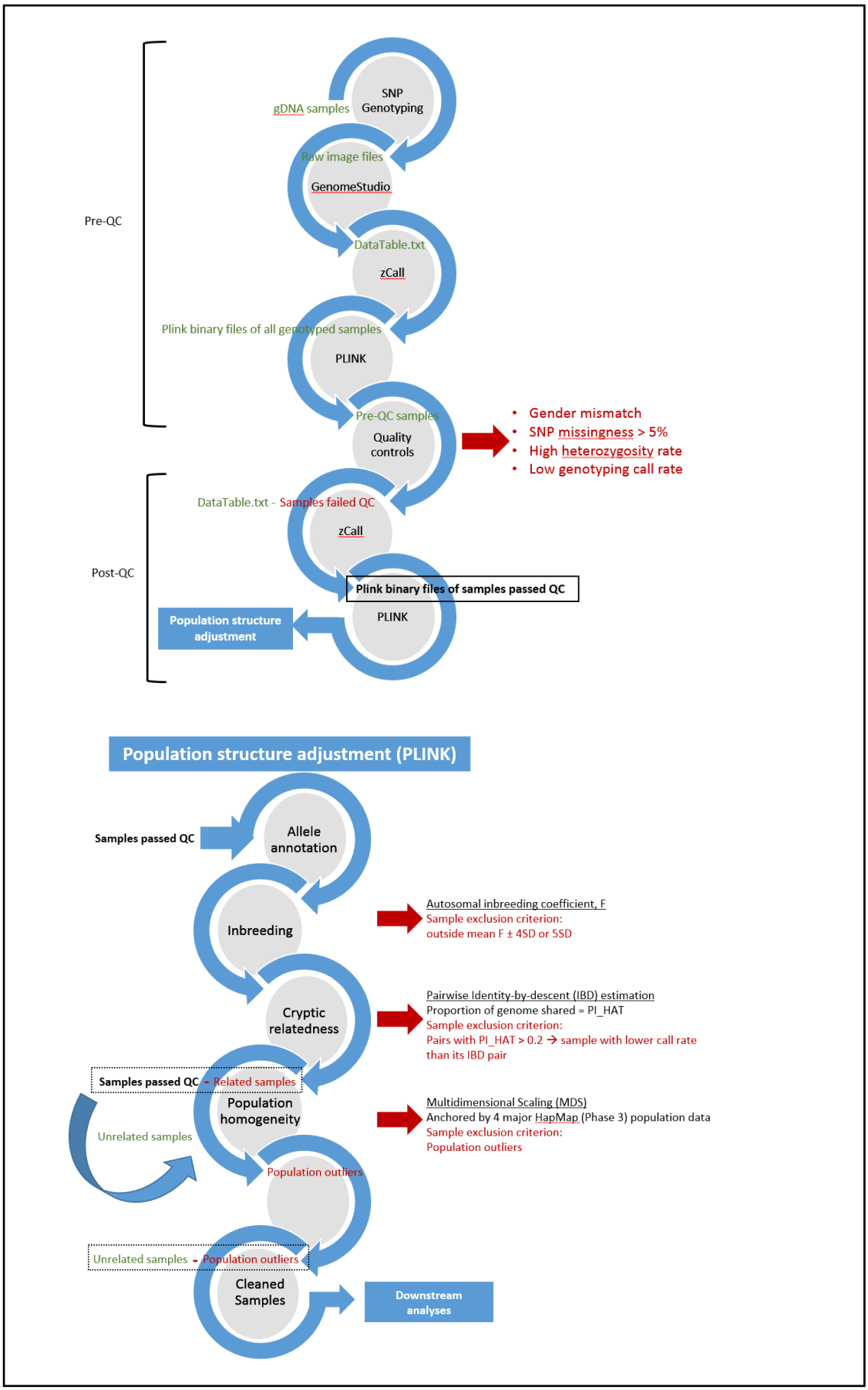


**Figure 3-1:** Analysis workflow in processing SNP genotyping data

Prior to loading the raw image data into the GenomeStudio software, a SNP manifest file and a cluster file specific for the array chip were obtained from Illumina website (URL: https://support.illumina.com/downloads.html). The files are listed as below:

i. PMMO and Nutritech cohorts (Genome-wide plus exons):
   - HumanOmniExpress-24 v1.0 Manifest File (BPM Format)
   - HumanOmniExpress-24 v1.0 Cluster File (EGT Format)

ii. NFBC1966 cohort (Genome-wide):
   - HumanCNV370 Manifest File (BPM Format)
   - HumanCNV370 Cluster File (EGT Format)

iii. NFBC1966 cohort (Exon-specific):
   - HumanExome-12 v1.1 Manifest File (BPM Format)
   - HumanExome-12 v1.1 Beadchip Cluster File (EGT Format)

iv. NFBC1986 cohort (Genome-wide plus exons):
   - HumanOmniExpressExome-8 v1.2 Manifest File (BPM Format)
   - HumanOmniExpressExome-8 v1.2 Cluster File (EGT Format)

All the cohorts mentioned above underwent the same steps in GenomeStudio software. Briefly, in the 'Genotyping' module, I created a new repository directory for each cohort and 'load sample intensities' option was selected. Since I had already prepared the necessary files, I loaded the SNP manifest file as listed above. In order to locate the raw image data, I directed the path for the image data directory in the 'Image Data Repository' option. From here, I was able to add all the relevant image data folders with the file names reflecting the chip numbers. I also selected 'Import Cluster Positions' option to load the specific cluster file listed above.

For each SNP of a sample, a GenCall score was calculated to act as an accuracy parameter in determining the most probable genotype. Based on the clustering of the SNPs, the GenCall

score decreases with increasing distance from the middle of the cluster. For this PhD, I used a no-call threshold of 0.15, which is the most commonly used threshold in Infinium data analysis (Illumina, 2014). Any SNP with GenCall score below 0.15 were denoted as 'NC' (No Call). Since the GenCall scores are heavily depending on the clustering of SNPs and their frequencies in the cohort, rare variants such as exonic variants typically falls below the GenCall threshold. However, I overcame this issue by using a tool called 'zCall', in which will be described later in this thesis.

Back in GenomeStudio software, I proceeded with calculating sample and SNP statistics. These steps were uniformly implemented across different cohorts, and each cohort was processed independently from one another, with the exception of the NutriTech cohort since both PMMO and NutriTech cohorts (participants were all London-based) were genotyped together. At this point, I did not exclude any genotyped SNPs or any samples from analysis, because I wanted to include everything in the following steps for quality controls (QC).

Using GenomeStudio software, I generated two types of output tables from the Full Data Table tab; (1) 'For-QC Table' (2) 'For-CNV Table'. The former was generated by selecting Name, Chr, Position, all samples as 'displayed columns' and GType, X, Y as 'displayed subcolumns'. The latter was generated by selecting Name, Chr, Position, all samples as 'displayed columns' and GType, Log R Ratio, B Allele Freq as 'displayed subcolumns'. Both tables were text-based in tab delimited format. As the names implied, 'For-QC Table' was generated for QC purposes and For-CNV Table was generated for CNV calling purposes (refer **Chapter 3.9 CNV calling**).

### 3.5.3. Access to the genotyped data of UK Biobank cohort

The genotype calls (in binary PLINK format; .bed, .bim. and .fam files), Log R Ratio values (LRR), and B Allele Frequency values (BAF) of the markers on all chromosomes were systematically downloaded from the UKBB server. The downloaded data corresponded to 488,377 participants, with no genotype data available for the remaining 2.8% of the UKBB participants, either due to the internal QC by Affymetrix, the participant being of non-white British background, or insufficient quantity of gDNA (Clare Bycroft *et al.*, 2017). The data download process was performed in a Linux/Unix environment according to the instructions given on the UKBB website, and described elsewhere (Clare Bycroft *et al.*, 2017).

## 3.6. Quality controls

---

All the quality control (QC) work from **Chapter 3.6.1 Quality control 1: Sex discordance** to **Chapter 3.6.5 Rare variant calling** was performed by myself to generate clean genetic datasets for NFBC1966, NFBC1986, NutriTech and PMMO cohorts. Most of the steps for the QC were conducted in Linux/Unix environment using PLINK Version 1.07 (Purcell *et al.*, 2007).

The text-based 'For-QC table' had to be converted into PLINK-readable format prior to QC. I used a python script called `convertReportToTPED.py` from the zCall tool to convert the 'For-QC Table' into transposed binary format (Goldstein *et al.*, 2012). The zCall tool is a free python-based tool to call rare variants from array-based genotyping (URL: https://github.com/jigold/zCall) (Goldstein *et al.*, 2012). Within the same zCall directory, a command line was executed as below:

```
python convertReportToTPED.py –R For-QC_Table.txt –O For-QC_Table
```

This generated `For-QC_Table.tfam` and `For-QC_Table.tped` (Transposed binary format). Both files were relocated into PLINK directory and an additional command line was executed to convert transposed binary format into a standard binary format, by using the command line as below:

```
./plink --noweb --tped For-QC_Table.tped --tfam For-QC_Table.tfam --make-bed --out For-QC_Table
```

This generated `For-QC_Table.bed` (genetic binary data), `For-QC_Table.bim` (SNP information) and `For-QC_Table.fam` (individual data containing sex and phenotype information). Since I have the sex information from the clinical data for most of the individuals, I added the actual sex (coded as male=1, female=2, missing=-9) by manual editing of the `For-QC_Table.fam`. From here on, I will refer `.bed`, `.bim` and `.fam` together as PLINK-binary files.

### 3.6.1. Quality control 1: Sex discordance

The first step in QC is to check whether the self-reported sex information is in concordance with the sex estimated from the genomics data. This step is important as it could highlight plating error or sample mix-up.

Since males only have one copy of the X-chromosome, all SNPs on the X-chromosome are hemizygous, but called by the genotyping software as homozygous. Thus, it is expected that males to have a mean homozygosity rate = 1 for the X-chromosome. However, I have included rate variations of ± 0.2 due to possible genotyping errors. Therefore, samples were determined as males if their X-chromosome mean homozygosity rates ≥ 0.8. Females with two copies of X-chromosomes are expected to have very much lower homozygosity rates. In

this thesis, females were determined if their X-chromosome mean homozygosity rates fall between 0 and 0.2, which is the standard threshold suggested by PLINK (URL: http://zzz.bwh.harvard.edu/plink/summary.shtml#sexcheck).

The `QC_Table` PLINK-binary files were already in the PLINK directory. The mean homozygosity rate across all the SNPs on X-chromosome were calculated for each sample by using the command line as below:

```
./plink --noweb --bfile QC_Table --check-sex --out QC_Table.sex
```

This generated several sex-related output files, including `QC_Table.sex.sexcheck`. This file contained the list of all samples that had undergone comparison between self-reported sex and its genetically-estimated sex. Any samples that had sex discordance between the two data types were denoted as 'PROBLEM' in the concordance status column (column 5). Those samples were isolated in `QC_Table.sex.sexprobs` using the command line as below:

```
grep PROBLEM QC_Table.sex.sexcheck > QC_Table.sex.sexprobs
```

In addition, any samples that had mean homozygosity rates between 0.2 and 0.8 were considered as having ambiguous sexs. This was examined by extracting the X-chromosome mean homozygosity rates (Fx) for all samples in `QC_Table.sex.sexcheck` to build a scatter plot of Fx in an Excel spreadsheet. Those samples with ambiguous sex, including the ones isolated in `QC_Table.sex.sexprobs` were included in a text-based list of 'failed samples'.

### 3.6.2. Quality control 2: Missing genotype data, high autosomal heterozygosity rates

The next QC steps were to identify samples that (1) had an unexpectedly high proportion of their SNPs not genotyped (considered as missing), and/or (2) had elevated mean heterozygosity rates across all autosomal chromosomes. The accuracy and rates of genotype calls depends heavily on the quality of the DNA samples. In normal circumstances, low DNA quality or concentration would lead to below-than-average sample call rates and genotype accuracy. Therefore, the samples that had either too high genotype missing data or an excessive/reduced proportion of autosomal heterozygous genotypes were considered to have failed. The latter might indicate DNA contamination (excessive) or inbreeding (reduced).

In the PLINK directory, lists of individual (`.imiss`) and loci (`.lmiss`) missingness rates were generated from the `QC_Table` PLINK-binary files, by executing the command line as below:

```
./plink --noweb --bfile QC_Table  --missing --out QC_Table.missing
```

The `QC_Table.missing.imiss` file contained information on the proportion of missing SNPs per individual (F_MISS), and the number of missing SNPs per individual (N_MISS). In order to calculate heterozygosity rate per individual, a command was executed in PLINK:

```
./plink --noweb --bfile QC_Table  --het --out QC_Table.hetero
```

The resulting `QC_Table.hetero.het` file contained information regarding observed (O (Hom)) and expected (E (Hom)) homozygous genotypes, as well as the number of non-missing genotypes (N(NM)). Autosomal heterozygosity rate (Het) was calculated for each individual using the formula as below:

Het = [N(NM) – O(Hom)] / N(NM)

A scatter plot of F_MISS against Het was created for each cohort. The failure rate for F_MISS is typically >3-7% across different populations. For the work described in this thesis, a 5% missingness rate was tolerated (F_MISS > 0.05). The threshold level for Het depends upon visual inspection of Het clustering in the plot. For the work described in this thesis, a threshold level of 4 standard deviations (4SD) away from the mean of Het was used across all study cohorts. All individuals who were located outside of the thresholds specified on the scatter plot, were included in the list of 'failed samples'.

### 3.6.3. Quality control 3: Sample call rates

Sample call rate information was extracted directly from the GenomeStudio software for each study cohort. It was generated as a tab-delimited, text-based 'Sample Table' for all the samples in each dataset. The sample call rates were filtered in an Excel spreadsheet according to a user-specified threshold. In this thesis, any individuals with sample call rates < 98% were considered as failed samples and were included in the 'failed samples' list. An exception was considered for individuals in NFBC1966 cohort, in which its failed individuals had sample call rates < 0.97. This consideration was decided upon the relatively poor overall quality of the SNP genotyping of NFBC1966 cohort.

### 3.6.4. Sample exclusions

All individuals who failed at least one of the three QC were compiled in the 'failed samples' list, and then removed from 'For-QC Table' using zCall tool. Although PLINK has the function to remove individuals from PLINK-binary files, I used the zCall tool to remove them from the

'For-QC Table' instead. This is because zCall tool only accepts an input format such as in 'For-QC Table'.

The 'For-QC Table' and 'failed samples' files were first relocated into the zCall directory. In zCall directory, an updated python script for excluding samples from GenomeStudio-exported report (In this case, 'For-QC Table') was downloaded by executing the command line below:

```
wget
https://github.com/jigold/zCall/blob/master/additionalScripts/dropSamplesFrom
Report_FasterVersion.py
```

The updated script is called `dropSamplesFromReport_FasterVersion.py`. In order to make sure the total number of individuals in the 'For-QC Table', the columns were counted using the command line below:

```
awk '{print split($0,a,"\t"); exit}' For-QC_Table.txt
```

The number should be 3 + (3 x n), where n = total number of individuals in a cohort.

The failed individuals were then removed from 'For-QC Table' by executing `dropSamplesFromReport_FasterVersion.py` script as shown below:

```
python dropSamplesFromReport_FasterVersion.py For-QC_Table.txt
failedsamples.txt > QCed_zcall
```

The output (`QCed_zcall.txt`) contained all individuals who passed the QC in a study cohort. Prior to rare variant calling steps, the columns were counted again to make sure that the failed individuals have been properly removed in the cleaned output:

```
awk '{print split($0,a,"\t"); exit}' QCed_zcall.txt
```

The number should be 3 + (3 x n), where n = total number of individuals who passed QC in a cohort.

### 3.6.5. Rare variant calling

For the work described in this thesis, variants from SNP genotyping were initially assigned with specific genotypes based on the clustering of the alleles through the processing of raw image data in GenomeStudio (for Illumina arrays). Hence, the specificity and the sensitivity of the genotype assignment depends on how well different individuals with the same genotype cluster together as a specific genotype. This can be true for variants that are common in population frequency, but rarer variants usually cluster poorly with each other due to the lack of carriers in the population. Consequently, the normal method of genotype calling might provide false and non-calls (NC) for such rare variants within the same population. For calling rare variants with higher specificity and sensitivity, I used zCall tool (Goldstein *et al.*, 2012).

The zCall tool uses the intensity information of the most common allele of its homozygote cluster (assigned previously by either GenomeStudio or Affymetrix Power Tools) to generate a threshold along the x and y axis (two thresholds altogether). This results in 4 dividing quadrants within the x and y axis space. The allele A is represented on the x axis, while allele B is represented on the y axis.

Prior to using zCall tool, the R version 3.3.1 package was downloaded and installed under Linux platform, as per the developer's manual (Team, 2013). An executable file called `Rscript` can be found in the bin directory of the newly installed R directory. The `Rscript` file was copied into the zCall tool so that it could be used later.

**a) Calculating the means and SDs**

To begin, a background model has to be created for all the NC variants. This was done by creating homozygote clusters with genotype calls using stringent variant inclusion criteria; variant call rates > 99%, variants with minor allele frequency (MAF) > 5%, and variants with P value for Hardy-Weinberg Equilibrium ($P_{HWE}$) > 1 x $10^{-5}$. The command line used to execute this step is below:

```
python findMeanSD.py -R QCed_zcall.txt > QCed_zcall_meanSD.txt
```

The means and SDs of the clusters were calculated based on common sites (with no less than 10 points present in each cluster) and no significant deviation from HWE.

**b) Calculating the beta coefficients**

The next step was to calculate the beta coefficients from the means and SDs of x and y for all the variants. This was done by executing the commands below:

```
chmod a+x Rscript
./Rscript findBetas.r QCed_zcall_meanSD.txt  QCed_zcall_betas.txt 1
```

The argument '1' was included to use weighted linear regression rather than the regular one (coded as '0'). Weighted linear regression was used to account for small sample sizes. The weightage calculation is 1/sV[(1/nAA) + (1/nBB)], where nAA=number of homozygote A genotypes and nBB=number of homozygote B genotypes.

**c) Finding the thresholds for x- and y-axis**

For each variant, two thresholds were calculated (denoted as T_x and T_y). This was done by executing the commands as below:

```
python findThresholds.py -B QCed_zcall_betas.txt -R QCed_zcall.txt -Z 7 -I 0.2
> QCed_zcall_thres7.txt
```

The argument 'Z=7' shows the z-score that with the highest global concordance between the genotype calls made by the primary genotype calling algorithms (in this case, Illumina Infinium GenCall) and the z calls. The argument 'I=0.2' shows the minimum value of the mean signal intensity for any common sites to have a z call assigned. The values for both optional arguments were as recommended by the developer (Goldstein *et al.*, 2012). The two thresholds derived from Z=7 and the beta coefficients (thresholds=T7) were used to execute zCall and replace the NCs with genotypes.

**d) Executing zCall**

The non-calls (NCs) were replaced with genotypes that were assigned based on the locations of the points within the specified thresholds (in this thesis, all cohorts used T7 as the thresholds).

```
python zCall.py -R QCed_zcall.txt -T QCed_zcall_thres7.txt -O QCed_popstrat
```

The output files generated were in the transposed PLINK format (`QCed_popstrat.tped` and `QCed_popstrat.tfam`). Both files were relocated into the PLINK directory.

### 3.6.6. Quality controls in the UK Biobank

The sample QC for the UKBB cohort were conducted by my colleague, Andrianos Yiorkas (Department of Medicine, Imperial College London, UK) to replicate the QC results of Bycroft, C., *et. al.* (2017) (Clare Bycroft *et al.*, 2017).

A subset of high quality SNPs with MAF>0.01% (n=621,642; 2.5% were on the sex chromosomes) was initially selected to identify genotyped samples with poor genotyping call rates, genetically-related samples, samples with high proportion of SNPs missing, samples with elevated autosomal heterozygosity rates, and samples who had discordance between self-reported and genetically-inferred sex.

Briefly, the high quality SNPs on autosomal chromosomes (n=605,876) were used to calculate heterozygosity rates (--het command) and missingness rates (--miss command) per individual in PLINK Version 1.07 (Purcell *et al.*, 2007, Clare Bycroft *et al.*, 2017). The procedures were executed as in **Chapter 3.6.2 Quality control 2: Missing genotype data, high autosomal heterozygosity rates**. The heterozygosity rates were adjusted for the first six principal components from the PCA to obtain ancestry-corrected heterozygosity rates for each individual. This adjustment was done to control for population bias (Clare Bycroft *et al.*, 2017). Samples with extreme heterozygosity rates, and either had SNP missingness >5% or genotyping call rates >95%, were flagged as outliers. The exact methods and calculations for the outlier detection were described in detail elsewhere (Clare Bycroft *et al.*, 2017).

The cohort was subsequently filtered to only include samples with white British ancestry to avoid a strong confounding effect from the population structure. The filter included samples

with self-reported 'British' and had genetically similar ancestry based on the principal component analysis (PCA). The detailed procedures were as described by Bycroft, C., *et. al.* (2017) (Clare Bycroft *et al.*, 2017).

For the sex check, samples with sex chromosome aneuploidy were initially detected by using the mean LRR for each individual, from the high-quality SNPs on chromosome X and Y (n=15,766). The mean LRRs for chromosome X and Y were plotted against each other, and the scatter plot was visually inspected. Any samples with putative aneuploidy were flagged and removed from further analysis. For the rest of the samples, the sex check was performed using the --checksex command in PLINK Version 1.07 (Purcell *et al.*, 2007). The procedures were executed as described in **Chapter 3.6.1 Quality control 1: Sex discordance**. The detailed methods and criteria were explained by Bycroft, C., *et. al.* (2017) (Clare Bycroft *et al.*, 2017). Additionally, samples without a valid identification (ID) were removed prior to the population structure adjustment steps. At the end of this process, each study cohort is expected to have a clean and high-quality genetic dataset that is ready for population structure adjustments.

## 3.7. Population structure adjustments

Different populations contain different genetic structures and allele frequencies. Within each population, these differences are important in studying complex disorders such as obesity. Population substructure and ethnic admixture could reduce power in genetic studies, as well as leading to false-positive discoveries. Such artefactual discoveries could be due to the representation of genetic differences from ancestral background rather than the presence of disease-causing or disease-associated genes. For the work described in this thesis, I carried out various corrections for population structure in NFBC1966, NFBC1986, NutriTech, and PMMO cohorts, separately. The analyses were conducted in PLINK, unless stated otherwise. By the end of this process, each study cohort is expected to have a clean genetic data that had been adjusted for population structure.

### 3.7.1. A, T, C, G allele conversion

All alleles were initially classified as A and B in a genotype instead of the actual A, T, C and G alleles. Therefore, it is essential to convert the A and B alleles to the usual A, T, C and G alleles for all the genotypes of the SNPs presented on the arrays in each cohort prior to correcting for population structure.

In the PLINK directory, the conversion files that contain the allele annotations were downloaded and extracted from a website dedicated to providing genotyping chip strand and build files (http://www.well.ox.ac.uk/~wrayner/strand/). The specific files were obtained for the genotyping arrays used in each case.

The alleles were converted by executing the command line as below:

```
./plink --noweb --bfile QCed_popstrat --update-alleles
[arraychipname].update_alleles.txt --make-bed --out QCed_popstrat.atcg
```

The output files generated were `QCed_popstrat.atcg` PLINK-binary files. In addition, actual sex (coded as male=1, female=2, missing=-9) was manually edited into `QCed_popstrat.atcg.fam`.

### 3.7.2. Adjustment 1: Inbreeding

In an ideal population where Hardy-Weinberg Equilibrium (HWE) is true, the mating behaviour is assumed to be random. However, real population data may have a certain degree of deviation from random mating. Inbreeding, which is a non-random mating, is one of the causes for this deviation. Inbreeding occurs when there is an increased possibility of individuals mating with relatives (up to 3$^{rd}$ degree relatives), rather than with other individuals at random, within a specific population. This non-random mating increases the probability of having homozygous offspring, eventually increasing the number of homozygous genotypes within the genetic pool of the population. Increased homozygosity can be detrimental to the overall fitness of the population species, and a massive increase can lead to a complete infertility after only a few generations. In terms of genetic studies, it is crucial to identify individuals with high levels of inbreeding and remove them from further analyses to avoid any confounding bias.

In this thesis, the inbreeding check was conducted by calculating the inbreeding coefficient (F) of each individual within a study cohort, based on the observed versus expected number of homozygous genotypes on autosomal (non-sex) chromosomes. This was done twice;

before and after adjusting for population structure, to make sure of the homogeneity of the study cohort.

Since all the study cohorts were genotyped on arrays with genome-wide coverage, the inbreeding check was applied only to a subset of SNPs that were initially pruned to be in approximate linkage equilibrium. The pruning step was executed using the command line below:

```
./plink --noweb --bfile QCed_popstrat.atcg --indep-pairwise 50 5 0.2 --out QCed_popstrat.atcg.inbred
```

The analysis was based on the squared correlations ($r^2$) between pairs of variants in a window size of 50 variants. The window was shifted after every 5 variants. Any variant pairs that had squared correlation above a specified threshold ($r^2 \leq 0.2$) in a window were listed in `QCed_popstrat.atcg.inbred.prune.out`, and the variants were pruned continuously in the window until no such pairs remained. The listed variants were excluded in the next step to calculate the inbreeding coefficients (F), as shown by the command line below:

```
./plink --noweb --bfile QCed_popstrat.atcg --het --exclude QCed_popstrat.atcg.inbred.prune.out --out QCed_popstrat.atcg.inbred
```

This step skipped haploid markers (male X and Y chromosome markers) in calculating the F per individual. The result file, `QCed_popstrat.atcg.inbred.het` was created and contained the Fs for all individuals within the same study cohort. A scatter plot of F was constructed for each study cohort, and any individuals who were located outside of a specified threshold (4SD

away from the mean of F) on the graph matrix, were included in a list of 'failed population structure'.

### 3.7.3. Adjustment 2: Pair-wise Identity-by-descent (IBD) Estimation

Identity-by-descent (IBD) can be estimated between a pair of individuals to resolve the issue of sample contamination, plating errors during genotyping (sample swap, duplications etc.). In addition, pedigree errors and unknown familial relationships can be detected by using pair-wise IBD estimation. I conducted a pair-wise IBD estimation analysis for each study cohort to detect specific segments of DNA that are shared between distantly-related individuals within the same study cohort. Basically, this analysis enables detection of pairs of individuals who are more closely genetically related to each other than we would expect by chance in a random homogeneous population.

To reduce computational complexity, a subset of variants was pruned in using similar arguments as in **Adjustment 1: Inbreeding**, together with additional arguments as below:

```
./plink --noweb --bfile QCed_popstrat.atcg --indep-pairwise 50 5 0.2 --maf 0.01 --exclude high-LD-regions.txt --range ----out QCed_popstrat.atcg.ibd
```

The --maf argument was executed to select common variants with minor allele frequency (MAF) > 1%. A list of known genomic regions that have high linkage disequilibrium (LD) was previously downloaded and extracted from this link: http://www.well.ox.ac.uk/ggeu/NPqc/raw-GWA-data.tgz. The resulting output file (`QCed_popstrat.atcg.ibd.prune.in`) contained the list of variants that are common (MAF>1%) and independent to each other (r2≤0.2).

IBD was estimated by executing the command line as below:

```
./plink --noweb --bfile QCed_popstrat.atcg --no-parents --extract
QCed_popstrat.atcg.ibd.prune.in --Z-genome --out QCed_popstrat.atcg.ibd
```

The Z-genome argument was implemented so that PLINK could compress the massive results that were generated. In the results file (`QCed_popstrat.atcg.ibd.genome.gz`), the proportion of the genome shared between pairs were labelled as 'PI_HAT'.

I considered a pair to be highly related to one another if the pair had PI_HAT > 0.2 (i.e. at least 20% of the genome shared between the pair). Therefore, I only extracted the results for the pairs according to the above-mentioned threshold, using the command as below:

```
gunzip QCed_popstrat.atcg.ibd.genome.gz
```

and

```
awk '$10>0.2' QCed_popstrat.atcg.ibd.genome > QCed_popstrat.atcg.ibd.genome.read
```

In the results file (`QCed_popstrat.atcg.ibd.genome.read`), individuals who were listed for exclusion were selected according to the criteria as below (priority-wise), and added into the 'failed population structure' list:

i. the individual who had a lower call rate than its pair

ii. if the call rates were similarly high, individuals were selected based on sex so that the sex (of the excluded individuals) is balanced between males and females.

### 3.7.4. Adjustment 3: Multidimensional Scaling (MDS)

Multidimensional scaling (MDS) can be used to visualise the genetic distance and assess population structure in GWAS. This method is commonly used as it identifies useful components underlying the different dimensions governing where each individual point is located. The dimensions are the observed genetic distances between individuals based on the **Adjustment 2: Pair-wise Identity-by-descent (IBD) Estimation** section (Wang *et al.*, 2009).

To detect population outliers using MDS, I used HapMap Phase 3 (release 2) genotype data to anchor the dimensional scales using common variants that are shared between a study cohort and HapMap (MAF > 0.01, r2 < 0.2). Thus, HapMap PLINK-binary files were downloaded from its FTP directory, as executed below:

```
wget
ftp://ftp.ncbi.nlm.nih.gov/hapmap/genotypes/hapmap3/plink_format/draft_2/hapmap3_r
2_b36_fwd.qc.poly.tar.bz2
```

Separate PLINK format files were downloaded and extracted for eleven HapMap populations- African ancestry in Southwest USA (ASW), Utah residents with Northern and Western European ancestry from the CEPH collection (CEU), Han Chinese in Beijing, China (CHB), Chinese in Metropolitan Denver, Colorado (CHD), Gujarati Indians in Houston, Texas (GIH), Japanese in Tokyo, Japan (JPT), Luhya in Webuye, Kenya (LWK), Mexican ancestry in Los Angeles, California (MEX), Maasai in Kinyawa, Kenya (MKK), Toscani in Italy (TSI), and Yoruba in Ibadan, Nigeria (YRI)(International HapMap, 2003). All HapMap populations were merged based on a text file `hapmap3r2.allpop.txt` containing the filenames (2 columns, each column represented either `.map` or `.ped` file format).

```
./plink --noweb --file hapmap hapmap3_r2_b36_fwd.ASW.qc.poly --merge-list
hapmap3r2.allpop.txt --make-bed --out hapmap3_r2_b37_fwd.allpop
```

Our study cohorts were annotated based on GRCh37 (hg19) human genome build assembly. Prior to its use, the annotation for the HapMap dataset was updated to hg19 using the LiftOver facility in the UCSC Genome Browser (Hinrichs *et al.*, 2006).

A new HapMap dataset that included the variants featured in our study cohort was created by executing the command below:

```
/plink --noweb --bfile hapmap3_r2_b37_fwd.allpop --extract QCed_popstrat.atcg.bim
--make-bed --out hapmap3r2_b37_sharedSNPs
```

Both datasets for HapMap and the study cohort were merged in PLINK on the basis of shared variants, as shown below:

```
./plink --noweb --bfile hapmap3r2_b37_sharedSNPs --bmerge QCed_popstrat.atcg.bed
QCed_popstrat.atcg.bim QCed_popstrat.atcg.fam --extract
hapmap3r2_b37_sharedSNPs.bim --make-bed --out Merged_hapmap3r2_studycohort
```

If the process stopped due to mismatching SNPs, the strands between the two datasets were aligned by flipping the strand of one of the datasets, and they were merged again, as executed below:

```
/plink --noweb --bfile hapmap3_r2_b37_fwd.allpop --extract QCed_popstrat.atcg.bim
–-flip Merged_hapmap3r2_studycohort.missnp --make-bed --out
hapmap3r2_b37_sharedSNPs.flipped
```

And

```
./plink --noweb --bfile hapmap3r2_b37_sharedSNPs.flipped --bmerge
QCed_popstrat.atcg.bed  QCed_popstrat.atcg.bim QCed_popstrat.atcg.fam --extract
hapmap3r2_b37_sharedSNPs.bim --make-bed --out Merged_hapmap3r2_studycohort
```

If there were still a small number of variants listed as 'miss SNP' after flipping and merging three times, I excluded those variants altogether from both datasets, and re-merged them.

A subset of variants was pruned in using similar arguments as in **Adjustment 2: Pair-wise Identity-by-descent (IBD) Estimation**, as below:

```
./plink --noweb --bfile Merged_hapmap3r2_studycohort --indep-pairwise 50 5 0.2 --
maf 0.01 --exclude high-LD-regions.txt --range ----out
Merged_hapmap3r2_studycohort.ibd
```

One of the output files, `Merged_hapmap3r2_studycohort.ibd.prune.in` contained only common, unrelated variants (MAF > 0.1, r2 ≤ 0.2) that were in low LD with each other.

Using the merged dataset, I estimated the IBD for individual pairs by executing the command below:

```
./plink --noweb --bfile Merged_hapmap3r2_studycohort --no-parents --extract
Merged_hapmap3r2_studycohort.ibd.prune.in --Z-genome --out
Merged_hapmap3r2_studycohort.ibd
```

As described in **Adjustment 2: Pair-wise Identity-by-descent (IBD) Estimation**, the proportion of the genome shared between each individual pair was given as PI_HAT in the resulting file.

The MDS was performed on the first 10 dimensions.

The calculation was based on the common, unrelated variants and the PI_HAT of each individual pair, as shown below:

```
./plink --noweb --bfile Merged_hapmap3r2_studycohort --read-genome
Merged_hapmap3r2_studycohort.ibd.genome.gz --extract
Merged_hapmap3r2_studycohort.ibd.prune.in --cluster --mds-plot 10 --out
Merged_hapmap3r2_studycohort
```

The output `Merged_hapmap3r2_studycohort.mds` contained the MDS values for the first 10 components per individual in the merged dataset. Using these values, a scatter plot of MDS values of dimension 1 (D1) against MDS values of dimension 2 (D2) was constructed. The plot displayed the relative genetic distance between individuals in the study cohort, anchored by the eleven HapMap Phase 3 populations. Any individuals who can be visibly seen to be located outside of the study cohort cluster on the plot, were considered to be population outliers. Those individuals were added into the list of 'failed population structure'.

### 3.7.5. Sample exclusions

A list of individuals who were identified as having high inbreeding estimates, high cryptic relatedness, or being population outliers was compiled as 'failed population structure'. Those individuals were excluded from the study cohort using PLINK, as shown below:

```
./plink --noweb --bfile QCed_popstrat.atcg --remove failedpopulationstructure.txt
--make-bed --out Cleanedcohort
```

The output files were PLINK-binary files of the study cohort that had undergone QC and population stratification analysis.

### 3.7.6. Adjustments in the UK Biobank

Population structure adjustments for the UKBB cohort were conducted by my colleague, Andrianos Yiorkas (Department of Medicine, Imperial College London, UK) to replicate the QC results of Bycroft, C., *et. al.* (2017) (Clare Bycroft *et al.*, 2017). In **Chapter 3.6.6 Quality controls in the UK Biobank**, ethnic outliers had already been flagged and removed prior to this section. Hence, the UKBB cohort was already homogenously 'white British' at this point.

The same subset of autosomal high quality SNPs (SNP call rate>98%, HWE>0.01%; refer **Chapter 3.6.6 Quality controls in the UK Biobank**) was used to identify genetically related individuals in the UKBB cohort. Briefly, familial relationship inference was conducted by estimating kinship coefficients for all pairs of individuals using the KING tool (Manichaikul *et al.*, 2010). Kinship coefficient is defined as the probability of two randomly sampled individuals having the same two alleles, which are identical by descent (IBD) (Clare Bycroft *et al.*, 2017, Manichaikul *et al.*, 2010). Pairs were considered genetically related if they had kinship coefficients of 1/16 ($3^{rd}$ degree relatives) or higher ($1^{st}$ and $2^{nd}$ degree relatives). Additionally, one individual was randomly-selected from each IBD pair, and they were tested for IBD between the resulting pairs using the same procedures as the first round. This step was conducted at least 5 more rounds until no related pair remained (Clare Bycroft *et al.*, 2017). A complete method is described in detail by Bycroft, C., *et. al.* (2017) (Clare Bycroft *et al.*, 2017).

## 3.8. Principal component analysis (PCA)

Principal components are commonly used to correct for confounding effects in downstream association analyses. Each principal component represents an uncorrelated variation value, derived from seemingly correlated variables that are present strongly within a dataset. In order to reduce the degree of correlated variables, principal component analysis (PCA) was performed for each study cohort after it had been cleaned through QC procedures and adjusted for population structure. I carried out PCA for the NFBC1966, NFBC1986, NutriTech and the PMMO cohorts, by executing the --pca command in the PLINK Version 1.07 package (Purcell *et al.*, 2007). The PCA for the UKBB cohort was conducted by Andrianos Yiorkas (Department of Medicine, Imperial College London, UK) as part of his PhD project. Firstly, a subset of common, unrelated SNPs (from genomic regions that were not known to harbour high-LD SNPs) were selected by executing the command below:

```
./plink --noweb --bfile Cleanedcohort --maf 0.01 --exclude high-LD-regions.txt --
range --indep-pairwise 50 5 0.2 --out Cleanedcohort_pca
```

The resulting output file (`Cleanedcohort_pca.prune.in`) contained the list of SNPs that were common (MAF>1%) and independent to each other ($r^2 \leq 0.2$).

IBD was estimated for all pairs of individuals within the same study cohort, based on the pruned SNPs in `Cleanedcohort_pca.prune.in`, by executing the command below:

```
./plink --noweb --bfile Cleanedcohort --no-parents --extract
Cleanedcohort_pca.prune.in --Z-genome --out Cleanedcohort_pca
```

PI_HAT, which is the proportion of the genome shared between a pair of individuals, was estimated in the resulting compressed `Cleanedcohort_pca.genome.gz` file. This file was subsequently used to calculate the first 20 principal components based on the variance-standardised relationship matrix provided in PLINK, by executing the command as below:

```
./plink --noweb --bfile Cleanedcohort --no-parents --extract
Cleanedcohort_pca.prune.in --read-genome Cleanedcohort_pca.genome.gz --cluster --
pca 20 --out Cleanedcohort_pca
```

The resulting output files included `Cleanedcohort_pca.eigenval` and `Cleanedcohort_pca.eigenvec`. The former contained standardised-variance values (eigenvalues) that were clustered for each principal component (PC), while the latter contained all uncorrelated standardised-variance values (eigenvectors) for the first 20 principal components in all individuals. To reduce the number of principal components needed as covariates, all eigenvalues were plotted in ascending order of principal component. Standard error (SE) bars were calculated and shown on all eigenvalue points. The plot was visually inspected from the first principal component (PC1) until the point where there was a clear redundancy of the error bars from the flanking PCs was identified. That point was flagged as the final PCs to be included as covariates in downstream multivariate analyses for each cohort. This procedure is somewhat arbitrary, since it depends on visual observation of the plot. However, it is obvious to see the variations being reduced from the PC1 onwards until the point where the variations became negligible.

## 3.9. CNV calling

As described earlier in **Chapter 3.5.2 Processing of genotyped data**, a table called 'For-CNV Table' was generated by the GenomeStudio Version 2011.1 (Illumina Inc., San Diego, California, USA). The table contained the chromosomes and genomic positions, genotypes, Log R Ratio (LRR) values, and B Allele Freq (BAF) values of all the genotyped SNPs for all individuals in each study cohort prior to any QC. The LRR and BAF values represent the overall signal intensity and allelic intensity ratios for all SNP markers that were genotyped, respectively. These signal information for each SNP was used for the CNV calling process in this chapter.

Overall, I performed the CNV calling process for NFBC1966, NFBC1986, NutriTech and PMMO cohorts in Linux/Unix environment using the PennCNV (May 2014 version) tool according to the author's recommended protocols (Wang *et al.*, 2007). Any modifications are stated accordingly in the text. By the end of this process, each study cohort is expected to have a raw CNV catalogue that is ready to be assessed for its overall quality and confidence.

For the UKBB cohort, I received the most recent version of the LRR and BAF tables from the UKBB server, separately and very late into my final year. Substantial effort was made to modify the massive files and split them into several smaller parts so that they could be used by the PennCNV tool. However, technical issues such as insufficient storage memory, failure of data manipulation, and my own limited capabilities to handle such complicated conversions stunted the progress, so that no CNV calls were predicted in the UKBB cohort using PennCNV. However, the signal information data were used elsewhere in this thesis.

### 3.9.1. Split signal information data

The signal information table ("For-CNV Table") containing the LRR and BAF values for all the SNPs of each individual was split into separate, smaller tables each corresponding to an individual within the study cohort. This process was performed using the kcolumn.pl script in the PennCNV package. The executed command line is shown below:

```
./kcolumn.pl For-CNV Table split 3 –heading 3 -tab -out cohort
```

The output files generated were named `cohort.split`. Firstly, the kcolumn.pl script took the first 3 columns (SNP name, chromosome, genomic position) of all SNPs and printed them into output files as a 'header' for each individual file. Secondly, the script skipped the first 3 columns (that have been printed), and continued to print subsequent 3 columns (representing the genotypes, LRR, and BAF values for an individual) into the first split file `cohort.split1` (already containing the first 3 columns as mentioned above), as the signal information table for the first individual. Consequently, the script repeated the second step by skipping the first 6 columns and printed the subsequent 3 columns into the second split file `cohort.split2`. The script continued to generate output files until all individuals had their signal information printed into individual split files. Therefore, the order of split numbers matched the order of the individuals in the "For-CNV Table" column-wise. A list file that compiled all of the split file names was created by simply executing the following command:

```
find cohort.split* > cohort.splitlist.txt
```

and

```
wc cohort.splitlist.txt
```

### 3.9.2. Create cohort-specific population frequency of B allele

A data file containing the population frequency of the B allele (.pfb) specific to each study cohort, was generated from the collection of individual signal information files from the previous step. All .pfb files were stored in the penncnv/lib directory for practical purposes. The following commands were executed to create a.pfb file:

```
./compile_pfb.pl –listfile cohort.splitlist.txt –output cohort.pfb
```

and

```
mv cohort.pfb /..../penncnv/lib
```

### 3.9.3. Calling of CNVs

Prior to CNV calling by the PennCNV tool, an HMM file was initially located. The HMM file contained the hidden Markov model (HMM) needed for the script, detect_cnv.pl, to assess the copy number states and call for the CNVs in an individual-by-individual approach. The HMM file provided two types of information; 1) the expected signal intensity values for different copy number state, and 2) the probability of an expected transition of one copy number state to another copy number state (Wang *et al.*, 2007). As recommended by the author, the `hhall.hmm` file was used as the HMM file for all study cohorts, since all genotyping had been done on Illumina platforms (Wang *et al.*, 2007). The CNV calling process was performed by executing the following command:

```
./detect_cnv.pl -test -hmm ./lib/hhall.hmm -pfb ./lib/cohort.pfb –list
cohort.splitlist.txt –conf -log cohort.log -out cohort.rawcnv
```

At this stage, I did not put any filtering arguments, such as minimum confidence score (-minconf), into the above-mentioned command so that I could obtain a complete, raw CNV

catalogue for each study cohort. The resulting output files `cohort.rawcnv` and `cohort.log` contained raw CNV calls predicted by PennCNV for all individuals, and the signal quality information for each individual, respectively.

## 3.10. CNV quality controls

The CNVs that were called by the PennCNV tool were compiled as a raw CNV catalogue for each study cohort. Each CNV catalogue was subjected to several filtering processes in order to retain only high-quality, high-confidence CNV calls with dense coverage of SNPs. Overall, I performed the quality controls of raw CNV calls for NFBC1966, NFBC1986, NutriTech, and PMMO cohorts due to the availability of the data. By the end of this process, each study cohort was expected to have a clean, high-quality, high-confidence CNV catalogue that is ready to be analysed.

### 3.10.1. Filter 1: QC from SNP genotyping

For each study cohort, all individuals who failed the QC and population structure adjustments were removed from its raw CNV catalogue by manual filtering.

### 3.10.2. Filter 2: QC from CNV calling

For each study cohort, all individuals who passed the QC and population structure adjustments were subjected to a series of PennCNV-internal QC based on a number of signal quality parameters per individual; standard deviations of Log R Ratio (LRR_SD), standard deviations of B Allele Frequency (BAF_SD), the drifting of the observed B Allele Frequency

(BAF) from the expected BAF values (BAF_DRIFT), the amount of genomic wave dispersed in signal intensity (WF), the summary score of WF that is explained by the local GC content (GCWF) (Diskin *et al.*, 2008), and the number of CNVs (#CNV). The distribution of each signal quality parameter per cohort was plotted. Cut-off thresholds were determined based on visual inspections on each parameter, cohort-by-cohort.

**(a) Quality summary extraction**

The `detect_cnv.pl` script typically generates a `.log` file for every execution by PennCNV. The .log file, which contains most of the parameters mentioned above, represents the quality summary for each individual within a study cohort. The quality summary entries were extracted from the .log file using the command as below:

```
grep “quality summary” cohort.log > cohort.qualsum
```

The `.rawcnv` file generated from the `detect_cnv.pl` script contained raw CNV calls per study cohort. Both `.rawcnv` and `.qualsum` files were exported into Excel for easier handling of the parameters.

**(b) Determining parameter thresholds**

For each study cohort, LRR_SDs, BAF_SD, #CNV and BAF_DRIFT for all individuals who survived Filter 1 were sorted in ascending order and plotted as scatter plots in Excel. For each parameter, the value whereby the plot increased exponentially (typically at a point when the points were starting to break) was noted. I initially used a standard cut-off point of mean + 4 standard deviations (SD), 6SD, 2SD and 4SD, as the maximum inclusion thresholds for the parameters, respectively. The WF and GCWF parameters were plotted against #CNV for all

individuals who survived Filter 1. The standard cut-off points of mean ± 4SD was initially used as the minimum and maximum inclusion thresholds for the WF and GCWF parameters.

The inclusion thresholds are different between study cohorts due to the differences of genotyping chips used, overall DNA qualities, different human handling, different laboratory conditions, and other unknown variables. Unless stated otherwise, the inclusion thresholds for the above-mentioned parameters were consistent across study cohorts.

The removal of individuals based on the visual inspection of each parameter distribution plot was performed in PennCNV tool by executing running the `filter_cnv.pl` script according to PennCNV protocols(Wang *et al.*, 2007).

### 3.10.3. Filter 3: low-quality and low-confidence CNVs

In addition to the signal quality parameters provided by PennCNV for individual exclusion, any CNV calls that had internal confidence scores (conf) < 20, and/or had a low SNP density (<3 SNPs per 100 kb) were excluded from the CNV catalogue. A confidence score is an empirical value given by PennCNV software to reflect the degree of confidence that the detected copy number change of that region does exist. It also reflects the reliability of a CNV call to be included into a common CNV region. This step was conducted to reduce false positive CNV calls since CNVs with low density of SNPs have the tendency to overlap low copy repeats (LCRs) and other genomic regions that are difficult to genotype (Wang *et al.*, 2007). The resulting file after applying Filter 1 to 3 was called `cohort_cnvqced.cnv.`

### 3.10.4. Filter 4: spurious CNV regions

CNV calls with 50% of the lengths overlapping several genomic regions that have been previously known to harbour cell-line artifacts were excluded to minimise the number of false

positive calls. The regions (immunoglobin, centromeric and telomeric regions) were compiled in `immuno_centro_telo_regions_hg19.txt` file from UCSC Table Browser Data Retrieval Tool (hg19) (Karolchik *et al.*, 2004).

The removal of CNV calls in spurious regions was performed in PennCNV tool by running the `scan_region.pl` script according to PennCNV protocols (Wang *et al.*, 2007), as shown below:

```
./scan_region.pl cohort_cnvqced.cnv immuno_centro_telo_regions_hg19.txt -minqueryfrac 0.5 > cohort_cnvqced.imm; fgrep -v -f cohort_cnvqced.imm cohort_cnvqced.cnv > cohort_cnvqced.nonspurious.cnv
```

### 3.10.5. Filter 5: merge CNVs

As recommended by the PennCNV developer(Wang *et al.*, 2007), any two CNV calls that had a gap of ≤ 20% (of the total length between them) were merged together as a single CNV call due to the tendency of PennCNV algorithm to split large CNV calls into several segments. The CNV merging process was performed in PennCNV tool by running the `clean_cnv.pl` script according to PennCNV protocols (Wang *et al.*, 2007), as shown below:

```
./clean_cnv.pl combineseg --fraction 0.2 --signalfile ./lib/cohort.pfb cohort_cnvqced.clean.cnv > cohort_cnvqced.clean.gapmerged.cnv
```

## 3.11. Other methods

There are other methods that were implemented specific for several Chapters. Those methods are described in their own chapters for convenience, under the sub-title ‘Specific methods’.

## 3.12. Methods overview in each Result Chapter

| Section | Method name | Chapter 4 | Chapter 5 | Chapter 6 | Chapter 7 | Chapter 8 |
|---|---|---|---|---|---|---|
| **3.2** | Biological sample collection | ✓ | ✓ | ✓ | ✓ | ✓ |
| **3.3** | Genomic DNA extraction | ✓ | ✓ | ✓ | ✓ | ✓ |
| **3.4** | Obesity-related phenotypic measurements | ✓ | ✓ | ✓ | ✓ | ✓ |
| **3.5** | SNP genotyping | ✓ | ✓ | ✓ | ✓ | ✓ |
| **3.6** | Quality controls | ✓ | ✓ | ✓ | ✓ | ✓ |
| **3.7** | Population structure adjustments | ✓ | ✓ | ✓ | ✓ | ✓ |
| **3.8** | Principal component analysis | | | | | ✓ |
| **3.9** | CNV calling | | ✓ | | ✓ | ✓ |
| **3.10** | CNV quality controls | | ✓ | | ✓ | ✓ |

# CHAPTER 4. Genetic Data Processing

## 4.1. Introduction

This chapter describes the results of quality control procedures and population structure adjustments on the dataset from each study cohort.

The rapid progress of high-throughput technology has transformed the landscape of genomics research into new heights. The use of microarrays has provided us with a fast and economical method to genotype polymorphisms across the whole genome for thousands of individuals reliably. The SNP genotyping arrays that are commercially available include arrays that have genome-wide coverage, with the potential for customised/targeted genes, and ethnic-specific SNP content, as well as arrays designed for multi-ethnic populations.

Since its introduction, SNP genotyping has been the preferred tool in most genome-wide association studies (GWAS) for the past 10 to 15 years. The majority of the SNPs in many commercial SNP genotyping arrays are considered common in the general population, with minor allele frequencies (MAF) of more than 5%. Having said that, the most recent genotyping array technology introduces us to the use of exome arrays; a genotyping array containing a reduced number of single nucleotide variants, focusing on exonic variants across the whole genome. Depending on the specific arrays, a significant portion of these exonic variants (~65-70%) are considered rare, with MAF of less than 1% (Illumina, 2014). The differences between genome-wide SNP arrays and exome arrays in terms of the variant frequencies, require us to analyse the data in different ways (Guo *et al.*, 2014). However, the general approach to processing these datasets are similar: usually divided into two major phases based on the tools used; 1) processing of raw genomic data using manufacturer-provided software, and 2) quality controls of the raw genomic data using a specialised tool for genetic data analysis.

Most of the study cohorts involved in this thesis were genotyped on Illumina's genotyping platforms. Thus, I have used GenomeStudio Version 2011.1 software (Illumina Inc., San Diego, California, USA) for the first phase, and PLINK Version 1.07 (Purcell *et al.*, 2007) tool for the second phase.

The UK Biobank cohort (UKBB) was genotyped on Affymetrix's genotyping arrays (Affymetrix Inc., Santa Clara, California, USA). For the UKBB cohort, the first phase was conducted by Bycroft, C., *et. al.* (2017) (Clare Bycroft *et al.*, 2017), while the second phase was carried out by my colleague, Andrianos Yiorkas (Department of Medicine, Imperial College London, UK). Therefore, the genotyped data from the UKBB that I received were already processed and quality-controlled.

**Figure 3-1** (refer to **Chapter 3**) shows a simplistic workflow of the procedures involved in processing the genotyped data from the Illumina-genotyped cohorts. Briefly, the genotyping produced raw image files per individual. These raw image data were submitted into GenomeStudio software to obtain the SNP calls based on a specific Illumina clustering algorithm, GenCall. The calls were then transformed into a large data table in PLINK format to complete the first phase. The second phase began by feeding the data table into the PLINK tool after it had been transformed into a format readable by PLINK: in a binary format. The binary files were then used for quality controls: check for genotyping call quality; sex discordance; missing SNP data; and high autosomal heterozygosity rates, as well as correction for population structure.

## 4.2. Aim of this chapter

The main aim of the work described in this chapter is to take raw unreadable data generated from a biological source (typically DNA samples), and from it to generate a high-quality, clean genetic dataset that can be systematically analysed. The aim is the same across all study cohorts. The generation of clean genetic data for each cohort was addressed by the following questions:

1. *Do any of the study cohorts have DNA samples with low genotyping call rates?*
2. *Do any of the study cohorts have individuals with conflicting information between the self-reported sex and genetically-determined sex?*
3. *Do any of the study cohorts have individuals with high proportion of missing/not genotyped SNPs?*
4. *Do any of the study cohorts have individuals with elevated autosomal heterozygosity rates?*

I answered these questions by running appropriate commands provided in PLINK software; I looked through every individual's call rate, executed SNP missingness test, scanned the genetic information on chromosome X to determine the actual sex, and executed heterozygosity tests on autosomal chromosomes. I also created a list of individuals who did not pass any of the quality checks.

5. *Do any of the study cohorts have individuals who have high inbreeding rates?*
6. *Are any of the individuals genetically related to the others from the same study cohort?*
7. *Do any of the individuals belong to another ethnicity or population that is genetically different from the rest of their study cohort?*

I answered these questions by adjusting the structure of each study cohort through inbreeding, identity-by-descent estimation and multidimensional scaling components in PLINK software. I created a list of all individuals who failed any of the adjustments. Inbreeding rates were ascertained by calculating the inbreeding coefficients per individual, based on the SNPs on autosomal chromosomes.

The genetic data that have been cleaned and quality-controlled were subsequently used in downstream CNV analyses. Having individuals with high quality genetic information is crucial in CNV calling, since copy number analyses heavily depend on the overall quality of the data. In addition, any association tests could easily be compromised due to random errors and erroneous CNV calls, potentially resulting false-positive (Type I error) or false-negative (Type II error) associations.

## 4.3. General methods used in this chapter

| Section | Method name (Described in Chapter 3) | Chapter 4 | Chapter 5 | Chapter 6 | Chapter 7 | Chapter 8 |
|---|---|---|---|---|---|---|
| **3.2** | Biological sample collection | ✓ | ✓ | ✓ | ✓ | ✓ |
| **3.3** | Genomic DNA extraction | ✓ | ✓ | ✓ | ✓ | ✓ |
| **3.4** | Obesity-related phenotypic measurements | ✓ | ✓ | ✓ | ✓ | ✓ |
| **3.5** | SNP genotyping | ✓ | ✓ | ✓ | ✓ | ✓ |
| **3.6** | Quality controls | ✓ | ✓ | ✓ | ✓ | ✓ |
| **3.7** | Population structure adjustments | ✓ | ✓ | ✓ | ✓ | ✓ |
| **3.8** | Principal component analysis | | | | | ✓ |
| **3.9** | CNV calling | | ✓ | | ✓ | ✓ |
| **3.10** | CNV quality controls | | ✓ | | ✓ | ✓ |

# 4.4. Results: Sample Quality Controls

## 4.4.1. NFBC1966 cohort

The NFBC1966 participants were genotyped using two different genotyping arrays (as mentioned in **3.5.1 Arrays used in different cohorts**); a GWAS array chip was applied to 5,557 samples, and an Exome array chip was applied to 1,536 individuals. Ninety-four percent of individuals (n=1,443) genotyped using the Exome chip had already been genotyped using the GWAS chip. Therefore, for the purpose of identifying poor-performing DNA samples and population outliers, I only focused on the GWAS chip since it has more SNPs and a bigger sample. Moreover, the genotype assignment described in **Chapter 3.5 SNP genotyping** could only be conducted for the GWAS chip.

**a) Quality control 1: Sex discordance**

In **Chapter 3.6.1 Quality control 1: Sex discordance**, I have described that an individual was classified as a male if all of his SNPs on chromosome X were homozygous. This was because males supposedly have only one copy of chromosome X. Hence, the average homozygosity rate for those SNPs (Fx) was expected to be 1 for a male (with ± 0.2 accounted for any genotyping errors). On the other hand, females were expected to have the average homozygosity rates of 0 (with ± 0.2 accounted for any genotyping errors) due to the presence of another chromosome X in normal females. Therefore, individuals were classified as genetically males if their Fx were 1 ± 0.2, and as genetically females if their Fx were 0 ± 0.2. Any individuals with Fx between 0 and 1 were classified as having ambiguous sex and were excluded from further analyses.

A total of 10,145 X-chromosome SNPs were present on the GWAS array chip used to genotype NFBC1966 individuals (n=5,557). These SNPs were used to determine sex based on genetics. **Figure 4 -1** shows the Fx for all individuals in the NFBC1966 cohort (n=5,557). A total of 11 individuals (**Figure 4 -1**; red circles) had an Fx between 0.2 and 0.8, indicating that their sex could not be unambiguously determined from SNP genotyping data. Twenty-three additional individuals (n=23) had self-reported sex different from the sex determined by their genotyped data. In summary, a total of 34 individuals from the NFBC1966 cohort (0.61%) were flagged to have sex discordance.

**b) Quality control 2: Missing genotype data, high autosomal heterozygosity rates**

The proportion of genotyped SNPs that were not called (will be referred to as 'SNP missingness' henceforth) was inspected, together with the distribution of mean heterozygosity rate (Het) for the SNPs on autosomal chromosomes across all individuals. This step was conducted to provide an indication of DNA sample quality, sample contamination, inbreeding, and overall genotyping quality. The calculation for Het and the plotting of SNP missingness against Het are described in detail in **Chapter 3.6 Quality controls**. I used 5% SNP missingness, and ± 4 standard deviations (SD) away from the mean Het, as the standard inclusion thresholds across all study cohorts (unless otherwise stated). These thresholds were based on the recommendations and protocols from previous GWAS publications (Anderson *et al.*, 2010, Silverberg *et al.*, 2009, Wellcome Trust Case Control, 2007).

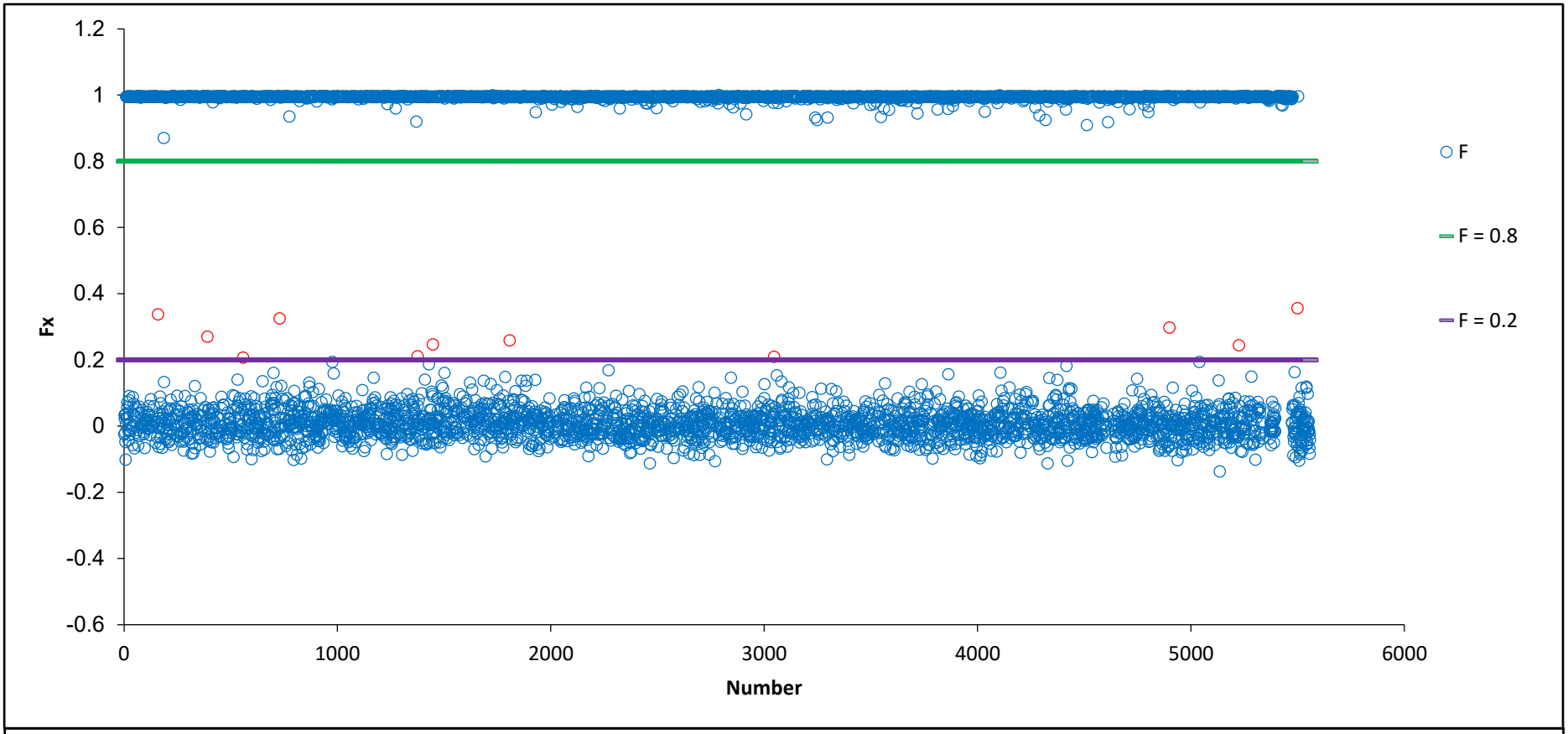


**Figure 4 -1:** Mean homozygosity rates of X-chromosome SNPs (Fx) for NFBC1966 individuals (n=5,557). Each blue circle represents an individual. The green line represents the lower Fx threshold (Fx=0.8) to infer an individual as a male. The purple line represents the upper threshold (Fx=0.2) to infer an individual as a female. Red circles represent individuals whose sex could not be genetically determined based on genotyping.

**Figure 4-2** shows the distribution of mean heterozygosity across all individuals on the x-axis and the proportion of SNP missingness for each individual in NFBC1966 cohort. Although all of them had high SNP call rates (>95% call rates), a total of 44 individuals (0.79%) were identified to have either excessive or reduced proportion of heterozygous genotypes (**Figure 4-2;** red circles).

**c) Quality control 3: Sample call rates**

The inclusion criterion for genotyping call rate in the NFBC1966 cohort was ≥97%, which was lower than used for the other study cohorts (i.e. ≥98%). To justify the relatively lower threshold, I looked at the number of samples that failed the 98% threshold in the NFBC1966 cohort: there were twice as many samples that failed the call rate threshold at 98% (216 individuals had call rates between 97% and 98%). I suspected that the reason for this being the old array's outdated SNP content applied on the NFBC1966 cohort had an effect in generating the relatively low genotyping quality. A personal communication with Illumina's technical support team had confirmed that Illumina Infinium® HumanCNV370-Duo v1.0 microarrays (NFBC1966) contained fewer SNPs with older information than in Illumina Infinium® Human OmniExpress arrays used in other cohorts. In summary, 106 individuals were identified to have genotyping call rate <97%.

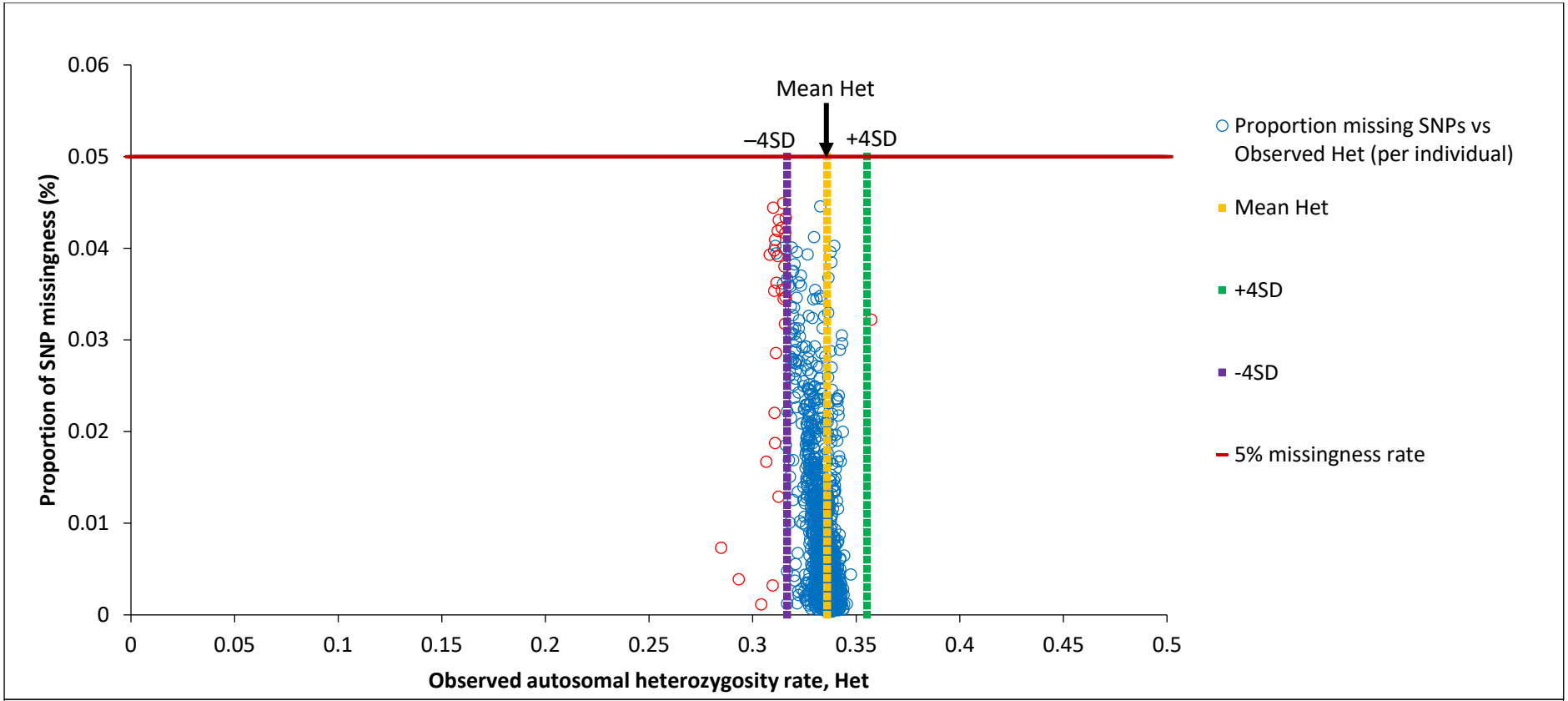


**Figure 4-2:** A scatter plot of the proportion of SNP missingness (%) against autosomal heterozygosity rate (Het) observed per individual in NFBC1966 cohort (n=5,557). Each blue circle represents an individual. The red circles represent the individuals who either had >5% SNP missingness, excessive (>mean+4SD) or reduced (<mean-4SD) proportion of heterozygote genotypes, or both. The red line represents 5% threshold level for SNP missingness.

**d) Sample exclusions**

A total of 194 samples failed at least one QC step. An additional sample was later identified as a duplicate of a failed sample. Upon compilation of these individuals, 36 had failed more than one QC steps, which resulted in a total of 158 NFBC1966 individuals (2.84%; male=70, female=86, missing sex=2) being excluded prior to the population structure adjustments.

### 4.4.2. NFBC1986 cohort

The NFBC1986 participants were genotyped in two batches on similar genotyping array chips containing GWAS SNPs and exonic variants (as mentioned in **Chapter 3.5.1 Arrays used in different cohorts**); the first batch (NFBC1986A) consisted of 2,016 individuals, while the second batch (NFBC1986B) consisted of 2,027 individuals (total n=4,043). The genotyping for both batches were carried out according to **Chapter 3.5 SNP genotyping**.

**a) Quality control 1: Sex discordance**

A total of 22,394 and 22,927 X-chromosome SNPs were present on the NFBC1986A and NFBC1986B array chips, respectively. These SNPs were used to ascertain the sex genetically. **Figure 4- 3** shows the mean homozygosity rates of the SNPs on chromosome X, for all individuals in NFBC1986 cohort. By using the same Fx thresholds described in **Chapter 4.4.1 NFBC1966,** a total of 7 individuals (**Figure 4- 3;** red circles) were identified to have undermined sex based on their genotyped data. In addition, there were 22 individuals with different sex ascertained between their clinical and genotyped data, while 6 individuals had no sex being reported in their clinical data. In summary, a total of 35 individuals (0.86%; 18 from NFBC1986A and 17 from NFBC1986B) were flagged to have sex discordance.

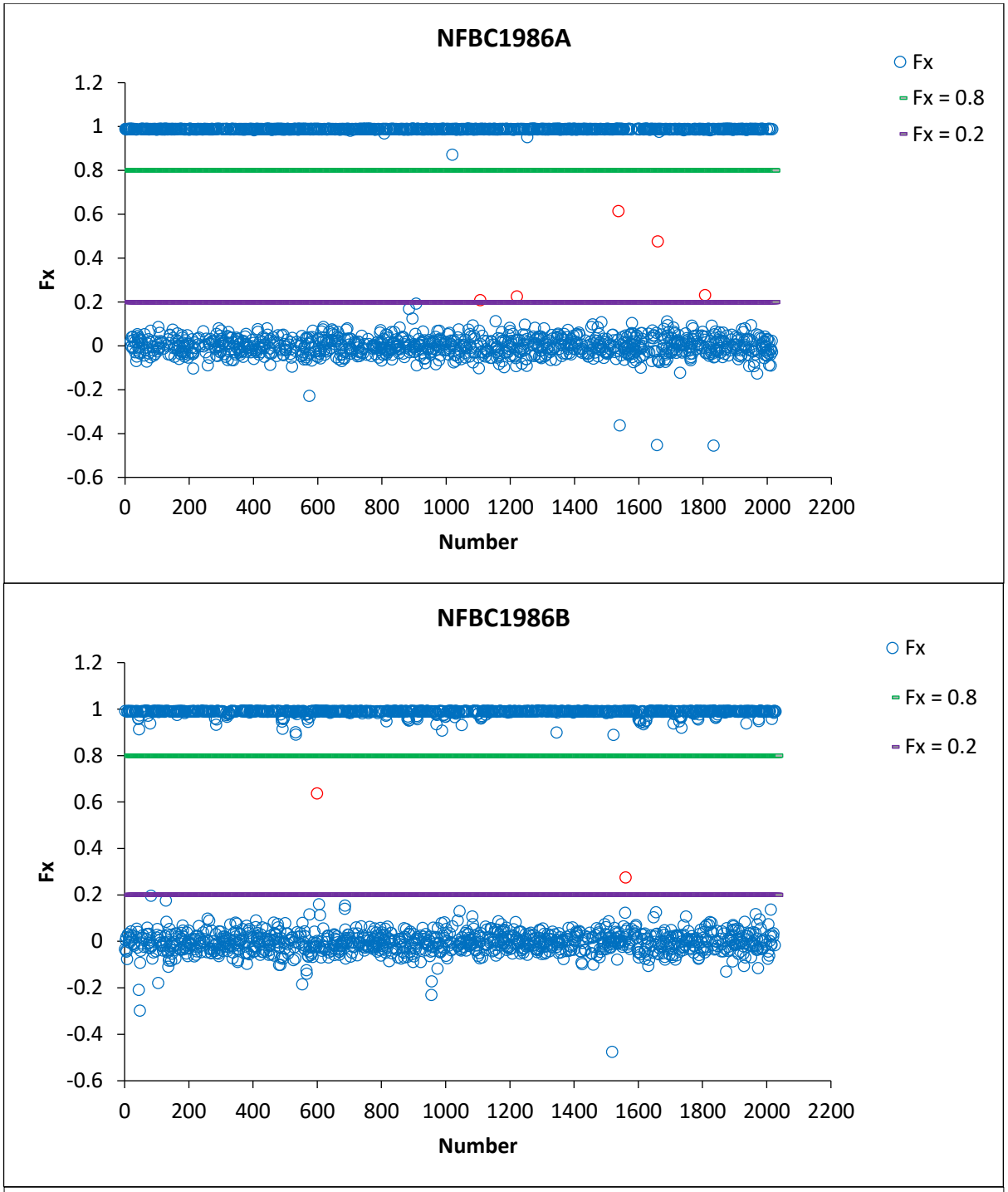


**Figure 4- 3:** Mean homozygosity rates of X-chromosome SNPs (Fx) for NFBC1986A (TOP) and NFBC1986B (BOTTOM) individuals. Each blue circle represents an individual. The green line represents the lower Fx threshold (Fx=0.8) to infer an individual as a male. The purple line represents the upper Fx threshold (Fx=0.2) to infer an individual as a female. Red circles represent individuals whom their sex could not be genetically determined based on genotyping.

**b) Quality control 2: Missing genotype data, high autosomal heterozygosity rates**

The thresholds used for both SNP missingness and Het were described previously (refer to **Chapter 4.4.1 NFBC1966**). As shown in **Figure 4- 4**, 15 (NFBC1986A) and 43 (NFBC1986B) individuals had either reduced or elevated proportion of heterozygous genotypes, or high SNP missingness (>5%), or both. Those individuals were depicted as red circles in **Figure 4- 4**. In summary, a total of 58 individuals (1.43%) from NFBC1986 cohort failed this QC step.

**c) Quality control 3: Sample call rates**

The inclusion threshold for the genotyping call rate used for the NFBC1986 cohort was ≥98%, as described in **Chapter 3.5.2 Processing of genotyped data**. In summary, 19 (NFBC1986A) and 189 (NFBC1986B) individuals were identified to have genotyping call rate <98%. Therefore, a total of 208 individuals (5.14%) from the NFBC1986 cohort failed this step.

**d) Sample exclusions**

**Table 4- 1** shows the summary of sample QC in the NFBC1986 cohort. Twenty-seven samples (n=27) were known to be parents of the other genotyped individuals. Additionally, 21 samples were identified as duplicates. Upon compilation of those individuals who failed the QC steps, a total of 86 individuals failed more than one QC step. Therefore, a total of 263 individuals from the NFBC1986 cohort (6.51%) were excluded prior to population structure adjustment.

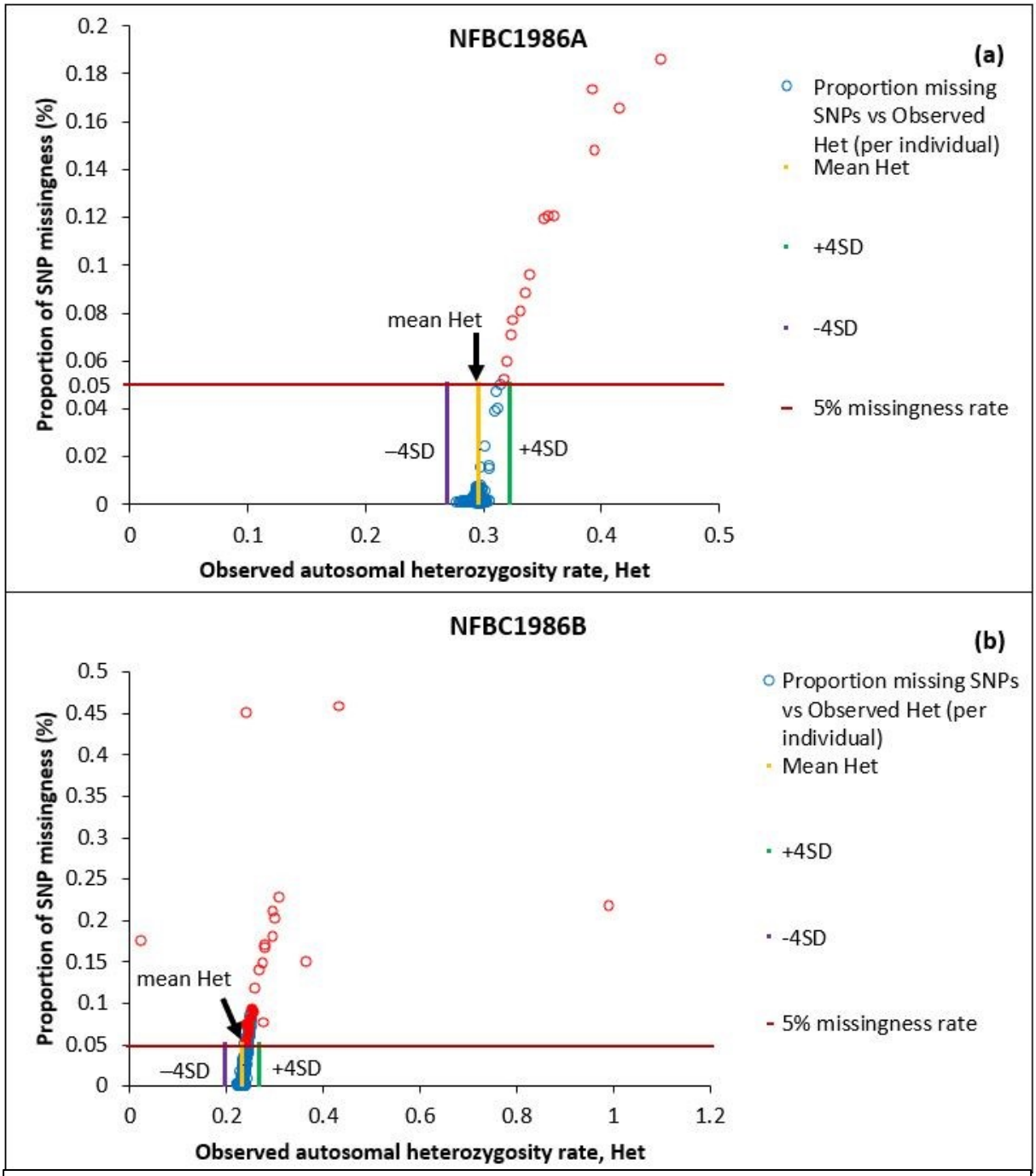


**Figure 4- 4:** A scatter plot of the proportion of SNP missingness (%) against autosomal heterozygosity rate (Het) observed per individual in (a) NFBC1986A, and (b) NFBC1986B batch. Each blue circle represents an individual. The red circles represent the individuals who either had >5% SNP missingness, excessive (>mean+4SD) or reduced (<mean-4SD) proportion of heterozygote genotypes, or both. The red line represents 5% threshold level for SNP missingness.

| Description | NFBC1986A | NFBC1986B |
|---|---|---|
| Samples with sex discordance | 18 | 17 |
| Missingness > 5% | 15 | 43 |
| Outlying heterozygosity | | |
| Call rate < 98% | 19 | 189 |
| Parents | 0 | 27 |
| Duplicated samples | 0 | 21 |
| QC failed | 52 | 297 |
| Shared > 1 failure criteria | 19 | 67 |
| **Final QC failed samples** | **33** | **230** |

**Table 4- 1:** Summary table showing the number of NFBC1986 individuals (genotyped in two batches, namely NFBC1986A and NFBC1986B) who failed each QC step. QC = Quality control

### 4.4.3. NutriTech cohort

The NutriTech cohort (n=74) was genotyped together with the first batch of PMMO cohort (n=21) on Illumina's Human OmniExpress microarray platform. The genotyping for both batches was described in **Chapter 3.5.1 Arrays used in different cohorts**. In order to analyse the individuals from the NutriTech cohort only, all individuals from the PMMO cohort were removed prior to QC checks.

**a) Quality control 1: Sex discordance**

The Fx thresholds used in this cohort were the same as in the previous cohorts. Briefly, Fx = 1 ± 0.2 were classified as males, Fx = 0 ± 0.2 were classified as females, and 0.2 < Fx < 0.8 were classified as undetermined/ambiguous sex. Individuals with ambiguous sex were excluded from further analyses.

Initially, 36 males and 38 females were identified from the clinical data. A total of 17,707 chromosome-X SNPs were presented on the genotyping array, for which were used to determine the sex of each NutriTech individual. **Figure 4- 5** shows the Fx for all individuals of the NutriTech cohort. All samples had Fx of either above 0.8 or below 0.2. Upon cross-checking with the sex in the clinical data, only one individual (sample ID = NT103, Fx = 0.02794) had a mismatch between its clinically and genetically determined sex. In summary, one NutriTech individual (1.35%) was flagged to have sex discordance.

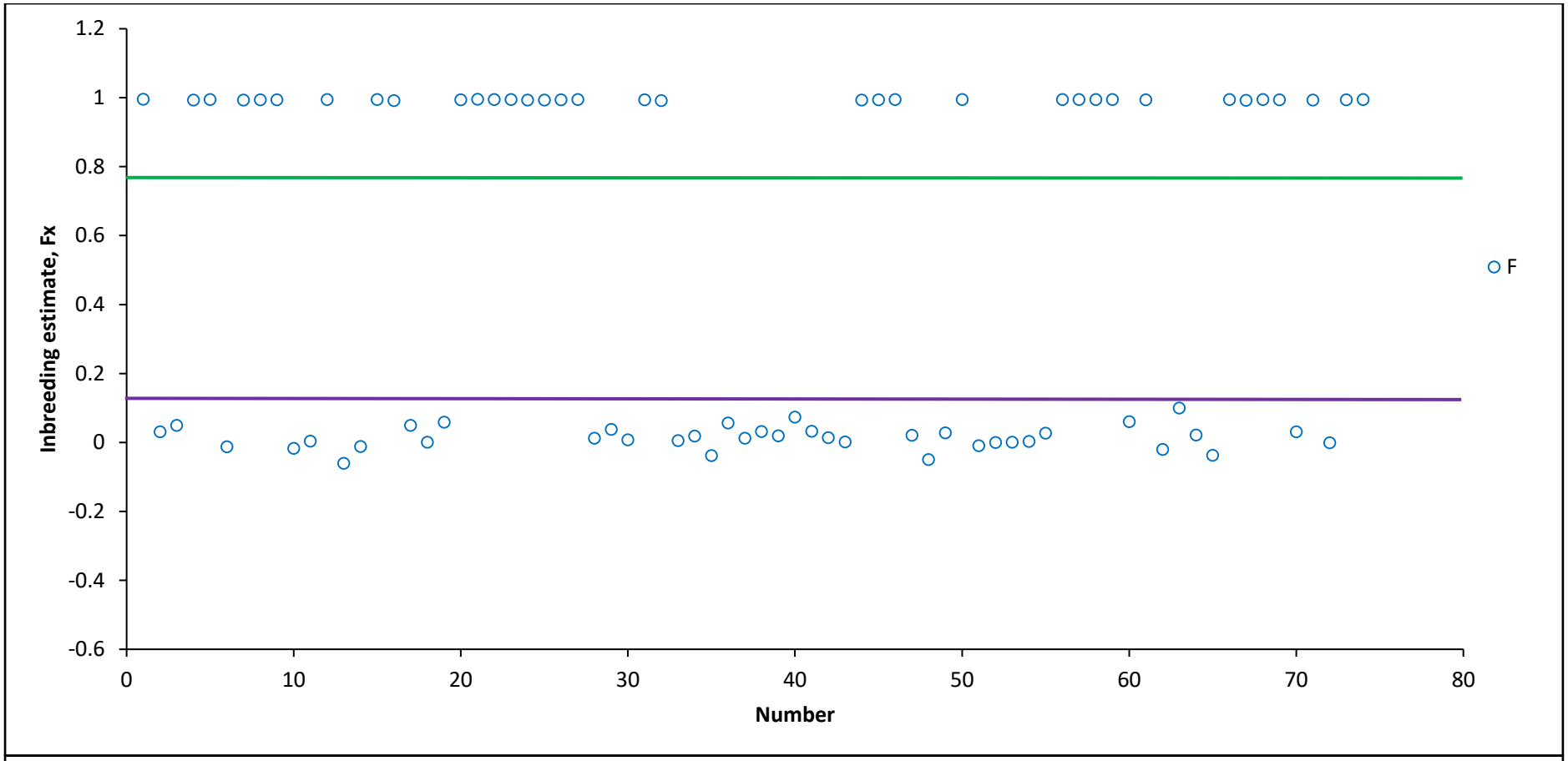


**Figure 4- 5:** X-chromosome inbreeding estimates (Fx) for 74 NutriTech samples. Each blue circle represents an individual. The green line represents the lower Fx threshold (Fx=0.8) to infer an individual as a male. The purple line represents the upper Fx threshold (Fx=0.2) to infer an individual as a female.

**b) Quality control 2: Missing genotype data, high autosomal heterozygosity rates**

The thresholds used for both SNP missingness and distribution of mean heterozygosity are as described in **Chapter 4.4.1 NFBC1966**. As shown in **Figure 4- 6**, only one individual (sample ID = NT103) had either reduced or elevated proportion of heterozygous genotypes, or high proportion of their SNPs not called (>5%), or both (**Figure 4- 6;** red circle). This individual also failed the previous sex discordance step.

**c) Quality control 3: Sample call rates**

The inclusion criterion for genotyping call rate used in this study cohort was ≥99%, a relatively more stringent than the one described in **Chapter 3.5.2 Processing of genotyped data**. This is because the overall genotyping quality of this cohort was very high. Only one NutriTech individual was identified to have a genotyping call rate of <99% (sample ID=NT103). This individual also failed the other QC steps, as stated previously.

**d) Sample exclusions**

A total of 1 individual from the NutriTech cohort failed at least one QC step. No duplicated sample was identified. The same individual failed all QC steps, potentially due to contamination or sampling error (1.35%; sample ID=NT103, genetically female but self-reportedly male). Hence, that individual was excluded prior to the population structure adjustment.

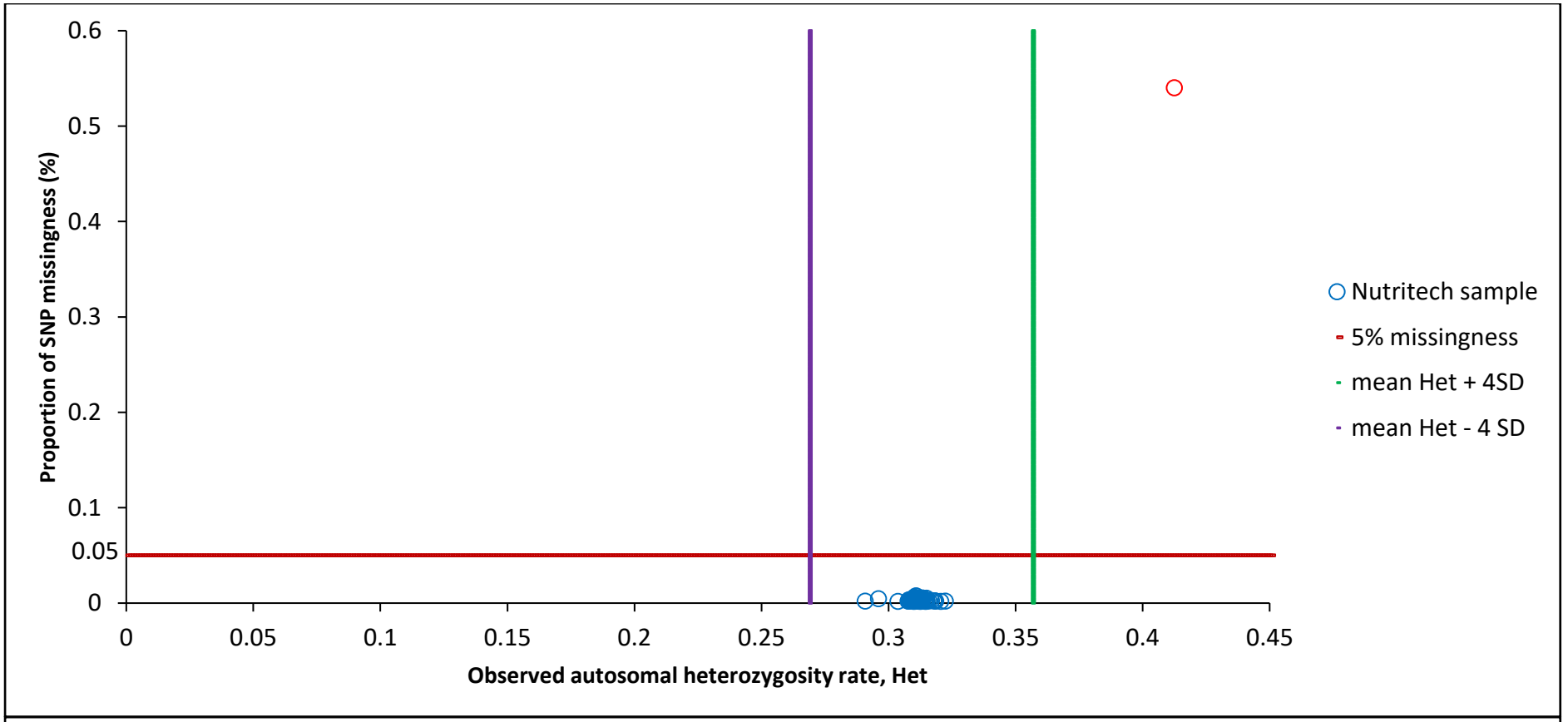


**Figure 4- 6:** A scatter plot of the proportion of SNP missingness (%) against autosomal heterozygosity rate (Het) observed per individual in the NutriTech cohort (n=74). Each blue circle represents an individual. The red circles represent the individuals who either had >5% SNP missingness, excessive (>mean+4SD) or reduced (<mean-4SD) proportion of heterozygote genotypes, or both. The red line represents 5% threshold level for SNP missingness.

### 4.4.4. PMMO cohort

As described in **Chapter 3.5.1 Arrays used in different cohorts**, the first 21 individuals were genotyped together with individuals from the NutriTech cohort on the same genotyping array. Another genotyping was conducted for the second batch of the PMMO cohort (n=96). I carried out the QC steps separately by batch to avoid any confusion, and to reduce batch effects. The QC steps were initially conducted on the second batch. I then merged the first batch (n=21) with the second batch (after QC and population structure adjustments), before conducting another round of QCs for the merged dataset. Thus, the following results are for the second batch of the PMMO cohort, unless otherwise stated.

**a) Quality control 1: Sex discordance**

The thresholds used for Fx are as described in **Chapter 4.4.1 NFBC1966**. Ninety-six individuals (n=96) consisting 36 males and 60 females were identified from the clinical database. A total of 22,691 chromosome-X SNPs were used to determine the sex of each PMMO individual from the second batch. **Figure 4- 7** shows that all individuals genotyped in the second batch of the PMMO cohort (n=96) had Fx of either above 0.8 or below 0.2, indicating that each individual had their sex ascertained from their genotyped data. Moreover, the ascertained sex were the same as reported in their clinical data. A detailed inspection on an individual (**Figure 4- 7**; orange circle) determined as a female but had a relatively low Fx than the mean Fx for females (mean ± 4SD = -0.003 ± 0.255), revealed no information on any possible sex-linked disorders being reported. In addition, the overall SNP call rate for this individual was high (>99%). Therefore, this individual was classified as a true female. In summary, none of them was found to have any sex discordance.

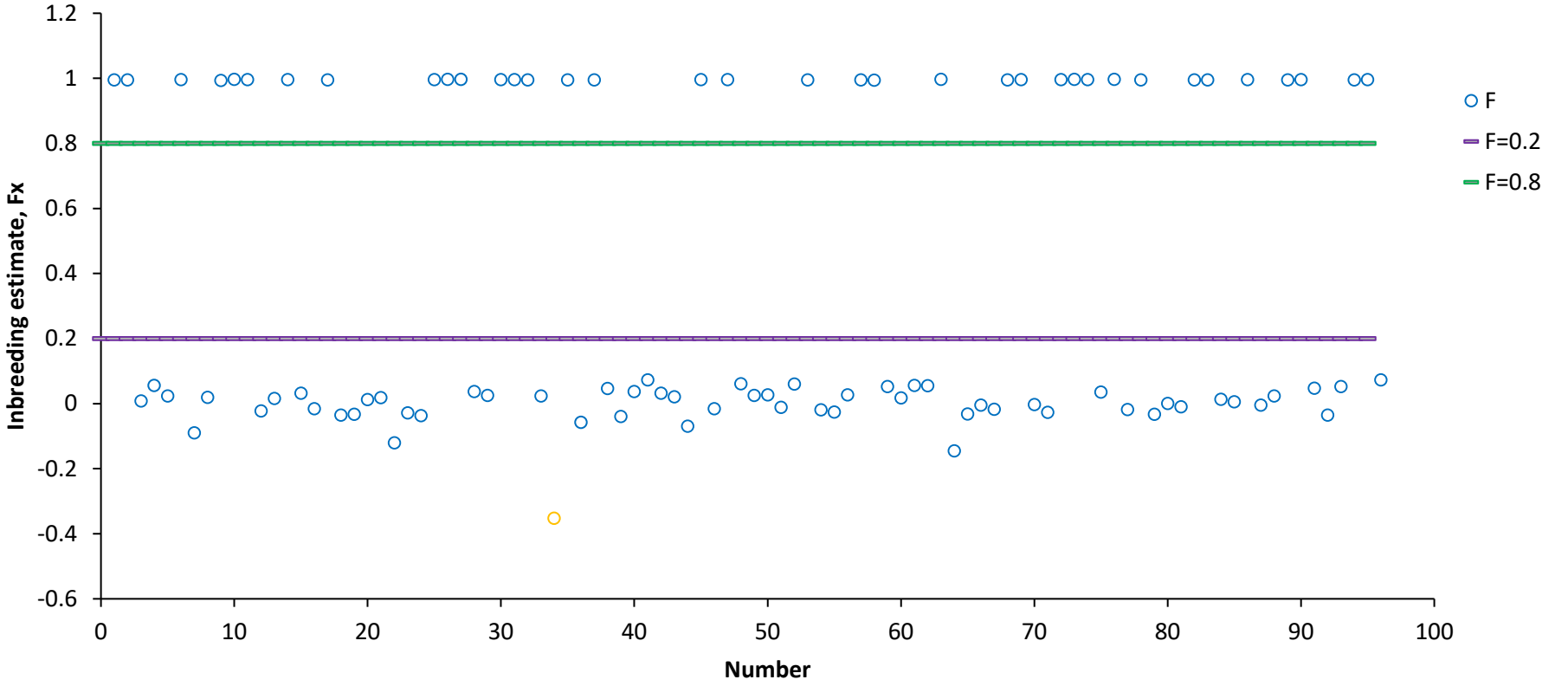


**Figure 4- 7:** X-chromosome inbreeding estimates (Fx) for 96 individuals from the second genotyping batch of PMMO cohort. Each blue circle represents an individual. The green line represents the lower Fx threshold (Fx=0.8) to infer an individual as a male. The purple line represents the upper Fx threshold (Fx=0.2) to infer an individual as a female. A genetically inferred female who had its Fx relatively lower than the mean Fx for females (orange circle) was assessed in detail to determine that the inferred sex was actually true.

**b) Quality control 2: Missing genotype data, high autosomal heterozygosity rates**

The thresholds used for both SNP missingness and distribution of mean heterozygosity are as described in **Chapter 4.4.1 NFBC1966**. **Figure 4- 8** shows that only one individual (n=1) had either reduced or elevated proportion of heterozygous genotypes, or high proportion of their SNPs not called (>5%), or both (**Figure 4- 8**; red circle).

**c) Quality control 3: Sample call rates**

The inclusion criterion for the genotyping call rate used in this study cohort was ≥99%, a relatively more stringent than the one described in **Chapter 3.5.2 Processing of genotyped data**. Just like in the NutriTech cohort, the overall genotyping quality of this cohort was very high. In summary, all individuals in the second genotyping batch of the PMMO cohort (n=96) had call rates > 99%.

**d) Sample exclusions**

A total of only 1 individual from the second genotyping batch of the PMMO cohort failed at least one QC step (1.04%). No duplicated sample was identified. Hence, that individual was excluded prior to population structure adjustment.

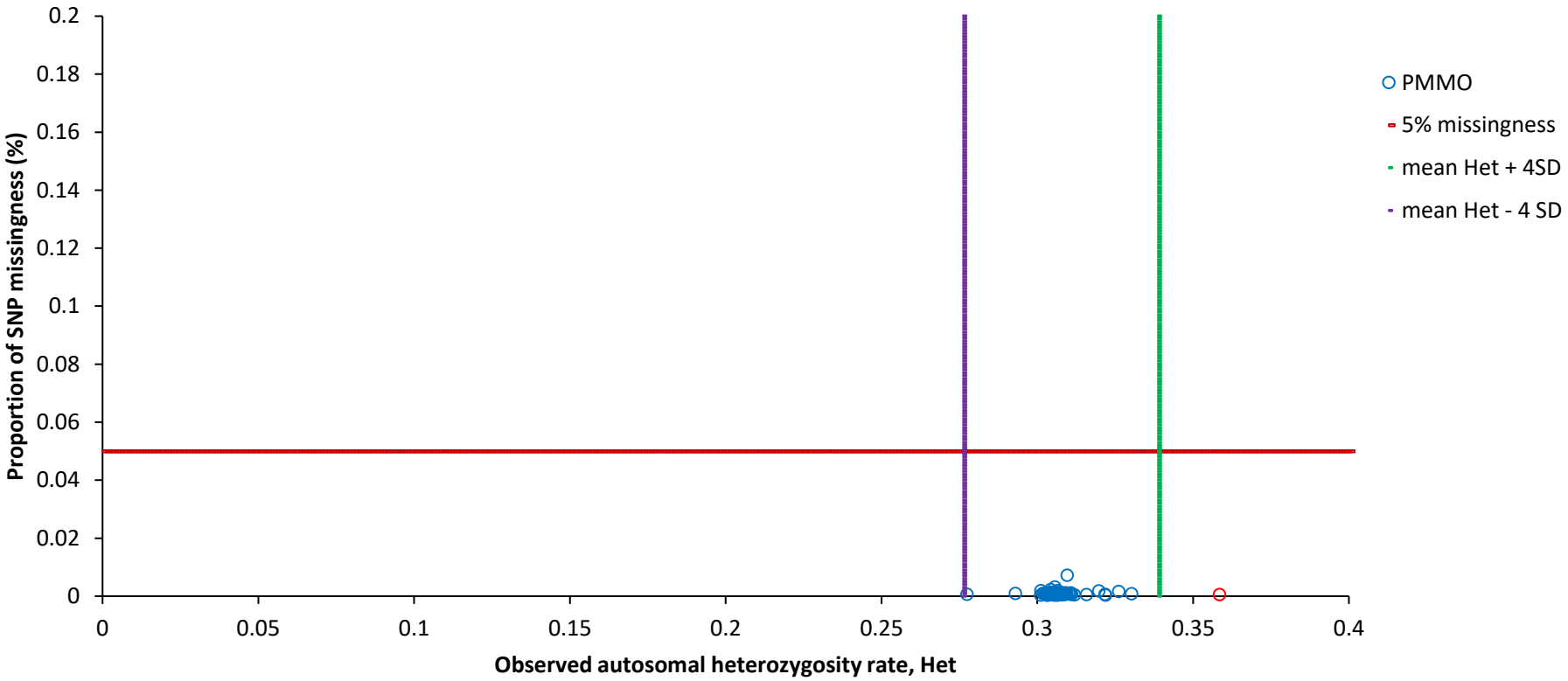


**Figure 4- 8:** A scatter plot of the proportion of SNP missingness (%) against autosomal heterozygosity rate (Het) observed per individual in the second genotyping batch of the PMMO cohort (n=96). Each blue circle represents an individual. The red circle represents the individuals who either had >5% SNP missingness, excessive (>mean+4SD) or reduced (<mean-4SD) proportion of heterozygote genotypes, or both. The red line represents 5% threshold level for SNP missingness.

### 4.4.5. UK Biobank cohort

As mentioned in the previous chapter, the sample QCs for the UKBB cohort were performed by Andrianos Yiorkas (Department of Medicine, Imperial College London, UK). Initially, a total of 107 batches of approximately 4,700 gDNA samples per batch (n=494,078) were sent for SNP genotyping (refer **chapter 3.5.1 Arrays used in different cohorts**).

A total of 4,866 participants (<1%) failed internal QCs from Affymetrix, and 835 participants (<0.2%) were identified as either sample duplicates, involved in laboratory sample handling errors, or withdrawn participants(Clare Bycroft *et al.*, 2017). Therefore, a total of 5,701 individuals (1.1%) were removed prior to sample QCs. The full genotyped data for 488,377 participants (~98.8%) underwent sample QCs.

Briefly, a total of 968 participants (~0.2%) were identified to either have low genotyping call rates (genotyping call rates <95% or SNP missingness rate >5%) or extreme heterozygosity rates. In addition, a total of 378 individuals (<0.1%) were identified to have sex mismatches between self-reported sex and genetically-determined sex. As stated by Bycroft C., *et. al.* (2017), a total of 652 participants (~0.1%) were later identified to have sex chromosome aneuploidies(Clare Bycroft *et al.*, 2017). Furthermore in **Chapter 3.6.6 Quality controls in the UK Biobank**, I mentioned that only participants with white British ancestry (i.e. self-reported 'British' or had genetically similar British ancestry) were included. Thus, a total of 78,674 participants (~16.1%) were identified as outliers from the white British ancestry group. As instructed by the UK Biobank(Clare Bycroft *et al.*, 2017), 9 more participants were removed from the dataset because they withdrew from the cohort. A total of 553 participants failed more than one QC step. In conclusion, a total of 80,130 participants (16.4%) were subsequently removed due to failing genotyping QCs.

# 4.5. Results: Population structure adjustments

The presence of inbreeding and high cryptic relatedness in a population could introduce a low concentration of genetic variations within the genetic pool of the population. This could drive the discovery of any genetic differences towards the ancestral background, thereby increasing the possibility of false discoveries in genetic studies (Zhang and Deng, 2010). In addition, ethnic heterogeneity could include subgroups with different allele frequencies in the genetic pool of a population, increasing the likelihood of false positive discoveries in genetic association studies (Wang *et al.*, 2018). Therefore, it is important to identify any individuals who were products of inbreeding (through inbreeding coefficients), genetically related to one another (through identity-by-descent estimation), or from an ethnic group that was different from the other individuals within the same study cohort (through multidimensional scaling).

## 4.5.1. NFBC1966 cohort

### a) Adjustment 1: Inbreeding

After removing data that failed the QCs, a total of 5,398 samples from the NFBC1966 cohort were subjected to population structure adjustments. As described in **Chapter 3**, the first step I took in adjusting for the population structure was to identify inbreeding within the NFBC1986 cohort. Inbreeding can significantly affect the overall fitness of a species in a population due to increased homozygous genotypes within the gene pool of that population.

According to the methods described in **Chapter 3.7 Population structure adjustments**, autosomal inbreeding coefficients, F, were estimated for all individuals who passed the QCs.

Due to poor overall genotyping quality and low SNP resolution in terms of its content, I used 5SD away from the mean F as the upper and lower inclusion thresholds for the NFBC1966 cohort. A total of 87,261 LD-pruned, autosomal SNPs were used to calculate the F per individual. **Figure 4- 9** shows the inbreeding estimates for all NFBC1966 individuals who passed the QCs (n=5,398). One sample (n=1) was identified to have an F below the lower threshold (. **Figure 4- 9**; red circle) and was excluded from the analysis.

**b) Adjustment 2: Pair-wise IBD Estimation**

Pair-wise identity-by-descent (IBD) estimation can identify pedigree errors and unknown familial relationships by the detection of specific DNA stretches that are shared between distantly-related individuals within the same study cohort. The methods are well-described in **Chapter 3.7 Population structure adjustments**.

Initially, there were 1,484 IBD pairs with P_HAT ≥ 0.2 (i.e. at least 20% of the genome shared between the individuals from each pair). Each pair was inspected for their call rates and sex. A few criteria were used to determine the individuals to be excluded based on the IBD estimations: 1) individuals with lower call rates than their respective IBD pairs, and 2) in an event where the call rates were similar between the IBD pairs, individuals were excluded based on sex such that the number of males and females of the excluded individuals were balanced. In addition, several individuals were found to have different IBD pairs. Therefore, a reduction process was conducted to address the issue. The priority criterion is described in **Chapter 3.7.3 Adjustment 2: Pair-wise Identity-by-descent (IBD) Estimation.** In summary, a total of 106 individuals (1.96%) were identified to be genetically related to at least one individual from the same cohort.

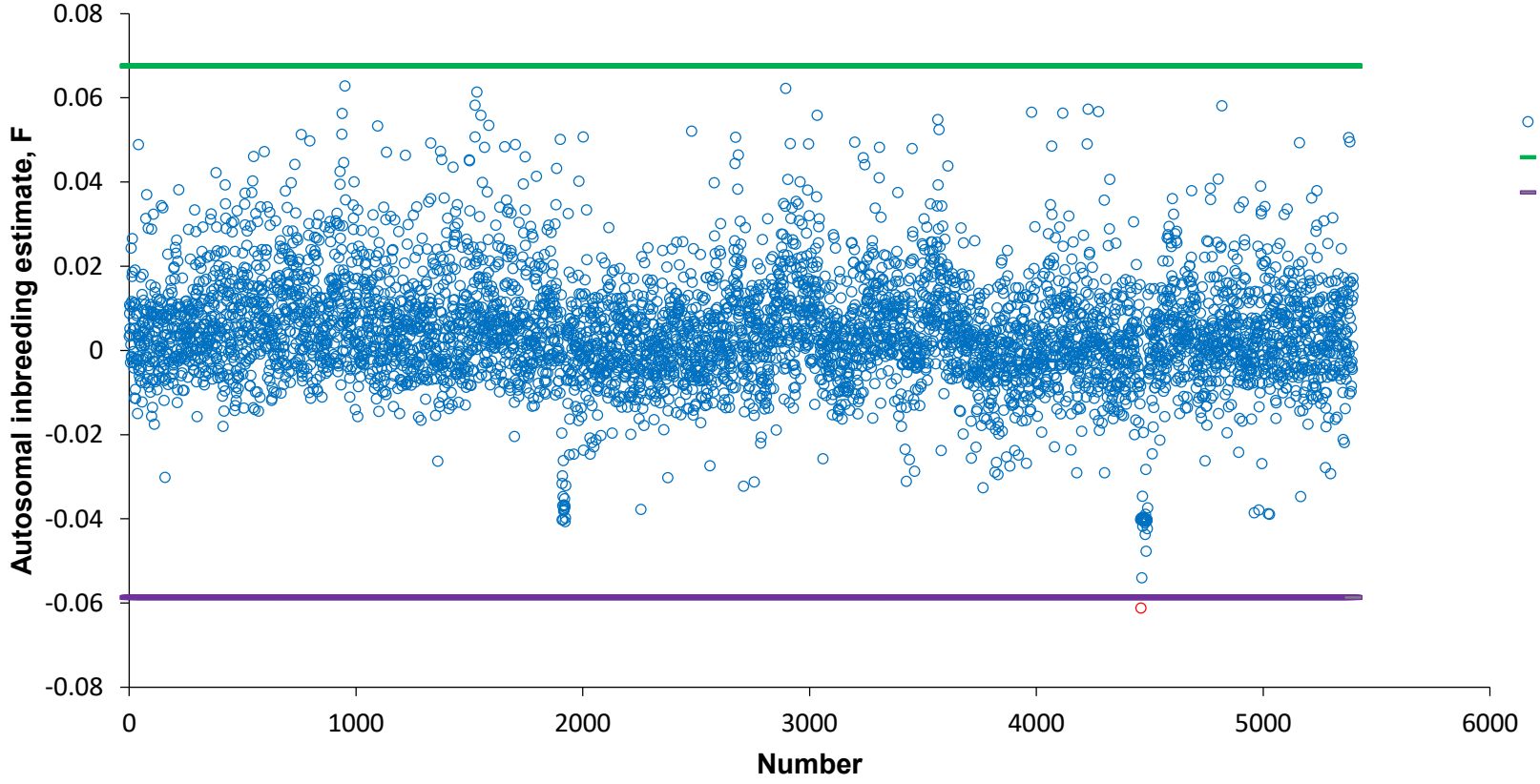


**Figure 4- 9:** Autosomal inbreeding estimates (F) for 5,398 individuals from the NFBC1966 cohort. Each blue circle represents an individual. The green line represents the upper F inclusion threshold. The purple line represents the lower F inclusion threshold. The red circle represents the samples with high inbreeding estimates.

**c) Adjustment 3: Multidimensional Scaling**

Multidimensional scaling (MDS) can be used as a method to identify individuals who belong to an ethnic subgroup that is different from the rest within a specific population structure. The MDS was carried out to identify different dimensions within the genotyped data of a study cohort. Each dimension was considered as a mathematical representation of a 'common' component shared between the individuals, for which it defined a specific part of the genetic structure within a study cohort. In the NFBC1966 cohort, all individuals who survived the QCs (n=5,398) were subjected to MDS according to the methods described in **Chapter 3.7 Population structure adjustments.** A total of 82,118 common SNPs (MAF>0.01) shared between the NFBC1966 cohort and 9 HapMap Phase 3 populations, were used to calculate the MDS. **Figure 4- 10** depicts the first two dimensions for each individual, plotted against each other. A total of 54 NFBC1966 individuals were found to be outliers from their population cluster (**Figure 4- 10**; red circles).

**d) Sample exclusions**

A total of 161 individuals failed at least one of the steps in adjusting for population structure. Among them, 54 individuals failed more than one step (duplicated samples). In summary, 107 NFBC1966 individuals (1.98%; male=78, female=29) were excluded prior to any statistical analyses.

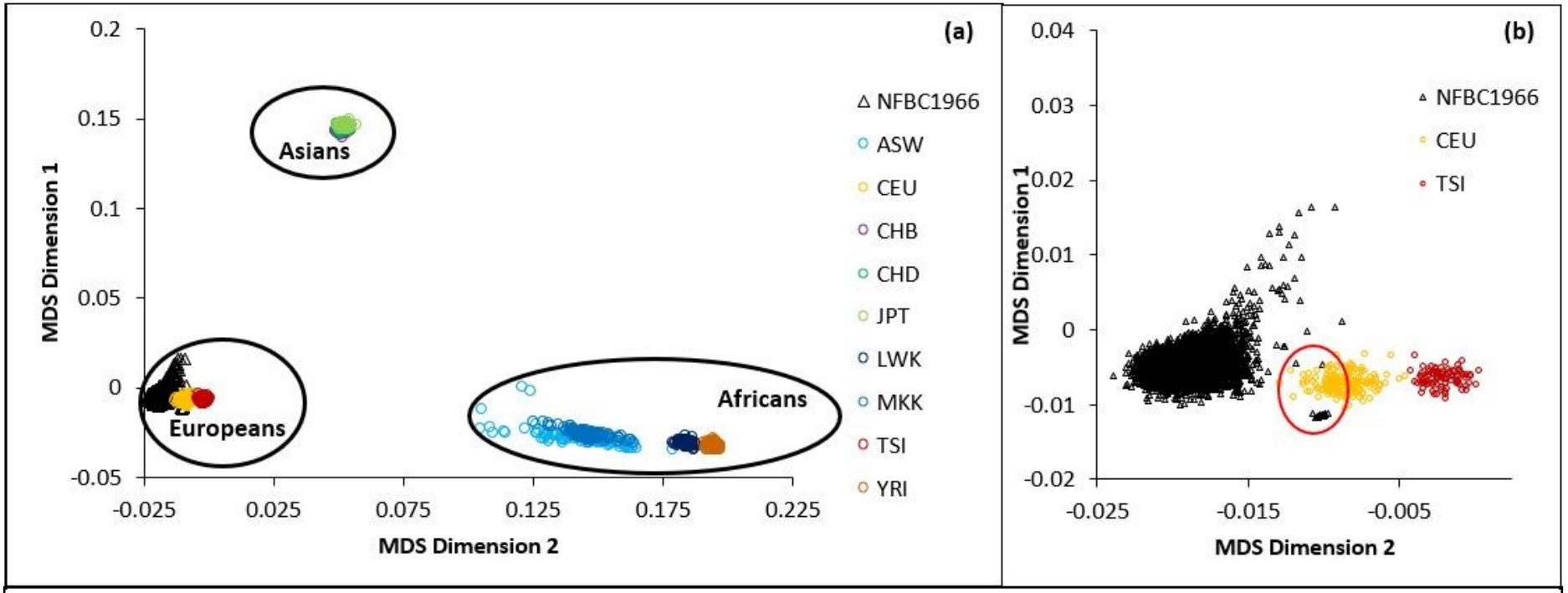


**Figure 4- 10: (a)** Results of multidimensional scaling (MDS) of the NFBC1966 cohort (black triangles; n=5,398) anchored by 9 major HapMap Phase 3 populations. A clear distinction of populations separated according to 3 major continents. **(b)** A zoomed-in MDS plot of the NFBC1966 and two European populations (CEU and TSI) that shows two NFBC1966 individuals within the CEU cluster, and a sub-cluster of 52 NFBC1966 individuals were visually detached from their cluster within the European continent (red circle). *[ASW: African ancestry in Southwest USA; YRI: Yoruba in Ibadan, Nigeria; CEU: Utah residents European ancestry from CEPH; CHB: Han Chinese in Beijing, China; CHD: Chinese in Denver, Colorado USA; JPT: Japanese in Tokyo, Japan; LWK: Luhya in Webuye, Kenya; MKK: Maasai in Kinyawa, Kenya; TSI: Toscani in Italy].*

### 4.5.2. NFBC1986 cohort

Here, I would like to remind readers that the NFBC1986 cohort was genotyped in two batches. According to **Table 4- 1**, a total of 33 and 230 individuals were removed from the NFBC1986A and NFBC1986B batches respectively. This has reduced the number of individuals for the population structure adjustments to 1,983 (NFBC1986A) and 1,797 individuals (NFBC1986B).

**a) Adjustment 1: Inbreeding**

For both batches, the upper and lower F thresholds (mean ± 4SD) were calculated according to the methods described in **Chapter 3.7 Population structure adjustments**. A total of 167,196 and 167,895 autosomal SNPs (pruned for LD) were used to calculate the F for each individual from the NFBC1986A and NFBC1986B batches, respectively. According to **Figure 4-11**, 8 (from NFBC1986A) and 9 (from NFBC1986B) individuals were identified to have high inbreeding estimates.

**b) Adjustment 2: Pair-wise IBD Estimation**

The methods to estimate pair-wise IBD among individuals from both batches of the NFBC1986 cohort are described in **Chapter 3.7 Population structure adjustments**. Initially, 9 and 3 IBD pairs with P_HAT ≥ 0.2 were identified for NFBC1986A and NFBC1986B respectively. Each pair was inspected for their call rates and sex. Priority criteria (as described for the NFBC1966 cohort) were also used. Since both individuals from each pair had high call rates (>99%), I selected the samples for exclusion based on the male-to-female ratio.

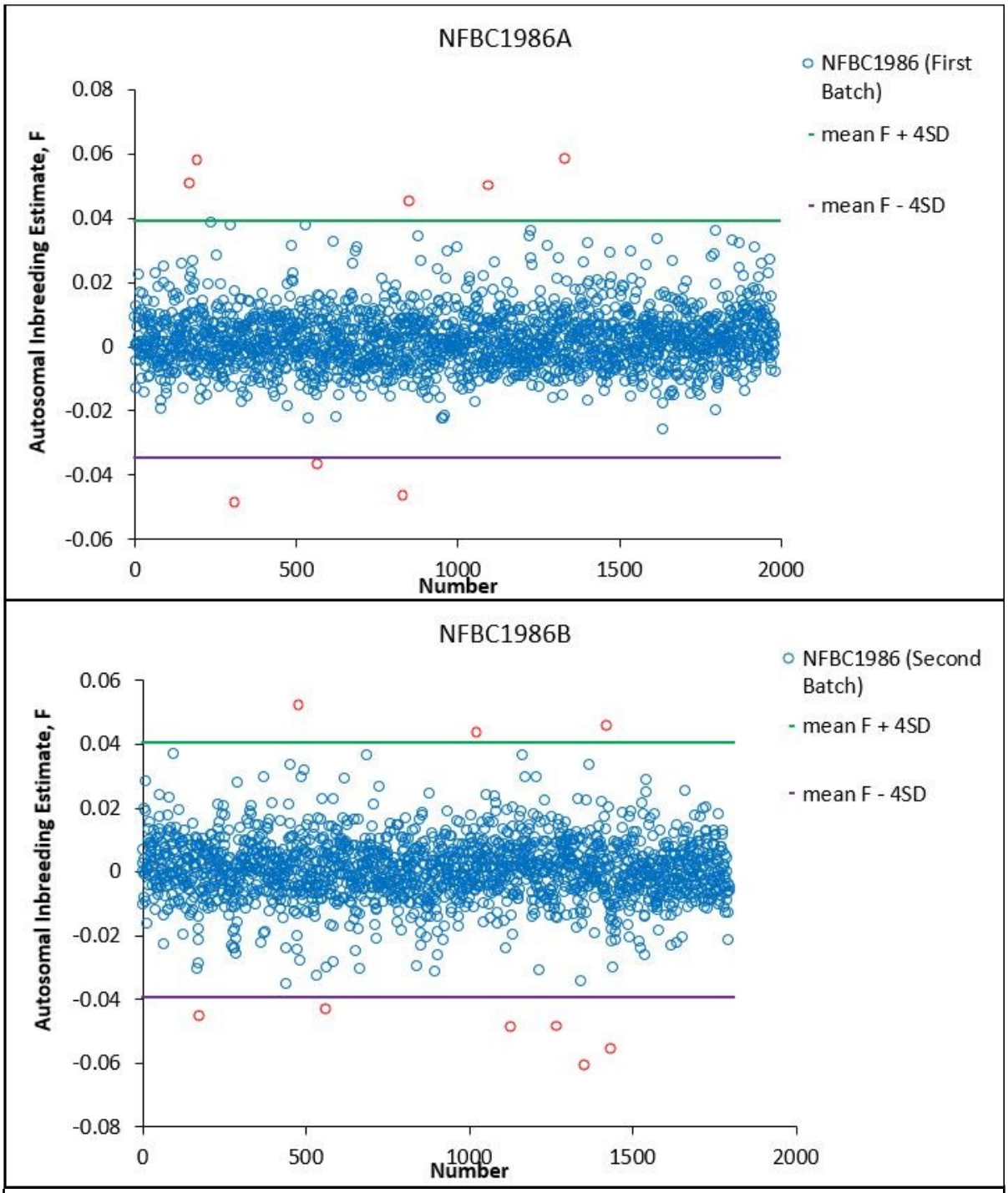


**Figure 4- 11:** Autosomal inbreeding estimates (F) for 1,983 and 1,797 individuals from the NFBC1986A and NFBC1986B cohorts respectively. Each blue circle represents an individual. The green line represents the upper F inclusion threshold. The purple line represents the lower F inclusion threshold. The red circles samples with high inbreeding estimates.

For the NFBC1986A batch, 116,564 independent SNPs (MAF>1%) were used for estimating the pair-wise IBD. A total of 4 males and 4 females were initially selected for exclusion. Since there were 3 females more than males in the post-QC dataset (**Table 4- 1**), I selected another female from the final IBD pair for exclusion. In summary, a total of 4 males and 5 females were identified to be genetically related to another individual within the same batch, and thus they were selected for exclusion.

For the NFBC1986B batch, 117,452 independent SNPs (MAF>1%) were used for pair-wise IBD estimation. A total of 3 IBD pairs were identified, with the first two pairs being male-to-male pairs. The final pair was a male-to-female pair, which meant that there were 5 males and 1 female among the genetically related pairs. By abiding to the same priority criteria, 2 males from the first two IBD pairs were randomly selected for exclusion, along with the only female from the final pair.

**c) Adjustment 3: Multidimensional Scaling**

The identification of population outliers in NFBC1986 cohort was conducted separately in NFBC1986A and NFBC1986B batches. A total of 1,983 NFBC1986A and 1,797 NFBC1986B individuals were subjected to MDS according to the methods described in **Chapter 3.7 Population structure adjustments. Figure 4- 12** and **Figure 4- 13** show scatter plots of the first two MDS dimensions for each individual in the NFBC1986A and NFBC1986B batch respectively, anchored by 9 HapMap Phase 3 population data. One individual from NFBC1986A batch (**Figure 4- 12**; red-circled) and 6 individuals from NFBC1986B batch (**Figure 4- 13**; red-circled) were identified as outliers.

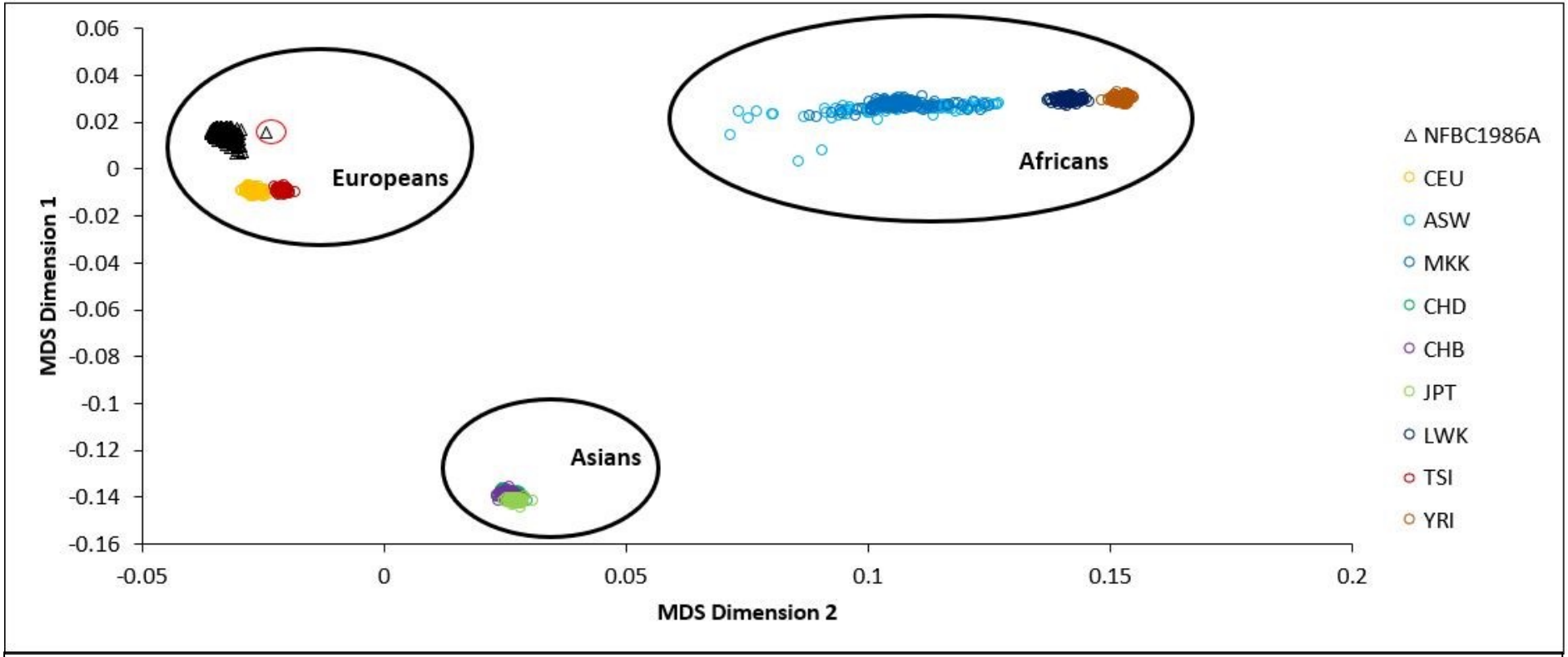


**Figure 4- 12:** Results of multidimensional scaling (MDS) of the NFBC1986A cohort (Black triangles; n=1,983) anchored by 9 major HapMap Phase 3 populations. A clear distinction of populations separated according to 3 major continents. One NFBC1986A individual who failed inbreeding check, was also identified as an outlier (red-circled) as it is detached from its MDS population cluster. *[ASW: African ancestry in Southwest USA; YRI: Yoruba in Ibadan, Nigeria; CEU: Utah residents European ancestry from CEPH; CHB: Han Chinese in Beijing, China; CHD: Chinese in Denver, Colorado USA; JPT: Japanese in Tokyo, Japan; LWK: Luhya in Webuye, Kenya; MKK: Maasai in Kinyawa, Kenya; TSI: Toscani in Italy].*

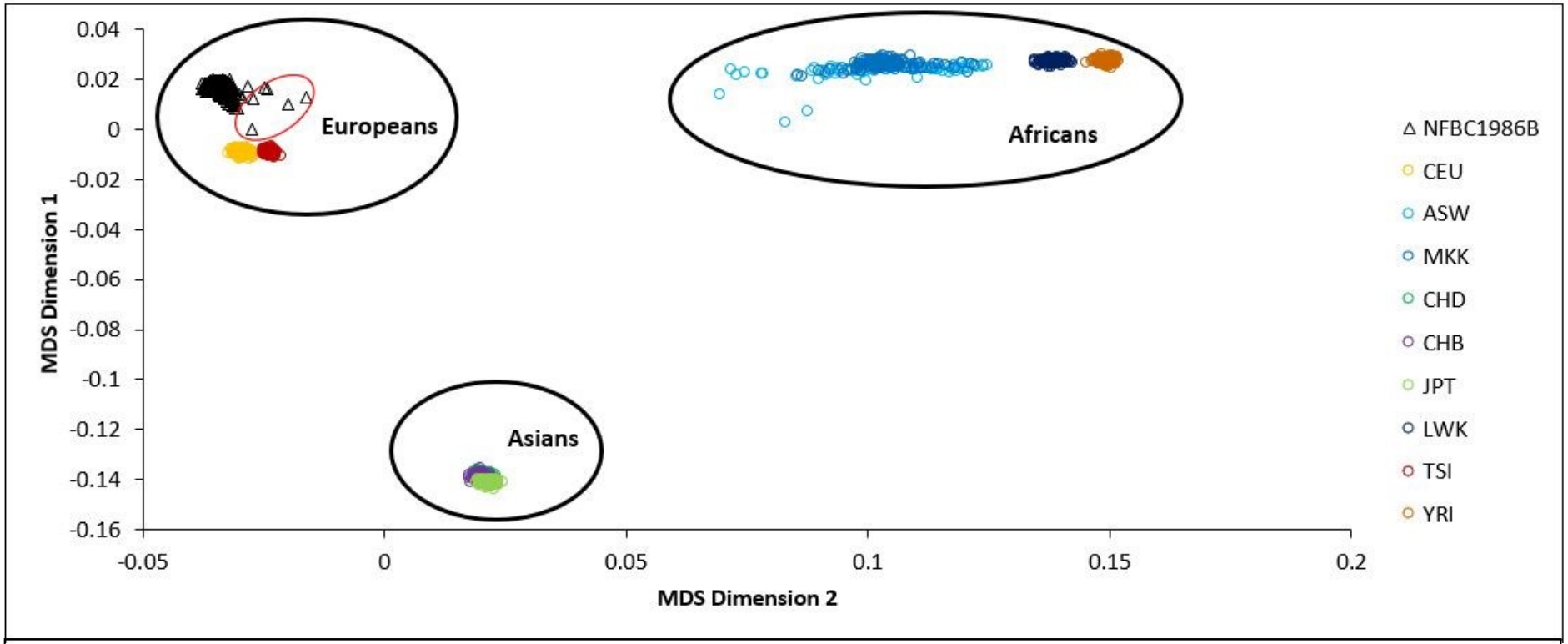


**Figure 4- 13:** Results of multidimensional scaling (MDS) of the NFBC1986B cohort (Blue circles; n=1,797) anchored by 9 major HapMap Phase 3 populations. A clear distinction of populations separated according to 3 major continents. Six NFBC1986B individuals were considered as outliers, including 4 with high inbreeding coefficients (red-circled). *[ASW: African ancestry in Southwest USA; YRI: Yoruba in Ibadan, Nigeria; CEU: Utah residents European ancestry from CEPH; CHB: Han Chinese in Beijing, China; CHD: Chinese in Denver, Colorado USA; JPT: Japanese in Tokyo, Japan; LWK: Luhya in Webuye, Kenya; MKK: Maasai in Kinyawa, Kenya; TSI: Toscani in Italy].*

**d) Sample exclusions**

For the NFBC1986A batch, a total of 18 samples failed at least one of the population structure adjustments. One duplicated sample was identified and removed as it failed in more than one adjustment. In summary, 17 NFBC1986A individuals (0.86%; male = 8, female = 9) had to be excluded from further analyses.

For the NFBC1986B batch, a total of 18 samples failed at least one of the adjustments, with 4 of them (duplicated samples) failed in more than one adjustment. In summary, 14 NFBC1986B individuals (0.78%; male = 6, female = 8) had to be excluded from further analyses.

### 4.5.3. Merged NFBC1986 cohort

After excluding individuals who failed at least one of the population structure adjustments, the two batches of NFBC1986 were merged based on the SNPs shared between both batches. The resulting NFBC1986 cohort consists of 1,966 and 1,783 individuals from NFBC1986A and NFBC1986B respectively (n = 3,749; 49% male; 964,063 SNPs). This new NFBC1986 cohort was subjected to population structure adjustments to identify inbreeding, cryptic relatedness and ethnic outliers between the individuals in NFBC1986A and NFBC1986B.

**a) Adjustment 1: Inbreeding**

The upper and lower F thresholds were according to the methods described in **Chapter 3.7 Population structure adjustments**. The calculation for F for each individual was based on 175,891 autosomal SNPs (pruned for LD). According to **Figure 4- 14**, 3 individuals had their F outside the pre-determined thresholds (red circles).

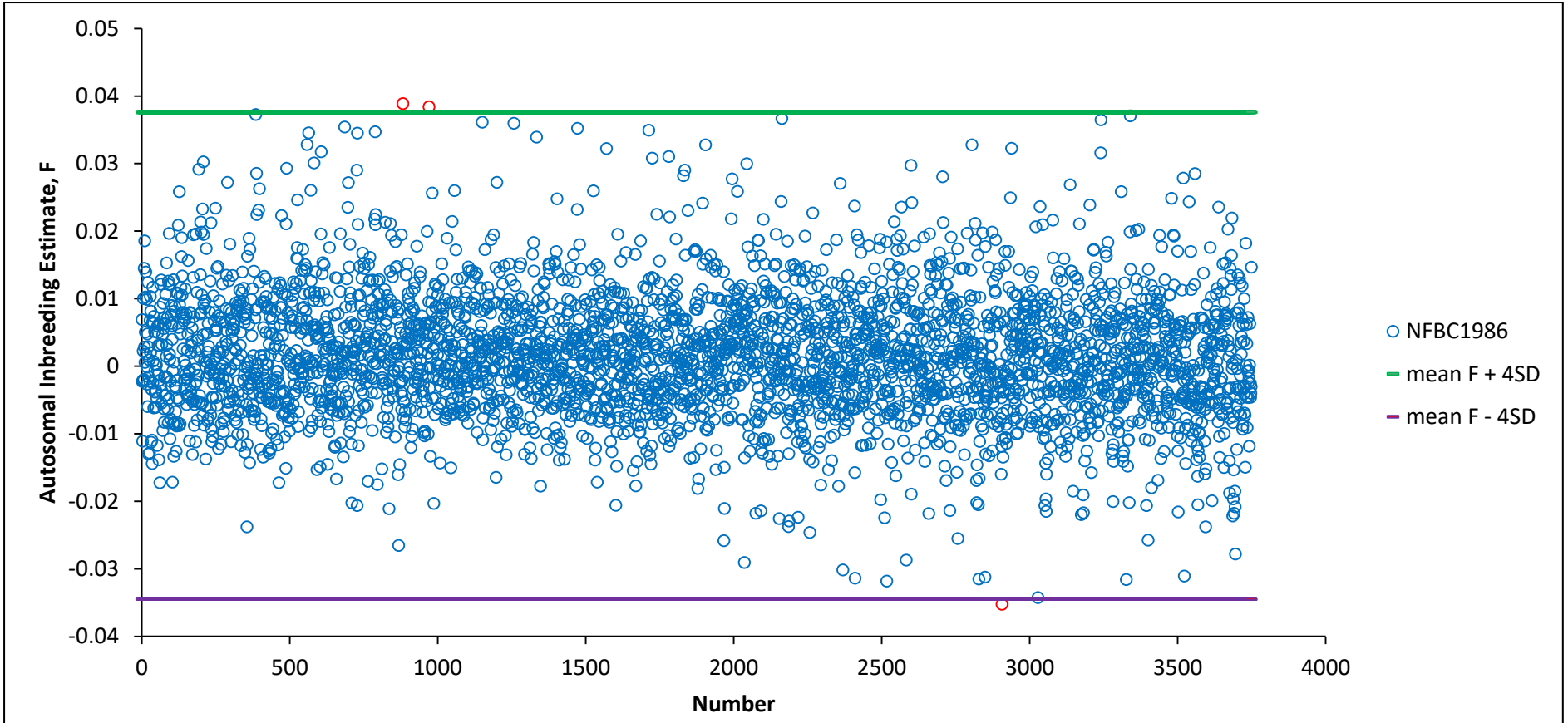


**Figure 4- 14:** Autosomal inbreeding estimates (F) for 3,749 individuals from the NFBC1986 cohort (merged). Each blue circle represents an individual. The green line represents the upper F inclusion threshold. The purple line represents the lower F inclusion threshold. The red circles represent the samples with high inbreeding estimates.

**b) Adjustment 2: Pair-wise IBD Estimation**

Ten IBD pairs with P_HAT ≥ 0.2 were identified for the merged NFBC1986 cohort. As expected each pair consisted of individuals from both batches. Each pair was inspected for their call rates and sex. The same priority criteria were used to select the individual to be excluded from each pair. A total of 6 males and 4 females were selected in a balanced 1:1 ratio between NFBC1986A and NFBC1986B batches.

**c) Adjustment 3: Multidimensional Scaling**

Multidimensional scaling was conducted for 3,749 individuals from the NFBC1986 cohort according to the methods described in **Chapter 3.7 Population structure adjustments. Figure 4- 15** shows a scatter plot of the first two MDS dimensions for the NFBC1986 cohort, anchored by the 9 HapMap Phase 3 populations based on 129,908 common SNPs (MAF>0.01) that were shared between all populations. The cluster of the NFBC1986 cohort (**Figure 4- 15**; blue circles) was concentrated without any visible outlier present.

**d) Sample exclusions**

In total, 13 individuals from the merged NFBC1986 cohort failed one of the population structure adjustments (no duplicated samples). These individuals had to be excluded from further analyses. After sample exclusion, the merged NFBC1986 cohort is now consisting of 3,736 individuals (49% male, 983,242 SNPs).

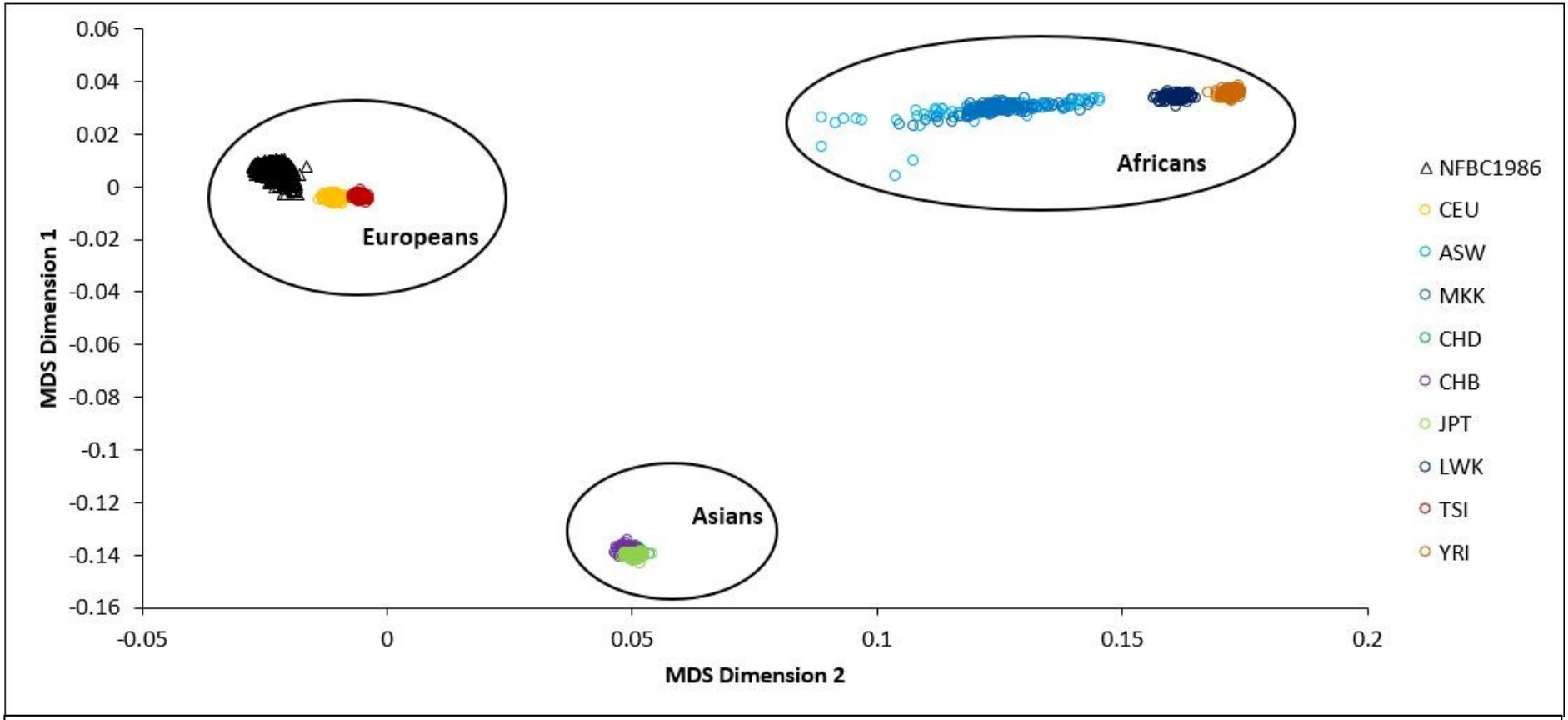


**Figure 4- 15:** Results of multidimensional scaling (MDS) of the merged NFBC1986 cohort (Black triangles; n=3,749) anchored by 9 major HapMap Phase 3 populations. A clear distinction of populations separated according to 3 major continents. *[ASW: African ancestry in Southwest USA; YRI: Yoruba in Ibadan, Nigeria; CEU: Utah residents European ancestry from CEPH; CHB: Han Chinese in Beijing, China; CHD: Chinese in Denver, Colorado USA; JPT: Japanese in Tokyo, Japan; LWK: Luhya in Webuye, Kenya; MKK: Maasai in Kinyawa, Kenya; TSI: Toscani in Italy].*

### 4.5.4. NutriTech cohort

After removing 1 individual who failed the QCs, a total of 73 samples from the NutriTech cohort were subjected to the population structure adjustments.

**a) Adjustment 1: Inbreeding**

The upper and lower F thresholds (mean ± 4SD) used for exclusion were calculated according to the thresholds stated in **Chapter 3.7 Population structure adjustments**. A total of 121,252 autosomal SNPs (pruned for LD) were used to calculate the F for each NutriTech individual. **Figure 4- 16** shows the F for all NutriTech individuals who passed the QCs (n=73). One individual was identified to have a high inbreeding estimate due to having an F higher than the upper threshold (**Figure 4- 16**; red circle).

**b) Adjustment 2: Pair-wise IBD Estimation**

Pair-wise IBD estimation was conducted according to the methods described in **Chapter 3.7 Population structure adjustments**. In summary, all individuals from the NutriTech cohort (n=73) were found to be genetically unrelated to one another (P_HAT < 0.2).

.

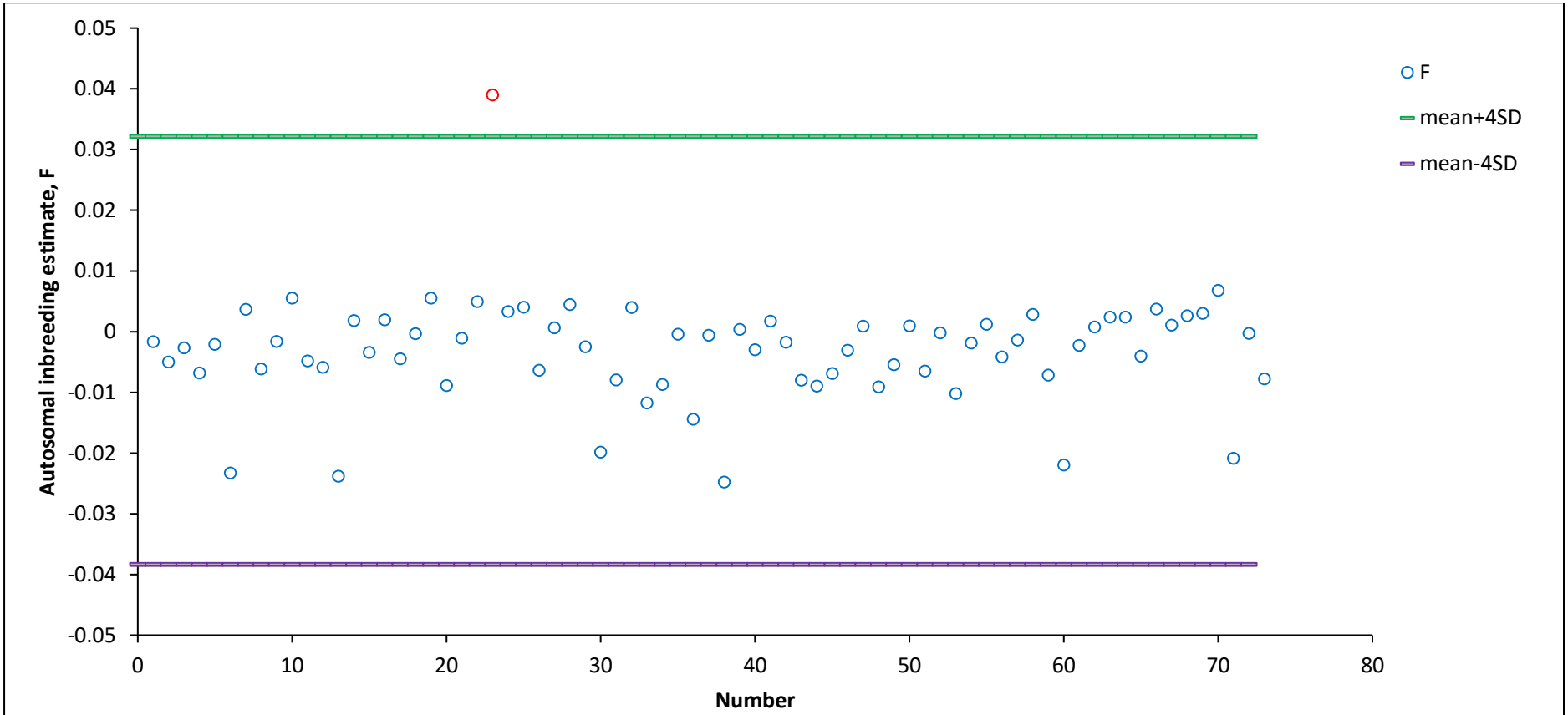


**Figure 4- 16:** Autosomal inbreeding estimates (F) for the individuals from the NutriTech cohort (n=73). Each blue circle represents an individual. The green line represents the upper F inclusion threshold. The purple line represents the lower F inclusion threshold. The red circle represents the sample with a high inbreeding estimate.

### c) Adjustment 3: Multidimensional Scaling

Multidimensional scaling was conducted for 73 individuals from the NutriTech cohort according to the methods described in detail in **Chapter 3.7 Population structure adjustments.** The MDS was calculated based on 171,523 common SNPs (MAF>0.01) that were shared between the 9 HapMap Phase 3 population data. **Figure 4- 17** shows the first dimension against the second dimension of the MDS for the NutriTech and HapMap Phase 3 populations. There is a clear distinction between the three major continents, with all NutriTech individuals (**Figure 4- 17**; black triangles) clustered together in close vicinity to the other European populations (TSI and CEU). In summary, there was no ethnic outlier amongst the NutriTech individuals.

### d) Sample exclusions

Only 1 individual from the NutriTech cohort failed at least one of the population structure adjustments (1.37%, male). Hence, that individual was excluded from any downstream analyses.

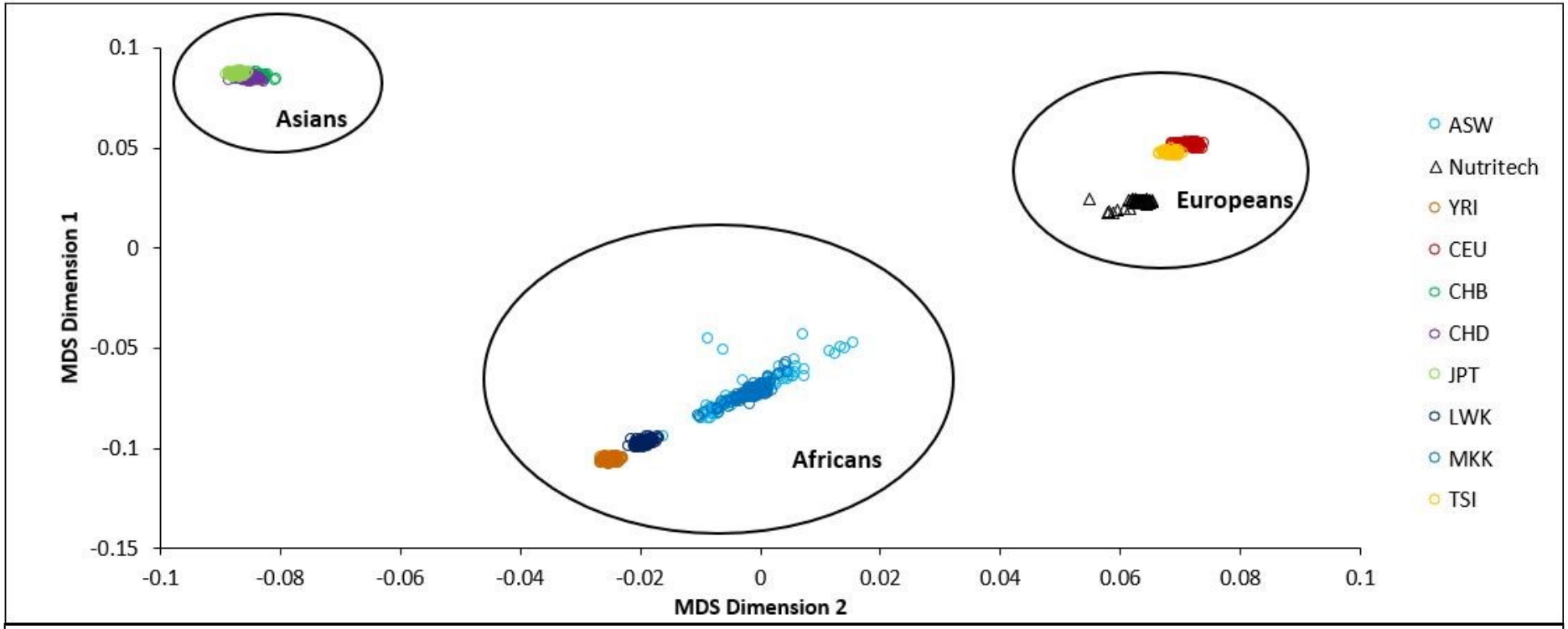


**Figure 4- 17:** Results of multidimensional scaling (MDS) of the NutriTech cohort (Black triangles; n=73) anchored by 9 major HapMap Phase 3 populations. A clear distinction of populations separated according to 3 major continents. *[ASW: African ancestry in Southwest USA; YRI: Yoruba in Ibadan, Nigeria; CEU: Utah residents European ancestry from CEPH; CHB: Han Chinese in Beijing, China; CHD: Chinese in Denver, Colorado USA; JPT: Japanese in Tokyo, Japan; LWK: Luhya in Webuye, Kenya; MKK: Maasai in Kinyawa, Kenya; TSI: Toscani in Italy].*

### 4.5.5. PMMO cohort

A total of 95 individuals from the PMMO cohort survived the QCs and were subjected to population structure adjustments.

**a) Adjustment 1: Inbreeding**

The upper and lower F thresholds (mean ± 4SD) were selected according to the methods in **Chapter 3.7 Population structure adjustments**. The F for each individual was calculated from 145,782 autosomal SNPs. **Figure 4- 18** shows the F for each individual from the PMMO cohort (n=95). In summary, no one had a high inbreeding estimate.

**b) Adjustment 2: Pair-wise IBD Estimation**

Pair-wise IBDs were estimated according to the methods described in **Chapter 3.7 Population structure adjustments**. A pair of individuals was identified to be genetically related to one another (PI_HAT = 0.51). Further investigations on the identity of the related individuals showed that they are in fact a mother (age during surgery = 57.8) and a daughter (age not known, no surgery) of the same family. Since PMMO is a cohort for bariatric surgery patients with an average age during surgery of 50.4, the daughter was selected to be removed from further analyses.

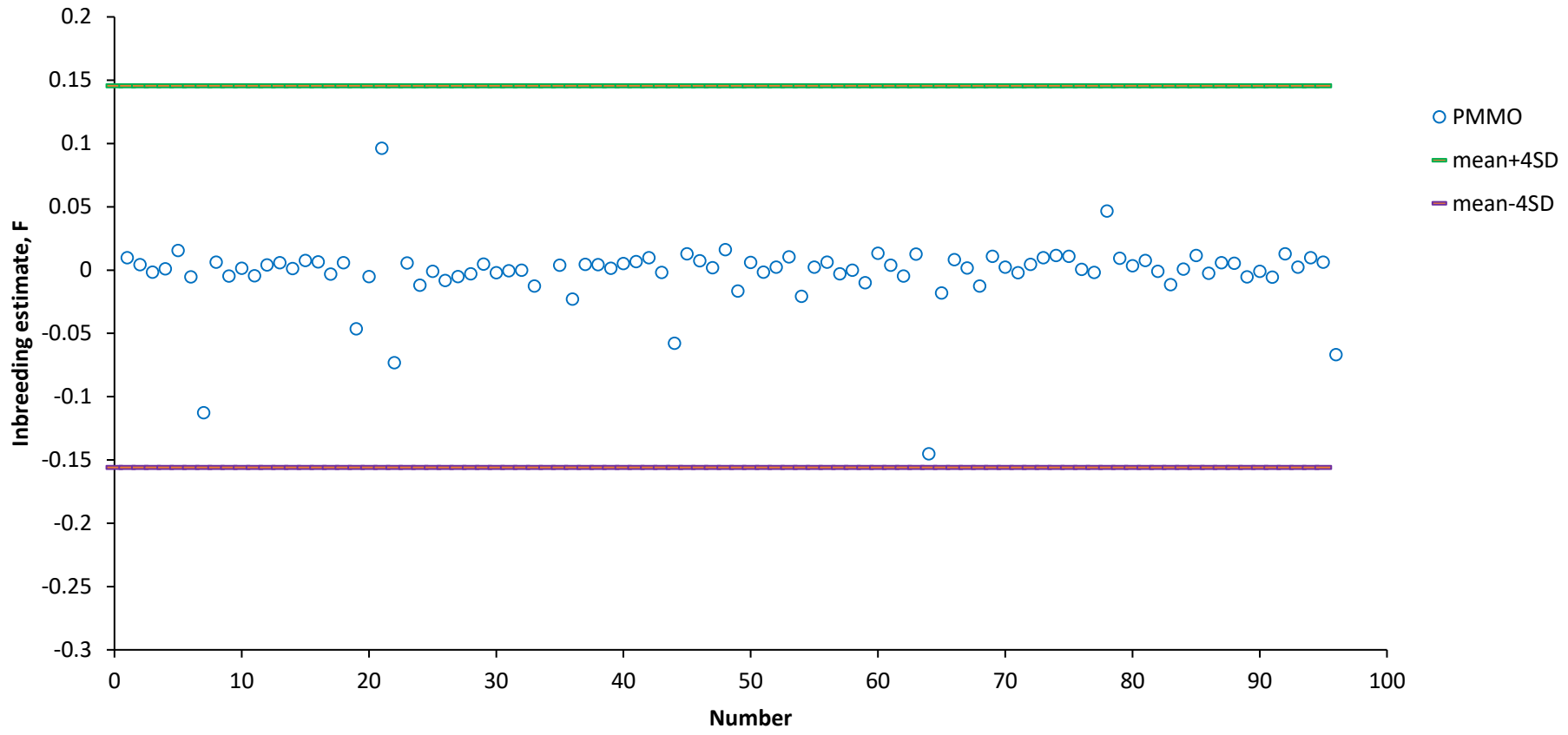


**Figure 4- 18:** Autosomal inbreeding estimates (F) for 95 individuals from the PMMO cohort. Each blue circle represents an individual. The green line represents the upper F inclusion threshold. The purple line represents the lower F inclusion threshold.

**c) Adjustment 3: Multidimensional scaling**

Multidimensional scaling of the PMMO cohort (n=95) was conducted accordingly (refer to **Chapter 3.7 Population structure adjustments**) to identify any population outliers in the PMMO cohort. **Figure 4- 19** shows the results of MDS from the first two principal components for each individual in the PMMO cohort (after QC). The scatter plot was based on 171,077 SNPs (previously pruned for LD) and anchored by 9 HapMap Phase 3 population data. As shown in the **Figure 4- 19**, only one PMMO individual (n=1) was identified to be an outlier as compared to the rest of the cohort (and to the other two European populations from the HapMap Phase 3 population data).

**d) Sample exclusions**

A total of 2 individuals from the PMMO cohort (2.1%; male = 0, female = 2) failed to survive any population structure adjustments. Both individuals were excluded from further analyses.

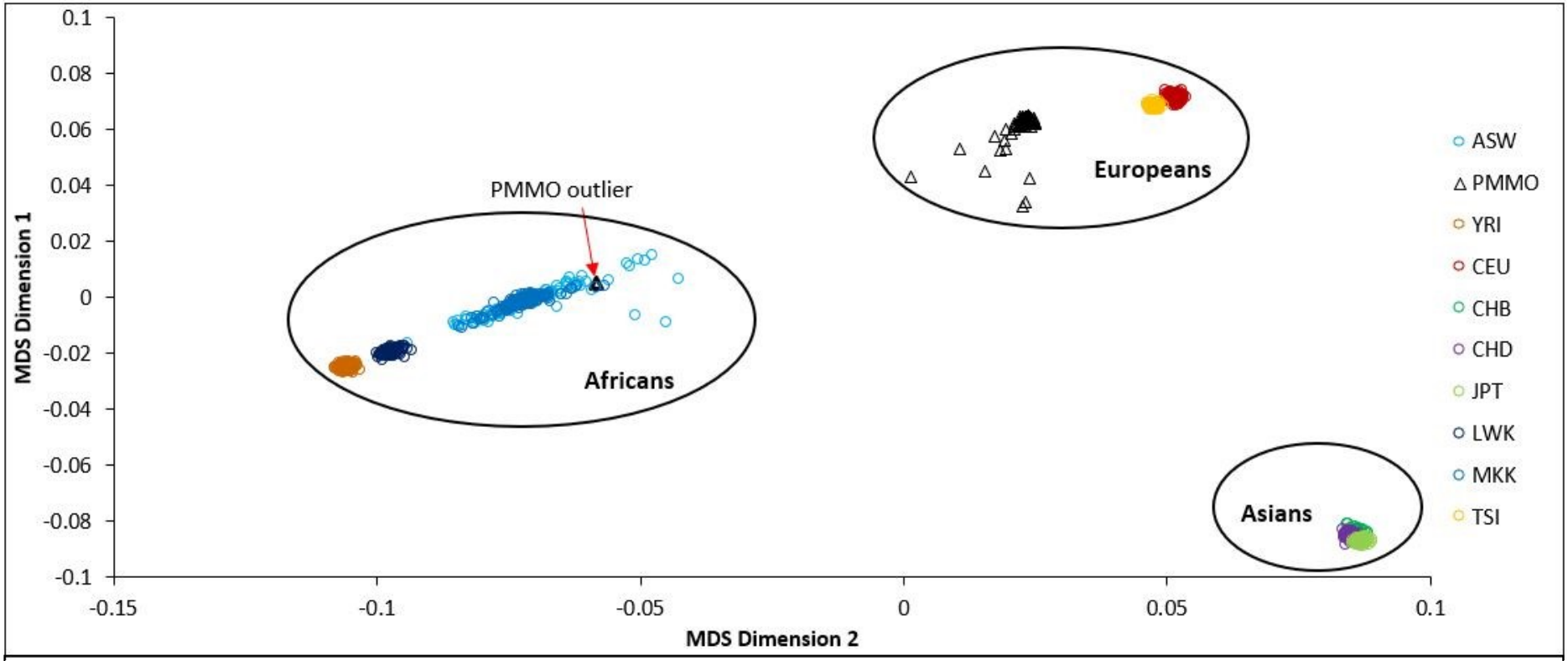


**Figure 4- 19:** Results of multidimensional scaling (MDS) of the PMMO cohort (Black triangles; n=95) anchored by 9 major HapMap Phase 3 populations. A clear distinction of populations separated according to 3 major continents. One PMMO individual was identified as an outlier (shown by the red arrow). *[ASW: African ancestry in Southwest USA; YRI: Yoruba in Ibadan, Nigeria; CEU: Utah residents European ancestry from CEPH; CHB: Han Chinese in Beijing, China; CHD: Chinese in Denver, Colorado USA; JPT: Japanese in Tokyo, Japan; LWK: Luhya in Webuye, Kenya; MKK: Maasai in Kinyawa, Kenya; TSI: Toscani in Italy].*

### 4.5.6. Merged PMMO cohort

After excluding individuals to establish a homogeneous population structure in the second batch (Batch 2), I then had to merge it with the first batch (Batch 1) of the PMMO cohort that was genotyped together with the NutriTech cohort (n=21) (refer to **Chapter 3.5.1 Arrays used in different cohorts**). Since both batches were genotyped on different arrays (with different SNP contents), the genetic data of the merged PMMO cohort now contained a total of 965,042 SNPs (709,810 shared GWAS SNPs, 6,693 additional GWAS SNPs and 248,539 additional exonic SNPs). A series of sample quality controls were conducted on the merged PMMO cohort, as described in **Chapter 3.6 Quality controls**. All individuals in the merged PMMO cohort (n=114; 37.72% male) survived the QCs and were subjected to population structure adjustments, as described in **Chapter 3.7 Population structure adjustments**.

**a) Adjustment 1: Inbreeding**

A total of 114 individuals from the merged PMMO cohort had their Fs calculated based on 146,486 autosomal SNPs that were pruned for LD. **Figure 4- 20** shows that 2 individuals had an F outside of the mean F ± 4SD thresholds (high inbreeding estimates).

**b) Adjustment 2: Pair-wise IBD Estimation**

A total of 117,966 common SNPs were used to estimate the PI_HAT for all individuals in the merged PMMO cohort (n=114). In summary, all of them were found to be genetically unrelated to each other (PI_HAT < 0.2).

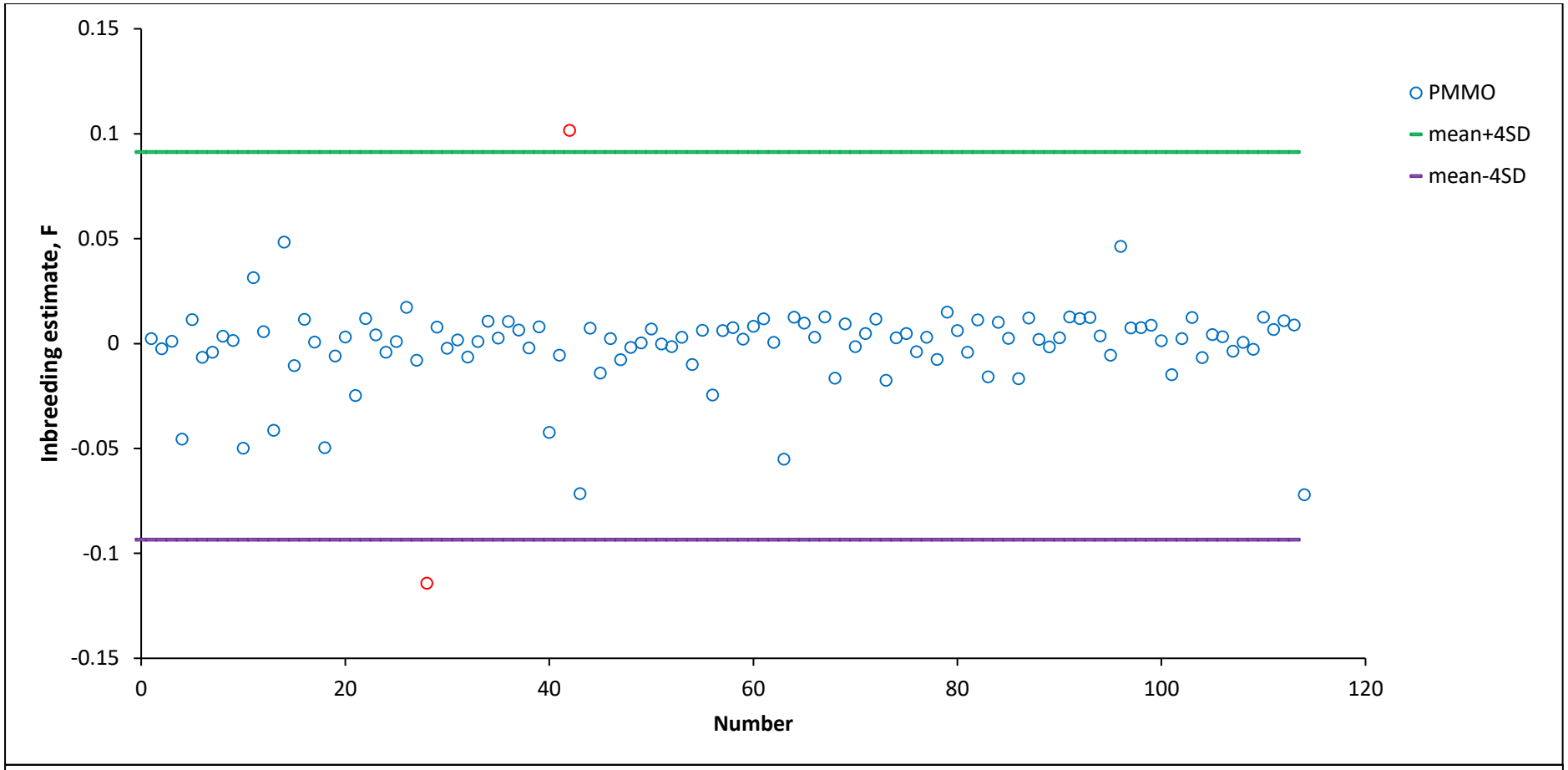


**Figure 4- 20:** Autosomal inbreeding estimates (F) for the individuals from the merged PMMO cohort (n=114). Each blue circle represents an individual. The green line represents the upper F inclusion threshold. The purple line represents the lower F inclusion threshold. The red circles represent the samples with high inbreeding estimates.

### c) Adjustment 3: Multidimensional Scaling

**Figure 4- 21** shows a scatter plot of the first two dimensions of the MDS for the merged PMMO cohort, anchored by 9 HapMap Phase 3 populations. The MDS was conducted based on 172,969 common SNPs (MAF>1%) that were shared between the merged PMMO cohort and the other 9 HapMap Phase 3 populations. From **Figure 4- 21**, all individuals from the merged PMMO cohort are clustered together (black triangles). The PMMO cluster is positioned closely to the two European cohorts from the HapMap Phase 3 (CEU and TSI). Thus, no outliers were identified from the MDS plot.

### d) Sample exclusions

A total of 2 individuals from the merged PMMO cohort were found to be inbred samples (1.75%; male = 0, female = 2). Therefore, both individuals were excluded from further analyses to produce a homogenous PMMO cohort (n=112; 965,042 SNPs, 38.4% male).

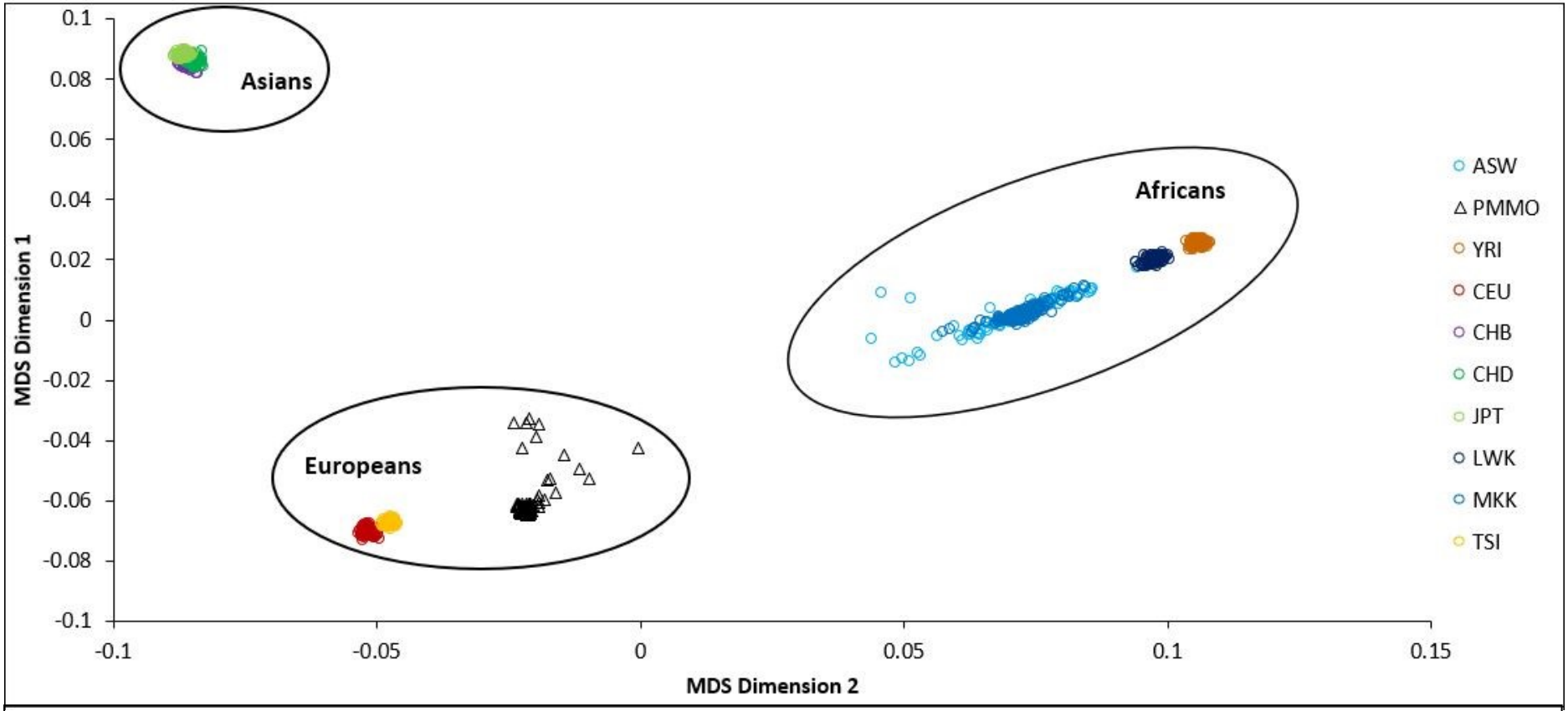


**Figure 4- 21:** Results of multidimensional scaling (MDS) of the merged PMMO cohort (Black triangles; n=114) anchored by 9 major HapMap Phase 3 populations. A clear distinction of populations separated according to 3 major continents. *[ASW: African ancestry in Southwest USA; YRI: Yoruba in Ibadan, Nigeria; CEU: Utah residents European ancestry from CEPH; CHB: Han Chinese in Beijing, China; CHD: Chinese in Denver, Colorado USA; JPT: Japanese in Tokyo, Japan; LWK: Luhya in Webuye, Kenya; MKK: Maasai in Kinyawa, Kenya; TSI: Toscani in Italy].*

### 4.5.7. UK Biobank cohort

The process of correcting the population admixture in the UKBB cohort was conducted by my colleague, Andrianos Yiorkas (Department of Medicine, Imperial College London, UK), as mentioned previously. Initially, a total of 408,247 participants (~83.6%) survived the genotyping QCs, as described earlier in **Chapter 4.4.5 UK Biobank cohort**.

Briefly, a total of 78,828 participants (~19.3% of the surviving participants) were identified as being genetically related to at least one of the remaining participants; they either had excessive number of relatives (>10 relatives) or were inferred as genetically related at up to 3$^{rd}$ degree level. Thus, they were removed from any statistical analyses to avoid confounding biases. In conclusion, a total of 329,419 participants (~67.4% of the total genotyped participants) survived both the genotyping QCs and population structure adjustments.

# 4.6. Discussions

## 4.6.1. Results overview

The genotyping QCs that I performed on both NFBCs, NutriTech and PMMO cohorts, identified <5% failure rates, except for the NFBC1986 cohort (~6.5%). The high percentage of failures in NFBC1986 cohort was likely due to several possibilities in genotyping the second batch, which contributed up to 87% of the failure count. This include possible sample-handling errors, pipetting errors, contaminations, or they were genotyped in a sub-optimal condition. In addition, 18% of the failure count consisted of parents of genotyped individuals and duplicates from the second batch of NFBC1986 cohort, of which there was none from the first batch. On the other hand, the adjustments for the population structure have identified <5% of the failures across all 4 study cohorts. To my knowledge, there has been no gold standard suggested previously regarding the maximum rate of failures that is expected from a SNP genotyping QC protocol. Nonetheless, I ran the QCs and the population structure adjustments again after removing the failed samples, during which no additional failed samples were detected across all study cohorts. It is a good practice to check them again to ensure that all samples with low genotyping quality have been removed from the dataset. Therefore, I am confident that the datasets are clean and homogeneous.

As for the UKBB cohort, the initial dataset available for QCs had been checked internally by Affymetrix prior to releasing the dataset for public use. The Affymetrix team excluded 0.97% of the genotype calls as part of their internal QC checks (Clare Bycroft *et al.*, 2017). The QC pipeline implemented by Andrianos Yiorkas identified ~0.2% of the genotyped participants as poor-quality samples. Overall, 16.4% of genotyped samples were excluded during the QC

pipeline process. However, the high percentage of failures was heavily contributed by the exclusion of non-white British participants (~98.2% of the samples removed were due to this cause). The rest of the QC criteria only resulted in <0.2% failure rate. By comparing with the other 4 study cohorts in this thesis, as well as considering the massive samples size of the UKBB cohort, the low failure rate indicates that the genotyped dataset of the UKBB cohort is of excellent quality.

### 4.6.2. The importance of doing quality controls on SNP genotyping data

A mismatch between the sex determined by genetic information and by clinical information could indicate various issues; plating error, DNA sample contamination, poor DNA quality or clinical data misplacement. Furthermore, it is important to check if any individual had a portion of their SNPs not genotyped since a large portion of SNPs missing could indicate low DNA quality/concentration, and overall genotyping quality. Additionally, heterozygosity rates on autosomal chromosomes have to be looked into since excessive autosomal heterozygous genotypes could indicate DNA contamination, and its reduction could suggest inbreeding. Interestingly, approximately 0.4% (NFBC1966), 0.5% (NFBC1986), 1.4% (NutriTech) and <1% (UKBB) of the individuals had sex discordance issues, in addition to the ~0.1% of the UKBB participants with sex chromosome aneuploidies (Clare Bycroft *et al.*, 2017). This could also mean that these sex-discordant individuals may have sex reversal conditions. In mammals, specifically humans, sex chromosome complements such as XX (female) and XY (male) determine the sex of an individual on genetic and molecular levels, while physiological, behavioural and anatomical presentations determine the sex on a phenotypic level (Sutton *et al.*, 2011). It has been known that chromosome Y holds the crucial signal in determining sex in humans, at which the sex-determining region Y gene (*SRY*; OMIM

*480000) is located (Sutton *et al.*, 2011). The *SRY* gene is thought to be the main component that initiates the activation of testis development and differentiation while suppressing the pathways specific for ovary development, via the upregulated expression of Sry-related HMG box–containing gene 9 (*SOX9*; OMIM* 608160) (Sutton *et al.*, 2011, Temel *et al.*, 2007). Despite that, there are cases whereby sex reversals have occurred, in which they are collectively called as disorders of sexual developments (DSDs) (Sutton *et al.*, 2011). In humans, DSDs are most commonly attributed to sex discordance with a frequency of 1% of live-births (Blackless *et al.*, 2000). There is little evidence to suggest any direct causes of most DSDs, but translocations of the *SRY* gene have been consistently associated with complete male-to-female sex reversal in mice and humans alike (Gubbay *et al.*, 1990, Sutton *et al.*, 2011, Temel *et al.*, 2007). Further studies on *Sry* knockout mice have demonstrated the crucial role of *SRY* gene in the development of testis, in which the *Sry*'s loss-of-function was sufficient to cause the sex reversal condition (Koopman *et al.*, 1991, Lovell-Badge and Robertson, 1990). Other than that, there are genomic syndromes caused by complete or partial genomic rearrangements of chromosome X: chromosome X monosomy causes Turner's syndrome (TS), while Klinefelter syndrome (KS) is characterised by the complete/partial chromosome X trisomy (Tyler and Edman, 2004). The frequency of TS among live female births is approximately 0.05%, in which the common clinical features include short stature and gonadal dysgenesis (Tyler and Edman, 2004). For KS, the prevalence is estimated at 0.16% among live male births (Bojesen *et al.*, 2003). The phenotypic expression of KS patients varied from developmental impairments to infertility (Tyler and Edman, 2004). By using the frequency of DSDs in live births as a reference and considering the prevalence rates for TS and KS, I would expect that ≤1% in the general population to truly have sex discordance issues. Among the datasets I have cleaned for this chapter, only the NutriTech cohort had a sex

discordance rate (~1.4%) more than that is expected in the general population. Additionally, the occurrence sex discordance is made complicated by the ethical and legal issues experienced by transgender individuals. In my understanding, if a person gets a gender recognition certificate in the UK, they might not be legally referred to the "gender assigned at birth" in any medical records. This could complicate the genetic data since a biologically-determined female may appear as a male in the clinical data. Therefore, it is essential that individuals with sex discordance and sex-linked aneuploidies be removed from any population-level genetic studies to avoid confounding.

To check for genotyping accuracy (and DNA quality), each individual has their own call rate; a proportion of SNPs being successfully called (assigned for their genotypes) per individual. The threshold by which researchers use tend to be different from one another, as it depends on the overall genotyping quality of the study. However, it is still possible to identify individuals with low call rates. Provided that the overall genotyping quality is good, these individuals need to be excluded from further analyses since low call rates could indicate poor DNA quality or DNA contamination.

### 4.6.3. Resolving admixture for population homogeneity

All individuals across all study cohorts who have passed quality controls, were subjected to another set of check points to adjust for population structure. As described in the **Chapter 3**, each population has its own genetic structure containing a pool of alleles with specific frequencies, different substructures and ethnic admixtures. These differences within a population could lead to false-positive discoveries since the differences in ancestry could confound the actual genetic differences. Inbreeding can significantly affect the overall fitness of a species in a population due to increased homozygous genotypes within the gene pool of

that population. Together with cryptic relatedness, inbreeding can cause population admixtures and creates a population with high LD levels, which will disrupt the distribution of runs of homozygosity (ROH) within the population (Lemes *et al.*, 2018). The level of inbreeding is therefore informative regarding the admixture events that took place in a population's history, and it can be related to the prevalence of diseases within a population (Lemes *et al.*, 2018). Thus, it is important to correct the population structure in each study cohort by looking at inbreeding, cryptic relatedness and ethnicity clustering. Population structure adjustments could provide us with a more homogeneous population data to represent each study cohort.

# CHAPTER 5. CNV Calling

## 5.1. Introduction

This chapter describes the results of CNV calling by using the PennCNV tool (Wang *et al.*, 2007). This study was carried out to generate a well-defined, high-quality CNV calls that will eventually be compiled as a catalogue of CNVs specific for the NFBC1966, NFBC1986, NutriTech and the PMMO cohorts. The raw genotyping data for these cohorts were readily available and data QC had been carried out in a uniform, systematic way as described in the previous chapter. Furthermore, the types of genotyping platforms (Illumina) that datasets were generated on were consistent across these study cohorts; this is important for the comparison analysis between the CNV catalogues later in this study. The cleaned, high-quality CNV data were subsequently used in the following chapters, mainly to be used 1) in interrogating any interested genomic regions previously known to be associated with obesity-related phenotypes, 2) as proxy-CNVs for permutation tests in identifying any cohort-specific genomic regions that are highly likely to be associated with obesity-related traits, and 3) as a reliable CNV catalogue specific for each study cohort.

Genotyping technology based on single nucleotide polymorphisms (SNPs) was initially designed to genotype common SNPs across the whole genome in a high-throughput manner. It has been successful in providing a robust genotyping system, able to simultaneously scan thousands of individuals efficiently. Recently, the use of SNP genotyping technology has been expanded to detect CNVs as it provides higher throughputs with lower costs compared to other approaches (i.e. comparative genomic hybridisation (CGH), DNA sequencing depth analysis).

Generally, a SNP genotyping technology uses probes labelled with fluorescent molecules that are excited to produce specific allelic signals during the probe hybridisation process. These allelic signals in combination with their intensity and differences between the two alleles of a given genotype are captured by an array scanner. The signal intensity data can be exploited in inferring the number of copy number of the probe sequence identifying a SNP for a given position. For an individual probe, the reliability of this may be questionable, but the presence of consecutive SNPs with similar signal intensity data potentially indicate the presence of an aberrant copy number for the given genomic location.

For the work described this chapter, I used the GenomeStudio Version 2011.1 software (Illumina Inc., San Diego, California, USA) to obtain the necessary signal data from the SNP genotyping, since the NFBC1966, NFBC1986, NutriTech and the PMMO cohorts were genotyped on Illumina genotyping arrays.

I then used the PennCNV tool (Version May 2014) (Wang *et al.*, 2007) and basic linux commands for the CNV calling processes. Briefly, the raw image files from the SNP genotyping were collected and compiled for each study cohort. The GenomeStudio software was fed with the raw image files to obtain the SNP calls, as outlined previously (refer to **Chapter 3.5.2 Processing of genotyped data**). The signal data were also collected and processed by the software to obtain two specific types of signal information; a normalised total signal intensity data (as a ratio of logged intensity value, Log R Ratio) and allelic intensity ratio at each SNP marker (as a frequency value, B Allele Frequency). Both types of signal information were transformed into a large data table consisting of the genotype, Log R Ratio (LRR) and B Allele Frequency (BAF) per SNP for all genotyped individuals. This data table (in comma-delimited format, csv) was fed into the PennCNV software as the main input for CNV calling. After CNVs

were called, a series of filtering and quality controls were conducted to obtain only reliable, high-quality CNVs that were cohort-specific.

## 5.2. Aim of this chapter

The main aim of this chapter was to use the signal data (from the SNP genotyping) in the PennCNV tool for CNV calling and generate a catalogue of high-quality CNVs that were specific for each study cohort under study. The aim is the same across all study cohorts. The generation of the high-quality CNV catalogue per cohort was addressed by the following questions:

1) *How many raw CNVs (stratified by CNV types) were inferred by the PennCNV tool for each study cohort?*
2) *What are the cut-offs for sample inclusion for each study cohort, based on the signal parameters given in the PennCNV tool?*
3) *How many individuals did not survive the signal quality checks in each study cohort?*

I answered these questions by visually inspecting empirical plots of signal quality parameters given by the PennCNV tool. I filtered out all individuals who did not pass the cut-offs determined from the plots.

4) *How many inferred CNVs were considered low in terms of quality, confidence, and SNP density?*

I answered this question by calculating the SNP density of each CNV as well as determining a standard cut-off for the internal confidence score per CNV call. I filtered out all CNV calls that did not pass the CNV quality checks.

5) *What are the percentage of CNVs (stratified by CNV types) that were retained from the raw CNV list?*

I answered this question by providing bar charts representing the proportions of CNVs retained in each study cohort under study, before and after applying various signal quality filters.

6) *Are there any significant differences in terms of the number of CNVs (stratified by CNV types) between all study cohorts?*

I answered this question by executing sample t-tests to determine whether a significant difference is present among the proportions of each CNV type between all study cohorts. I also compared the proportions of the CNV calls in each study cohort, with the findings from other studies.

## 5.3. General methods used in this chapter

| Section | Method name (Described in Chapter 3) | Chapter 4 | Chapter 5 | Chapter 6 | Chapter 7 | Chapter 8 |
|---|---|---|---|---|---|---|
| **3.2** | Biological sample collection | ✓ | ✓ | ✓ | ✓ | ✓ |
| **3.3** | Genomic DNA extraction | ✓ | ✓ | ✓ | ✓ | ✓ |
| **3.4** | Obesity-related phenotypic measurements | ✓ | ✓ | ✓ | ✓ | ✓ |
| **3.5** | SNP genotyping | ✓ | ✓ | ✓ | ✓ | ✓ |
| **3.6** | Quality controls | ✓ | ✓ | ✓ | ✓ | ✓ |
| **3.7** | Population structure adjustments | ✓ | ✓ | ✓ | ✓ | ✓ |
| **3.8** | Principal component analysis | | | | | ✓ |
| **3.9** | CNV calling | | ✓ | | ✓ | ✓ |
| **3.10** | CNV quality controls | | ✓ | | ✓ | ✓ |

## 5.4. Specific methods

### 5.4.1. Statistical analyses

For the comparisons between the study cohorts in this study, I carried out independent t-tests (two-tailed) on the proportions of homozygous deletions, hemizygous deletions, 1-copy duplications, and 2-copy duplications (in %), the ratio of deletions-to-duplications, as well as the average number of CNVs per person (n). All statistical analyses were conducted using IBM SPSS Statistics for Windows Version 24.0 (Released 2016; IBM Corp., Armonk, New York. USA). The t-tests were carried out to determine if the proportion of a CNV type in the final CNV catalogue of a study cohort, was significantly different than the proportions of the same CNV type in other study cohorts. P-values $< 0.05$ were considered as statistically significant. I also conducted independent t-tests on the means and medians of CNV lengths for each CNV type to compare the proportions of CNV calls according to lengths. In addition, I also compared the findings from the study cohorts involved in this study, with the other CNV studies that have been previously reported in different populations.

# 5.5. Results

## 5.5.1. NFBC1966 cohort

*Filter 1: QC from SNP genotyping* - Based on 370,404 SNPs acting as probes, the PennCNV tool was used to infer 287,513 segments as CNVs on all autosomal chromosomes (chromosome 1 to 22) in 5,557 individuals. According to **Chapter 4.4 Results: Sample Quality Controls** and **Chapter 4.5 Results: Population structure adjustments**, a total of 265 NFBC1966 individuals (with 7.2% of the total number of CNVs) were removed from the raw CNV catalogue. In addition, 43 samples had no CNVs called by PennCNV algorithm. Therefore, a total of 266,703 CNVs were retained from 5,249 NFBC1966 individuals (**Figure 5- 2**; red bars).

*Filter 2: QC from CNV calling* – All 5,249 individuals were subjected to a series of internal QCs based on a number of signal quality parameters per individual, as described in **Chapter 3.10.2 Filter 2: QC from CNV calling. Figure 5- 1** shows the distribution of each signal quality parameter for 5,249 NFBC1966 individuals. Briefly, the inclusion threshold point for each parameter was determined by visually inspecting each plot to identify a point whereby all the points started to break while it increased exponentially. In this study cohort, the standard inclusion thresholds of 4SDs away from the mean were chosen for LRR_SD, BAF_DRIFT, WF and GCWF parameters, whereas 2SDs and 6SDs away from the mean were selected for #CNV and BAF_SD respectively (refer **Chapter 3.10.2 Filter 2: QC from CNV calling**).

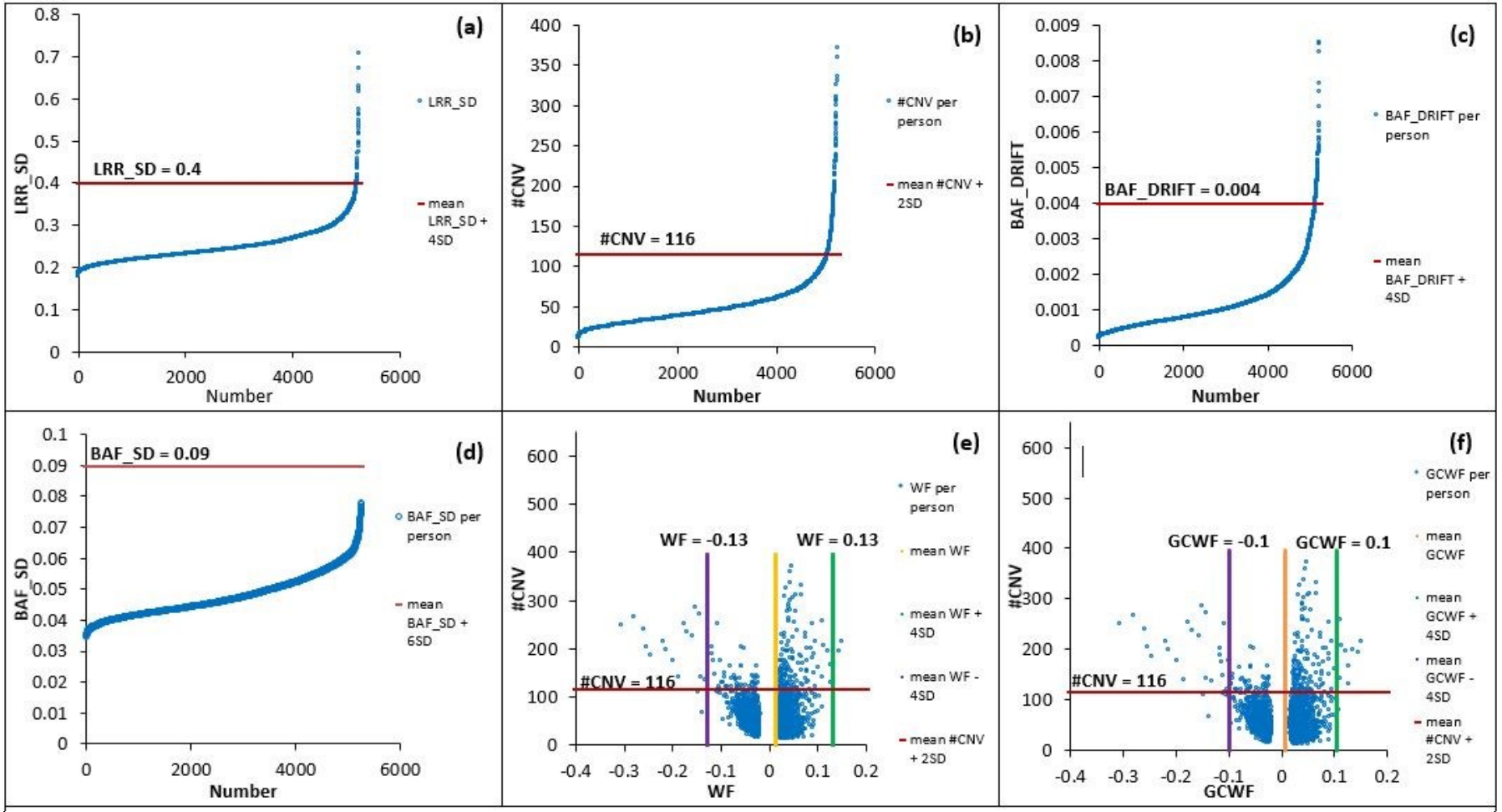


**Figure 5- 1:** Empirical plots for (a) LRR_SD (b) #CNV (c) BAF_DRIFT (d) BAF_SD (e) WF and (f) GCWF parameters, for 5,249 NFBC1966 individuals. LRR_SD = standard deviation of LRR; BAF_SD = standard deviation of BAF; BAF_DRIFT = departure score of observed BAFs from expected BAF; #CNV = Total CNVs; WF = Waviness Factor; GCWF = GC-adjusted WF.

An inclusion cut-off was selected for each parameter based upon visual inspections on the empirical plots in **Figure 5- 1** (**Table 5- 1**). A total of 331 individuals were deemed as outliers from at least one parameter distribution and removed from further analyses. An additional 13 individuals were excluded because they did not give consents for their genetic information to be used in this study. Therefore, a total of 344 individuals were excluded based on the CNV quality control parameters (~6.5%).

| **Parameter** | **Parameter cut-off criterion** | **Cut-off/Inclusion value(s)** | **PennCNV default settings for the inclusion value(s)** |
|---|---|---|---|
| LRR_SD | μ + 4SD | value ≤ 0.4 | value ≤ 0.3 |
| BAF_SD | μ ± 6SD | value ≤ 0.1 | value ≤ 0.1 |
| BAF_DRIFT | μ + 4SD | value ≤ 0.004 | none |
| #CNV | μ + 2SD | value ≤ 116 | value ≤ 100 |
| WF | μ ± 4SD | -0.13 < value < 0.13 | -0.05 < value < 0.05 |
| GCWF | μ ± 4SD | -0.1 < value < 0.1 | none |

**Table 5- 1:** The cut-off criteria for sample inclusion in the NFBC1966 cohort, based on visual inspection of empirical plots for the parameters provided by PennCNV tool. SD = standard deviation; μ = mean value per cohort; LRR_SD = standard deviation of LRRs per individual; BAF_SD = standard deviation of BAFs per individual; BAF_DRIFT = departure score of observed BAFs from expected BAFs per individual; #CNV = Total number of CNVs inferred by PennCNV per individual; WF = Waviness Factor per individual; GCWF = GC-adjusted WF per individual. NFBC166 = Northern Finland Birth Cohort 1966.

*Filter 3-4: Poor quality, spurious CNV calls* – Low confidence CNV calls were defined as CNV calls that had internal confidence scores (conf) < 20, while low-quality CNV calls were defined as CNV calls that had a low SNP density (<3 SNPs per 100 kb). Any CNV calls that were in either or both categories were excluded from the CNV catalogue. In this cohort (n=5,249), a total of 227,114 CNV calls (~85%) were considered as low confidence calls, while 1,517 CNV calls (<1%) were deemed as low-quality calls. In addition, CNV calls located in immunoglobin, centromeric and telomeric regions were flagged as spurious (Karolchik *et al.*, 2004). A total of 2,319 CNV calls (<1%) were identified as spurious.

After applying the filters (Filter 2 to Filter 4; refer to **Chapter 3.10 CNV quality controls**), a total of 344 individuals (6.5% from 5,249 individuals) and 237,030 CNVs (82.4% from raw CNV catalogue) were excluded from any downstream analyses. Any overlapping CNV calls were merged according to the methods described in **Chapter 3.10 CNV quality controls**. In summary, the final CNV catalogue for the NFBC1966 cohort consisted of 25,878 high-confidence and SNP-dense CNVs from 4,905 individuals (**Figure 5- 2**), which accounted for ~9.7% of the raw CNV calls prior to the CNV filtering process.

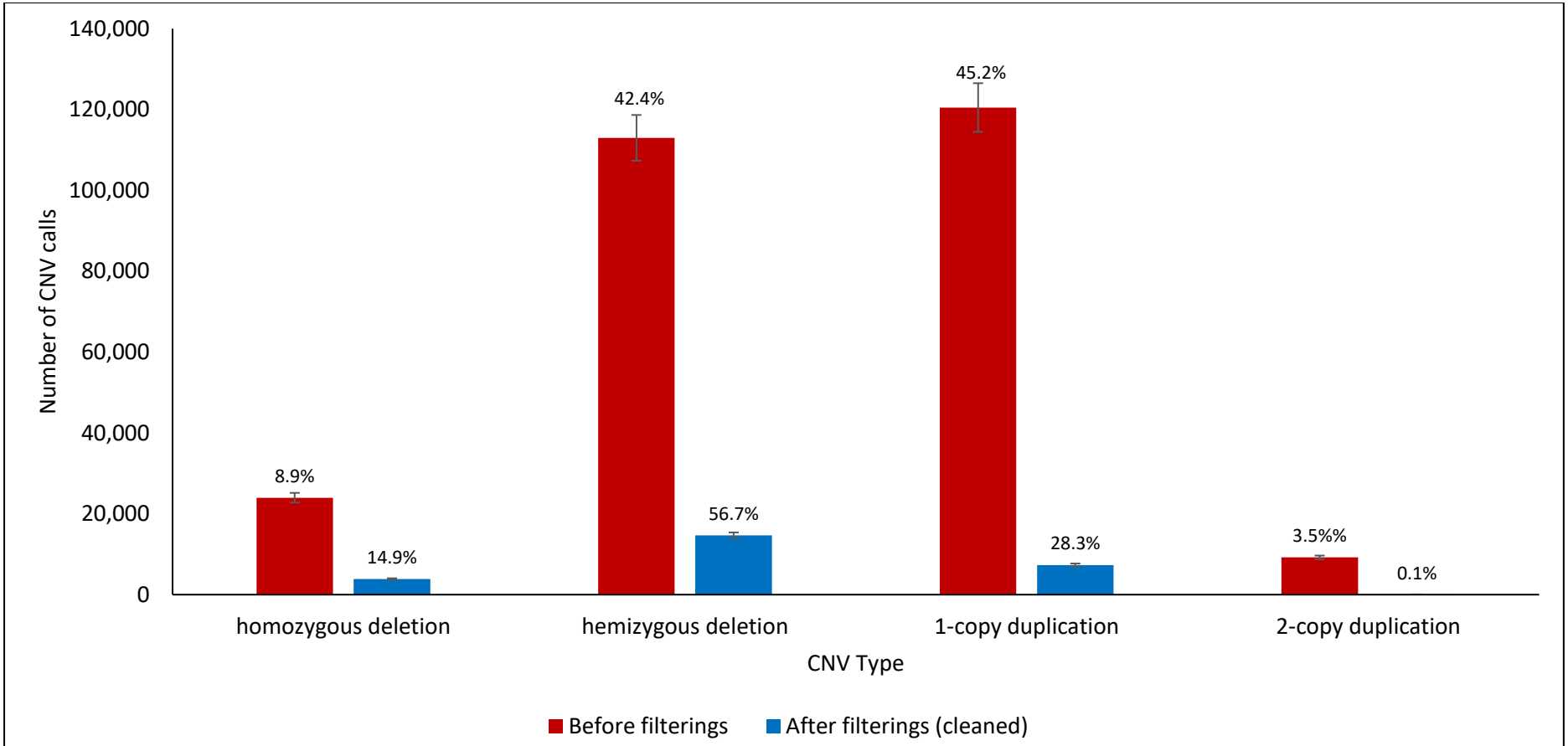


**Figure 5- 2:** The number of CNV types that were called by PennCNV for the NFBC1966 cohort, before (Red) and after (Blue) removing individuals with poor signal quality and low quality CNV calls. The percentages shown were the proportions of CNV types from the total number of CNV calls. CNVs = Copy number variants. Error bars = 5% error

### 5.5.2. NFBC1986 cohort

Since NFBC1986 cohort was genotyped in two batches and they differ in terms of genotyping time, each batch was treated separately before combining. Both batches were subjected to the same filter criteria (but with different cut-offs for each signal parameter) to reduce CNV quality differences between batches.

*Filter 1: QC from SNP genotyping* - Based on 949,326 (NFBC1986A) and 962,312 (NFBC1986B) autosomal SNPs acting as probes (>99% identical probes between batches), PennCNV was used to infer 68,305 (in 2,016 individuals from the NFBC1986A batch) and 206,483 (in 2,027 individuals from the NFBC1986B batch) segments as CNVs on all autosomal chromosomes. It is important to note that there were 3 times more CNV calls in the second batch (NFBC1986B) compared to the ones being called in the first batch (NFBC1986A). Both batches were genotyped on genotyping chips that were highly identical in terms of the probe contents (>~99% identical; personal communication with Illumina's technical support group), which included exonic probes as well. Thus, the reason for the difference observed in the number of CNVs was unknown; it could be the differences in lab personnel, lab conditions, or other uncontrolled variables. Of note, the mean WF (and GCWF) of the NFBC1986B were not close to zero as compared to that of the first batch (NFBC1986A) and the NFBC1966 cohort (**Figure 5- 4**). Thus, I suspected that most of the false positive calls came from a group of samples with high levels of genomic waves in the signal background, which presumably led to false calling of technical artefacts. After removing individuals who did not survive genotyping QCs, a total of 65,612 and 141,404 CNV calls were retained for the NFBC1986A and NFBC1986B batches respectively.

*Filter 2: QC from CNV calling* – Both batches of the NFBC1986 cohort (post-QC) were subjected to the same internal QCs as with the NFBC1966 cohort. **Figure 5- 3** and **Figure 5- 4** show the distributions of each signal quality parameter for 1,966 NFBC1986A and 1,783 NFBC1986B individuals, respectively. Since the total number of CNV calls from the second batch were potentially too high (3 times higher than the first batch (NFBC1986A), with an abundance of 1-copy duplications), I suspected that the overall genotyping quality of the NFBC1986B batch was relatively poor. The results in **Chapter 4.4 Results: Sample Quality Controls** also demonstrated a 7-fold increase in the number of NFBC1986B individuals who failed the sample QCs as compared to the NFBC1986A batch (230 against 33 individuals). Thus, all cut-offs for signal quality parameters in the second batch (NFBC1986B) were devised by placing twice the stringency that was used on the first batch (i.e. 2SDs away from the mean is a more stringent cut-off point than 4SDs away from the mean). **Table 5- 2** shows the inclusion cut-off points that were based on the empirical plots in **Figure 5- 3** and **Figure 5- 4**. A total of 92 NFBC1986A (~4.7%) and 430 NFBC1986B (~24%) individuals failed in at least one parameter criterion and were removed from further analyses.

*Filter 3-4: Poor quality, spurious CNV calls* – For the NFBC1986A batch (n=1,966), a total of 32,286 CNV calls (~49%) were classified as low in calling confidence (conf <20), 85 CNV calls (<1%) were considered as low-quality (SNP density <0.00003), while 1,355 calls (~2%) were deemed to be spurious CNV calls (Karolchik *et al.*, 2004). On the other hand, the NFBC1986B batch (n=1,783) had 112,993 CNV calls (~80%) that were of low confidence, 160 CNV calls (<1%) with low SNP density, and 439 (<1%) spurious CNV calls.

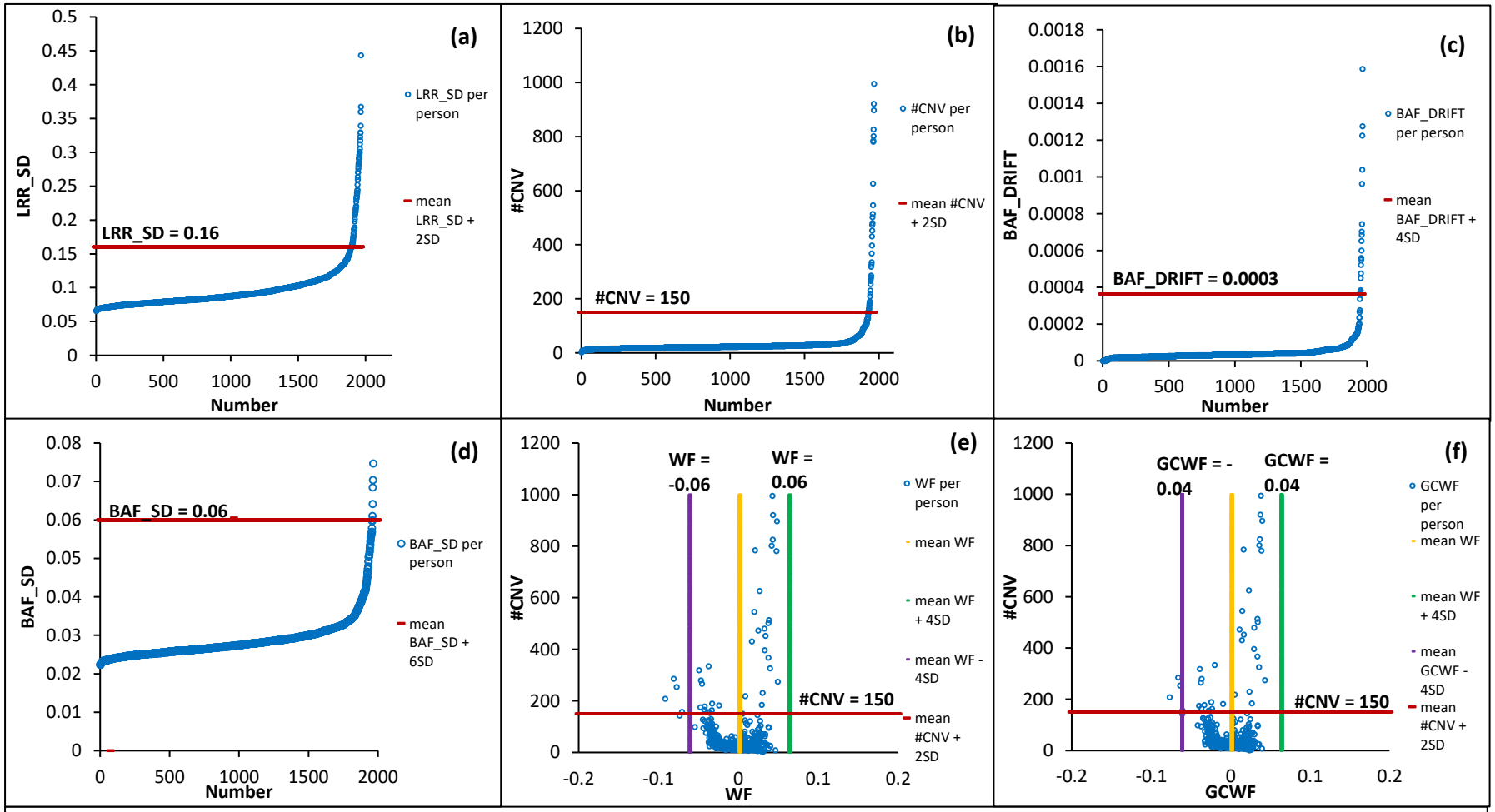


**Figure 5- 3:** Empirical plots for (a) LRR_SD (b) #CNV (c) BAF_DRIFT (d) BAF_SD (e) WF and (f) GCWF parameters, for NFBC1986A individuals. LRR_SD = standard deviation of LRR; BAF_SD = standard deviation of BAF; BAF_DRIFT = departure score of observed BAFs from expected BAF; #CNV = Total CNVs; WF = Waviness Factor; GCWF = GC-adjusted WF.

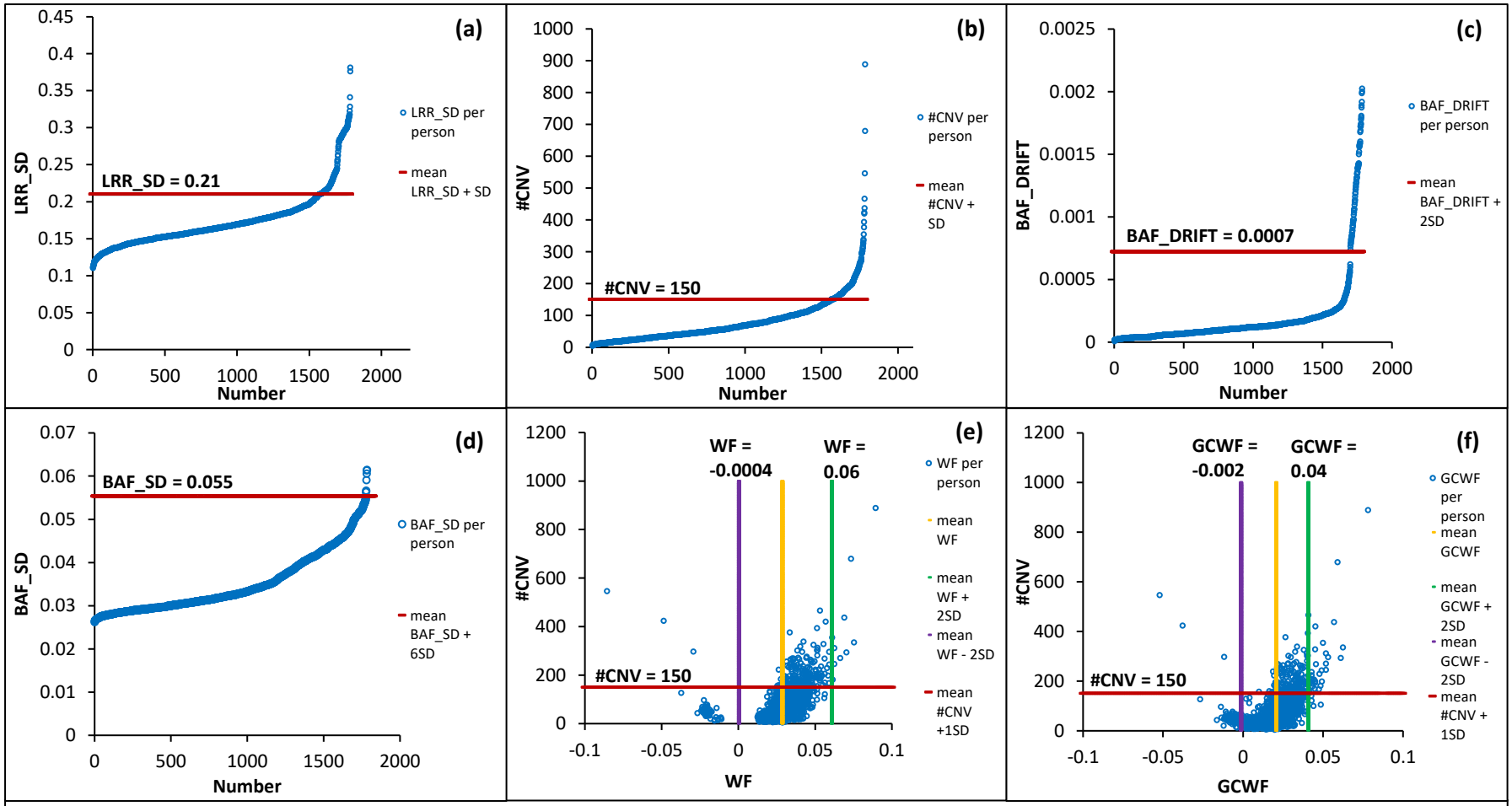


**Figure 5- 4:** Empirical plots for (a) LRR_SD (b) #CNV (c) BAF_DRIFT (d) BAF_SD (e) WF and (f) GCWF parameters, for NFBC1986B individuals. LRR_SD = standard deviation of LRR; BAF_SD = standard deviation of BAF; BAF_DRIFT = departure score of observed BAFs from expected BAF; #CNV = Total CNVs; WF = Waviness Factor; GCWF = GC-adjusted WF.

| Parameter | Parameter cut-off criterion | | Cut-off/Inclusion value(s) | | PennCNV default settings for the inclusion value(s) |
|---|---|---|---|---|---|
| | NFBC1986A | NFBC1986B | NFBC1986A | NFBC1986B | |
| LRR_SD | μ + 2SD | μ + 1SD | value ≤ 0.4 | value ≤ 0.4 | value ≤ 0.3 |
| BAF_SD | μ ± 6SD | μ ± 3SD | value ≤ 0.1 | value ≤ 0.1 | value ≤ 0.1 |
| BAF_DRIFT | μ + 4SD | μ + 2SD | value ≤ 0.004 | value ≤ 0.004 | none |
| #CNV | μ + 2SD | μ + 1SD | value ≤ 116 | value ≤ 116 | value ≤ 100 |
| WF | μ ± 4SD | μ ± 2SD | -0.13 < value < 0.13 | -0.13 < value < 0.13 | -0.05 < value < 0.05 |
| GCWF | μ ± 4SD | μ ± 2SD | -0.1 < value < 0.1 | -0.1 < value < 0.1 | none |

**Table 5- 2:** Cut-off criteria for sample inclusion for the first (NFBC1986A) and second (NFBC1986B) batches of the NFBC1986 cohort, based on visual inspection of empirical plots for the parameters provided by the PennCNV tool. SD = standard deviation; μ = mean value per cohort; LRR_SD = standard deviation of LRRs per individual; BAF_SD = standard deviation of BAFs per individual; BAF_DRIFT = departure score of observed BAFs from expected BAFs per individual; #CNV = Total number of CNVs called by the PennCNV tool per individual; WF = Waviness Factor per individual; GCWF = GC-adjusted WF per individual.

The stringent QC parameter filters were applied on the raw CNV lists from both batches of the NFBC1986 cohort (Filter 2 to Filter 4; refer to **Chapter 3.10 CNV quality controls**). For the first batch (NFBC1986A), the filters identified 92 individuals who failed the QC filters (~4.7%) and poor quality 42,456 CNV calls (~65%). For the second batch (NFBC1986B), 432 individuals (~24%) failed the QC filters, and 125,916 CNV calls (~89%) were of low quality. They were subsequently removed from any downstream analyses. Any overlapping CNV calls were merged according to the methods described in **Chapter 3.10 CNV quality controls**. **Figure 5-5** and **Figure 5- 6** show the percentages of CNV calls predicted by the PennCNV tool (stratified by CNV types), before and after QC filtering on the first and the second batch of the NFBC1986 cohort, respectively. The filtering processes reduced the amount of 1-copy duplications in the second batch by nearly 12-folds (from 119,604 to 10,176 duplication calls), as shown in **Figure 5- 6**.

In summary, I managed to retain 22,360 CNVs (32.7% from raw CNV list) in 1,874 individuals from the first batch, and 15,310 CNVs (7.4% from raw CNV list) in 1,353 individuals from the second batch. The CNV catalogues from both batches were then combined to produce a total of 37,670 CNV calls for 3,227 individuals in the NFBC1986 cohort, as shown in **Figure 5- 7**.

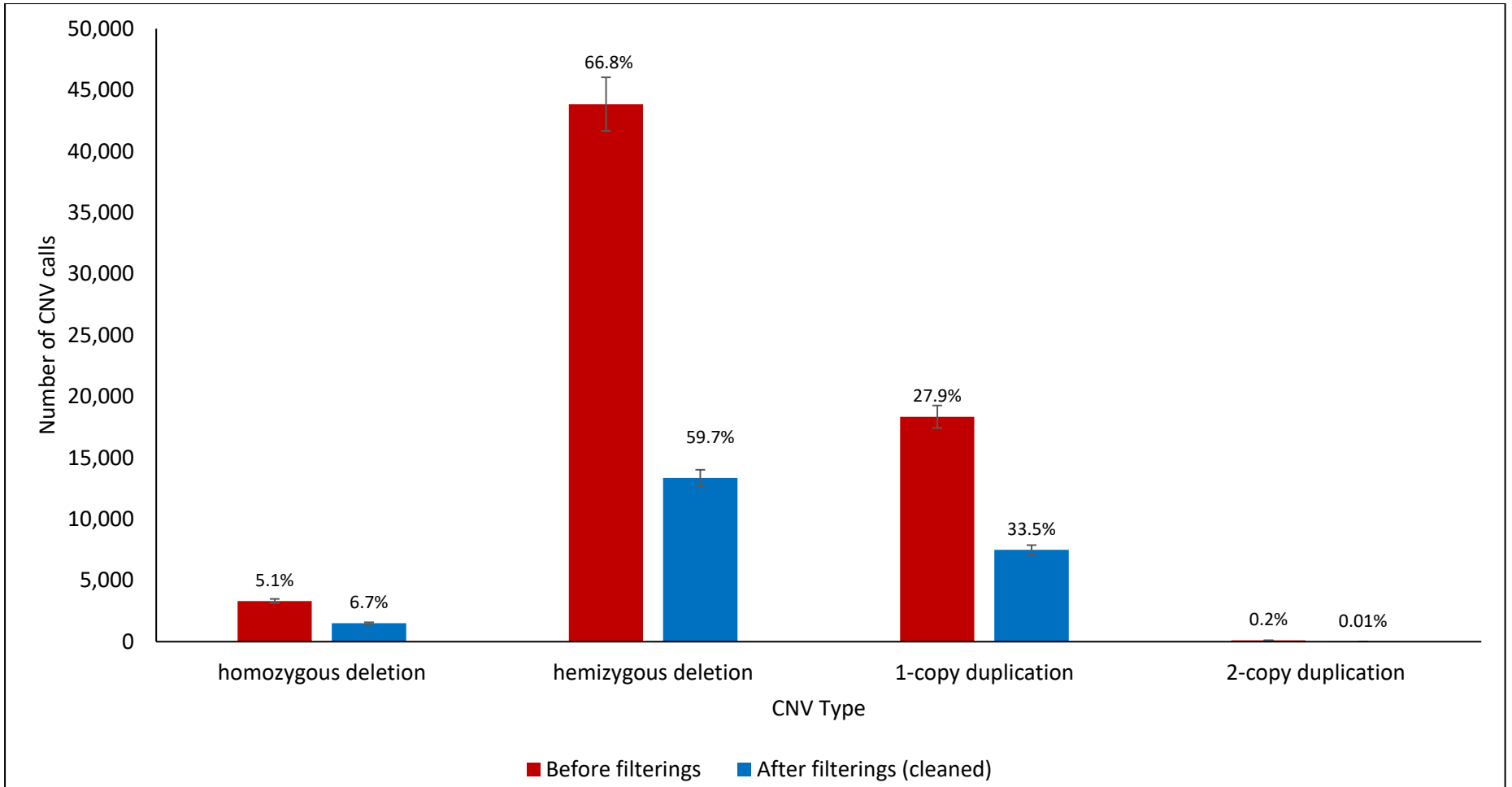


**Figure 5- 5:** The number of CNV types that were called by PennCNV for the first batch of the NFBC1986 cohort (NFBC1986A), before (Red) and after (Blue) removing individuals with poor signal quality and low quality CNV calls. The percentages shown were the proportions of CNV types from the total number of CNV calls. CNVs = Copy number variants. Error bars = 5% error

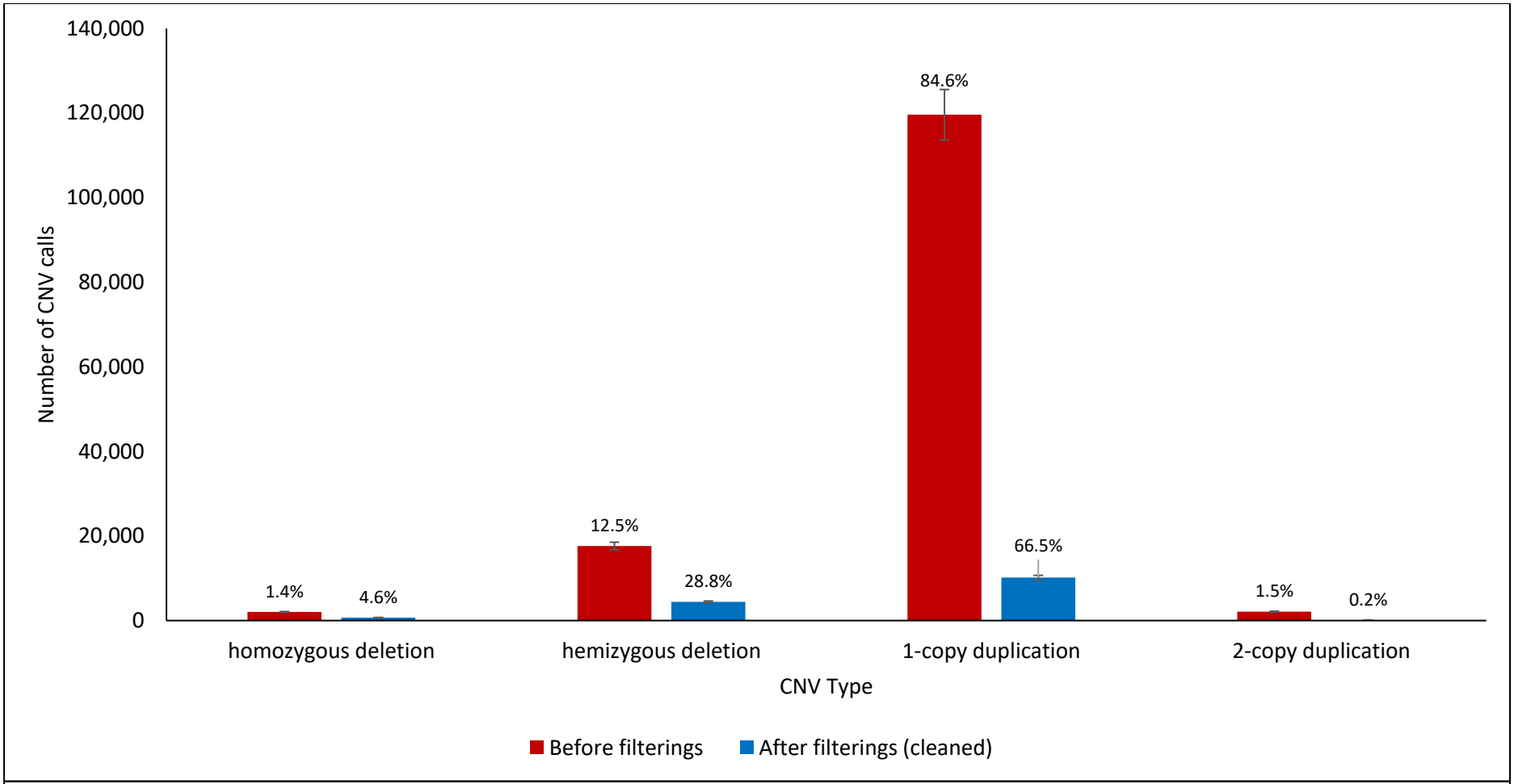


**Figure 5- 6:** The number of CNV types that were called by PennCNV for the second batch of the NFBC1986 cohort (NFBC1986B), before (Red) and after (Blue) removing individuals with poor signal quality and low quality CNV calls. The percentages shown were the proportions of CNV types from the total number of CNV calls. CNVs = Copy number variants. Error bars = 5% error

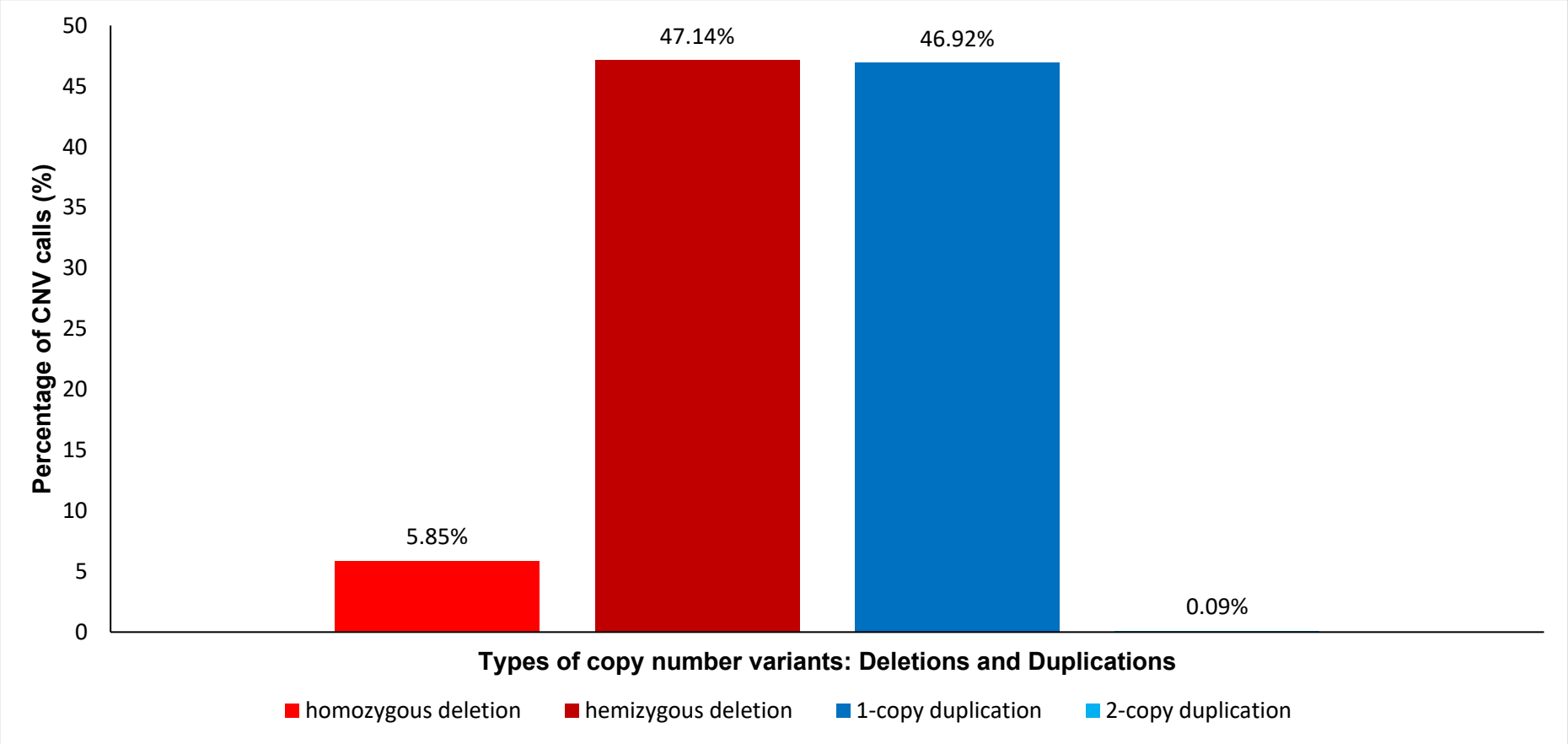


**Figure 5- 7:** The percentages of CNV types (called by PennCNV) in the final CNV catalogue for the merged NFBC1986 cohort. A total of 37,670 CNV calls were retained in 3,227 individuals for the NFBC1986 cohort. These high-quality CNV calls were inferred based on good genotyping signal information. CNVs = Copy number variants.

### 5.5.3. NutriTech cohort

*Filter 1: QC from SNP genotyping* – A total of 714,863 SNPs were used by the PennCNV tool to call for CNVs in the NutriTech cohort (n=74). Prior to removing 2 participants who failed the genotyping QC (refer to **Chapter 4.4 Results: Sample Quality Controls** and **Chapter 4.5 Results: Population structure adjustments)**, 2,239 CNVs were initially predicted. The two participants (accounted for ~12% of the raw CNV calls) were subsequently removed from the cohort, which reduced the amount of calls to 1,967 CNVs in the remaining participants (n=72). The proportions of CNV calls in each CNV type for the remaining participants are shown in **Figure 5- 9**, represented as red bars (before CNV QC filtering).

*Filter 2: QC from CNV calling* – The same signal quality parameter filtering process that was used for the NFBCs, was also applied to the NutriTech participants (n=72) (refer to **Chapter 3.10.2 Filter 2: QC from CNV calling)**. **Figure 5- 8** displays the parameter distributions and inclusion thresholds specific for each parameter, for the NutriTech cohort (n=72). The standard inclusion thresholds of 2SDs, 4SDs and 6SDs away from the means were selected upon visual inspections on the #CNV, LRR_SD and BAF_SD distribution plots respectively. However, 2SDs away from the mean was chosen as the inclusion thresholds for BAF_DRIFT, WF and GCWF parameters, rather than using the standard inclusion thresholds of 4SDs away from the means (**Figure 5- 8**).

**Table 5- 3** shows a summary of the inclusion thresholds (and the cut-off values) for these signal quality parameters. Four NutriTech participants (n=4) failed to survive at least one parameter threshold, while 2 others had missing clinical data. Thus, a total of 6 participants (~8%) had to be excluded from further analyses.

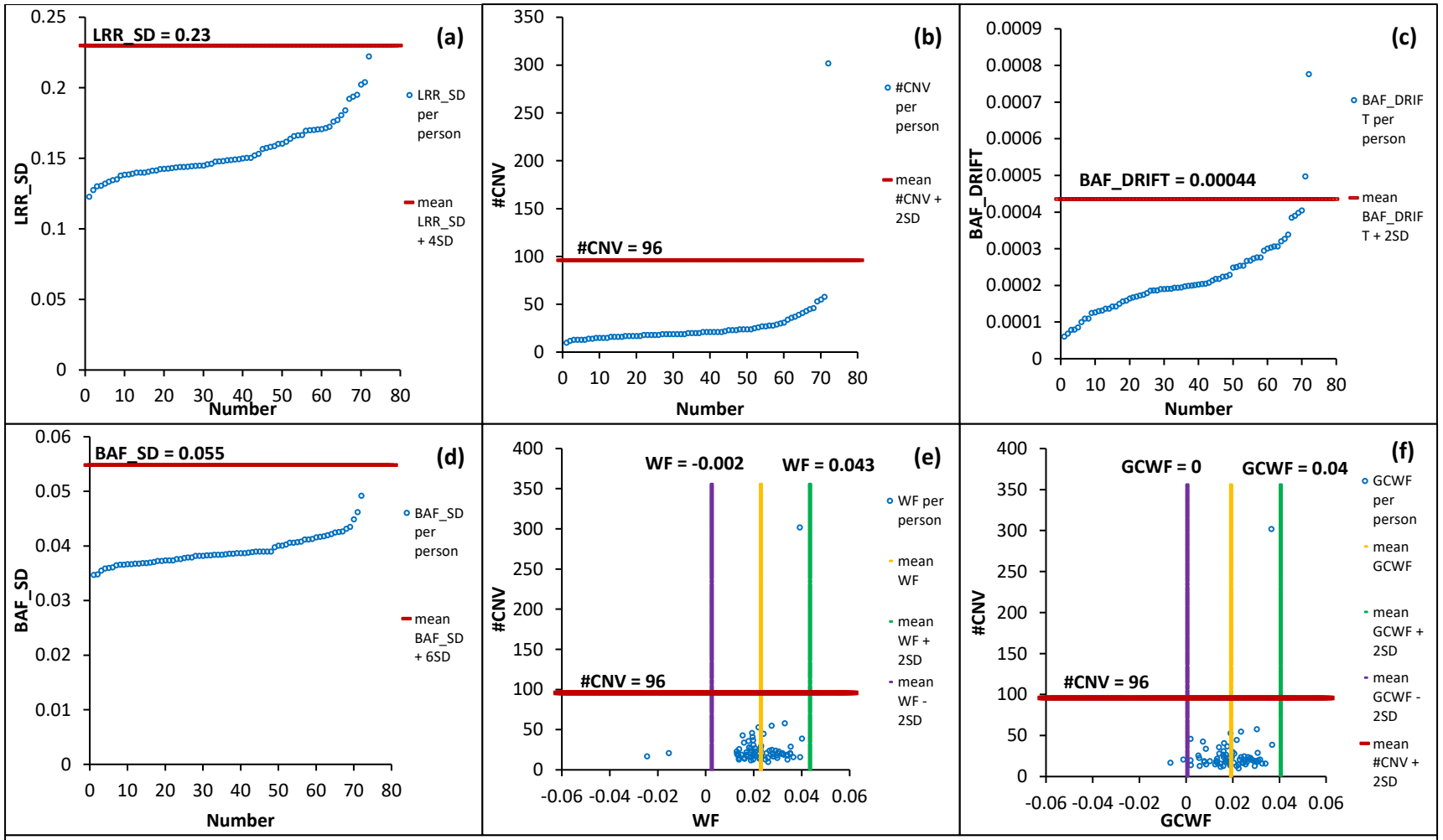


**Figure 5- 8:** Empirical plots for (a) LRR_SD (b) #CNV (c) BAF_DRIFT (d) BAF_SD (e) WF and (f) GCWF parameters, for 72 NutriTech participants. LRR_SD = standard deviation of LRR; BAF_SD = standard deviation of BAF; BAF_DRIFT = departure score of observed BAFs from expected BAF; #CNV = Total CNVs; WF = Waviness Factor; GCWF = GC-adjusted WF.

| Parameter | Parameter cut-off criterion | Cut-off/Inclusion value(s) | PennCNV default settings for the inclusion value(s) |
|---|---|---|---|
| LRR_SD | μ + 4SD | value ≤ 0.23 | value ≤ 0.3 |
| BAF_SD | μ ± 6SD | value ≤ 0.055 | value ≤ 0.1 |
| BAF_DRIFT | μ + 2SD | value ≤ 0.00044 | none |
| #CNV | μ + 2SD | value ≤ 96 | value ≤ 100 |
| WF | μ ± 2SD | -0.043 < value < 0.043 | -0.05 < value < 0.05 |
| GCWF | μ ± 2SD | -0.04 < value < 0.04 | none |

**Table 5- 3:** Cut-off criteria for individual inclusion from the NutriTech cohort, based on visual inspections of parameter distribution plots in **Figure 5- 8**. SD = standard deviation; μ = mean value per cohort; LRR_SD = standard deviation of LRRs per individual; BAF_SD = standard deviation of BAFs per individual; BAF_DRIFT = departure score of observed BAFs from expected BAFs per individual; #CNV = Total number of CNVs predicted by the PennCNV tool per individual; WF = Waviness Factor per individual; GCWF = GC-adjusted WF per individual.

*Filter 3-4: Poor quality, spurious CNV calls* – I have identified 1,249 CNV calls (~63.5%) that were considered as low-confidence based on the conf scores. In addition, 12 CNV calls (<1%) were deemed to be low-quality calls since these CNVs had low SNP density. In order to identify spurious calls, the CNV calls were scanned for any overlaps between the CNVs and immunoglobin, centromeric or telomeric regions (Karolchik *et al.*, 2004). The scan yielded a total of 20 spurious CNV calls (~1%).

After removing CNV calls that were identified from Filter 2, Filter 3 and Filter 4, a total of 6 NutriTech participants (~8% from 72 participants) and 1,450 CNVs (~73.7% from the raw CNV catalogue) were excluded from any downstream analyses. Subsequently, two CNV calls were merged as a single call because ≥20% of their lengths overlapped each other. **Figure 5- 9** shows the amount of CNV calls according to CNV types in the NutriTech cohort, before and after applying all the filters. In summary, the final CNV catalogue for the NutriTech cohort (n=66) contained 516 CNV calls that were high in quality and confidence, which accounted for ~26% of the raw CNV calls prior to any filtering process.

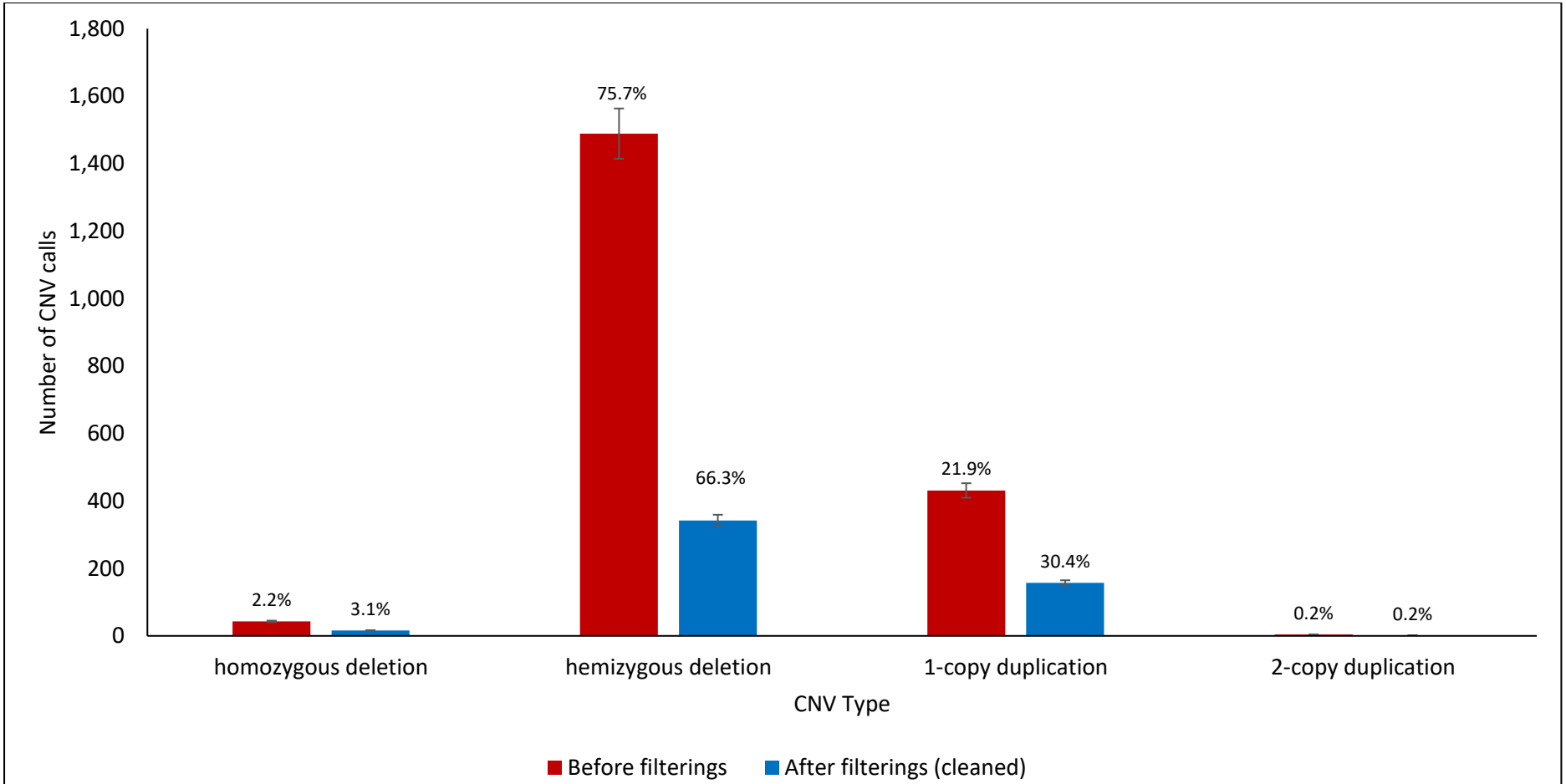


**Figure 5- 9:** The number of CNV types that were called by PennCNV for the NutriTech cohort, before (Red) and after (Blue) removing individuals with poor signal quality and low quality CNV calls. The percentages shown were the proportions of CNV types from the total number of CNV calls. CNVs = Copy number variants. Error bars = 5% error

### 5.5.4. PMMO cohort

*Filter 1: QC from SNP genotyping* – Based on 956,781 SNPs (1,716 SNPs were discarded because they were on mitochondrial and sex chromosomes), the PennCNV tool was used to infer 49,827 CNV calls in the PMMO cohort (n=118). According to **Chapter 4.4 Results: Sample Quality Controls** and **Chapter 4.5 Results: Population structure adjustments,** a total of 6 patients did not survive the genotyping QCs and the population structure adjustments. Those patients (accounting for ~3.4% of the raw CNV calls) were removed from the cohort, with 47,445 calls retained in 112 patients.

*Filter 2: QC from CNV calling* – All signal quality parameters (and thresholds) that have been described in the previous cohort, were also used in cleaning up the CNV list for the PMMO cohort (n=112). **Figure 5- 10** depicts the distribution of each parameter for the PMMO cohort (n=112). The inclusion thresholds and cut-off points for each parameter were summarised in **Table 5- 4**. A total of 6 patients (~5.3%) failed at least one parameter threshold, and hence were excluded from the next filter process. The excluded patients accounted for ~58% of the CNV calls.

*Filter 3-4: Poor quality, spurious CNV calls* – 19,870 CNV calls from the remaining patients (n=106) were checked for their quality and confidence by excluding those with conf scores <20, or those with low SNP density, as described in the previous cohort. In addition, 106 CNV calls were identified to overlap with genomic regions that were previously known to harbour spurious copy number variants (Karolchik *et al.*, 2004).

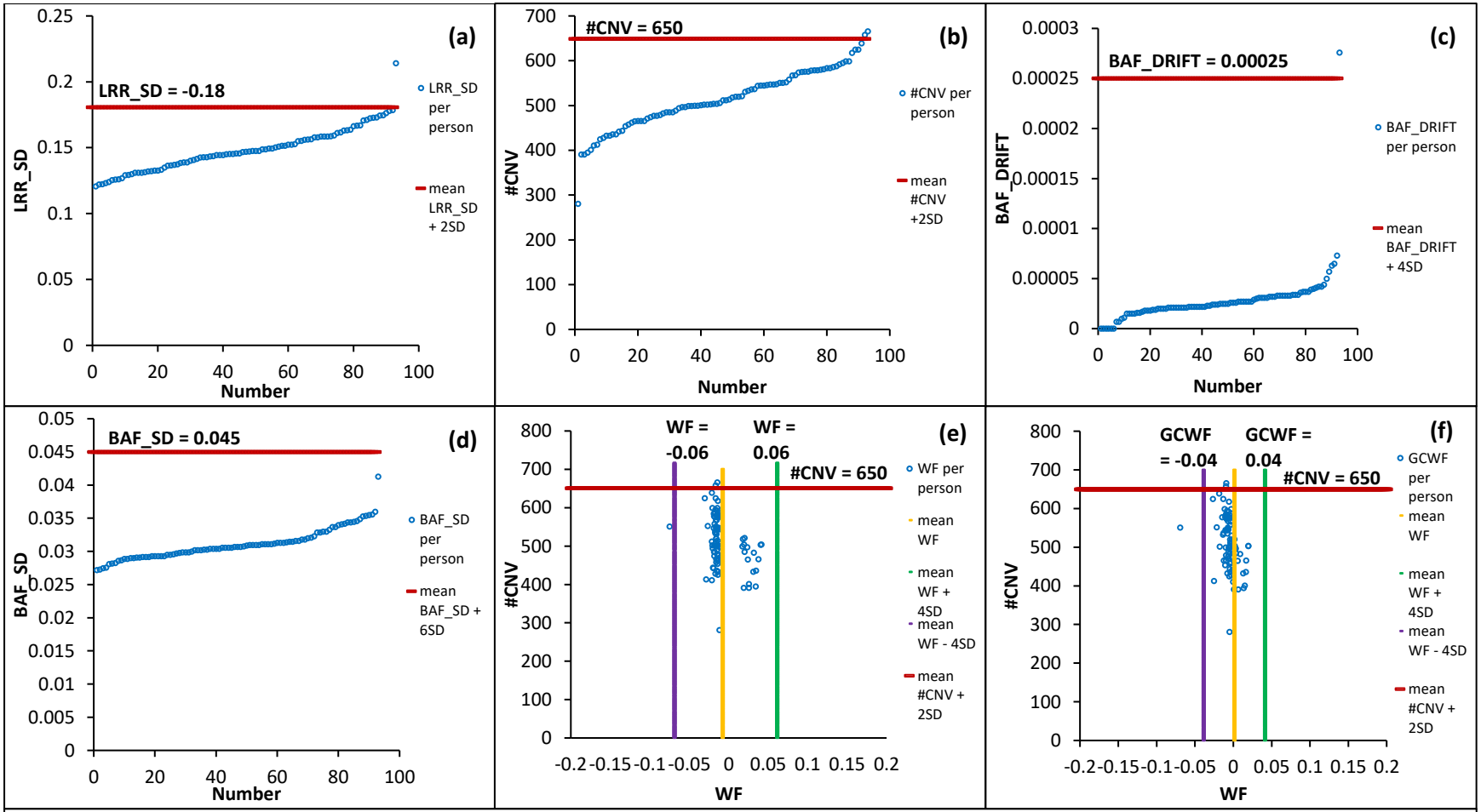


**Figure 5- 10:** Empirical plots for (a) LRR_SD (b) #CNV (c) BAF_DRIFT (d) BAF_SD (e) WF and (f) GCWF parameters, for 106 PMMO participants. LRR_SD = standard deviation of LRR; BAF_SD = standard deviation of BAF; BAF_DRIFT = departure score of observed BAFs from expected BAF; #CNV = Total CNVs; WF = Waviness Factor; GCWF = GC-adjusted WF.

| Parameter | Parameter cut-off criterion | Cut-off/Inclusion value(s) | PennCNV default settings for the inclusion value(s) |
|---|---|---|---|
| LRR_SD | μ + 2SD | value ≤ 0.18 | value ≤ 0.3 |
| BAF_SD | μ ± 6SD | value ≤ 0.045 | value ≤ 0.01 |
| BAF_DRIFT | μ + 4SD | value ≤ 0.00025 | none |
| #CNV | μ + 2SD | value ≤ 650 | value ≤ 100 |
| WF | μ ± 4SD | -0.06 < value < 0.06 | -0.05 < value < 0.05 |
| GCWF | μ ± 4SD | -0.04 < value < 0.04 | none |

**Table 5- 4:** Cut-off criteria for individual inclusion from the PMMO cohort (n=112), based on visual inspections of parameter distribution plots in **Figure 5- 11**. SD = standard deviation; μ = mean value per cohort; LRR_SD = standard deviation of LRRs per individual; BAF_SD = standard deviation of BAFs per individual; BAF_DRIFT = departure score of observed BAFs from expected BAFs per individual; #CNV = Total number of CNVs predicted by the PennCNV tool per individual; WF = Waviness Factor per individual; GCWF = GC-adjusted WF per individual.

Upon the exclusion of poor quality and spurious CNVs, 6 participants (~5.3%) and 27,699 CNV calls (~58.4%) were excluded from further analyses, while retaining 16,454 CNV calls in 106 participants. The CNVs were subsequently scanned for any overlaps between each other; those with ≥20% of their lengths overlapped, were merged as a single CNV call. The number of CNV calls (stratified by CNV types) before and after the filtering processes, are shown in **Figure 5- 11**. To summarise, the final CNV catalogue for 106 participants from the PMMO cohort contains 16,454 CNVs (~34.7%) that are considered good quality.

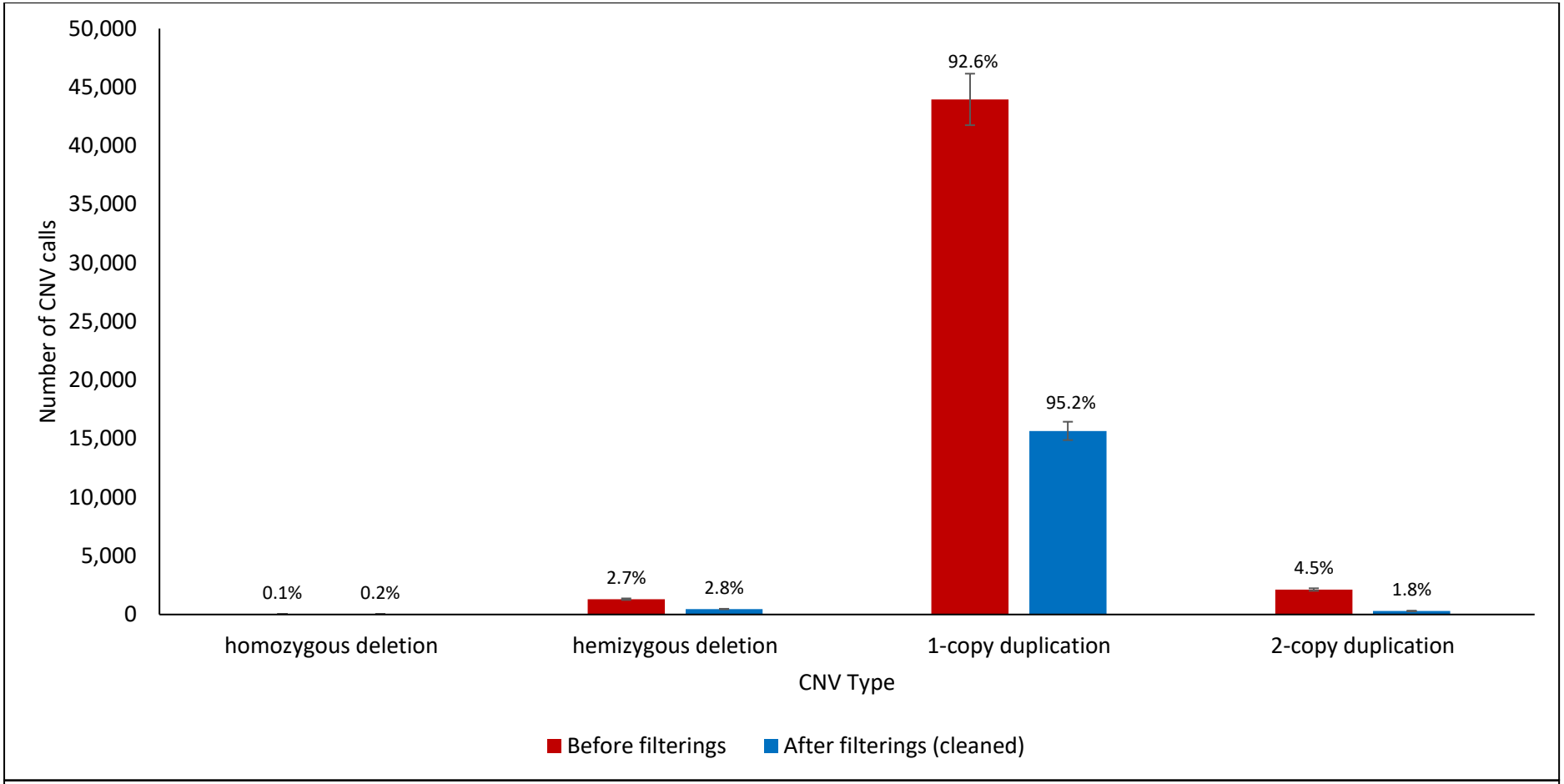


**Figure 5- 11:** The number of CNV types that were called by the PennCNV tool for the PMMO cohort, before (n=112; Red) and after (n=106; Blue) removing individuals with poor signal quality and low quality CNV calls. The percentages shown were the proportions of CNV types from the total number of CNV calls. CNVs = Copy number variants. Error bars = 5% error.

### 5.5.5. General characteristics of CNVs across study cohorts

As described in **Chapter 5.4.1 Statistical analyses**, sample t-tests were conducted to compare the proportions of each CNV type (%), average number of CNVs per person, and deletion-to-duplication ratio in the final CNV catalogues of NFBC1966, merged NFBC1986, NutriTech and PMMO cohorts. I observed no significant difference among the CNV types across the study cohorts, except for the CNV characteristics in the PMMO cohort ($p<0.05$). In the PMMO cohort, the deletions were significantly lower than the means in other cohorts, while the duplications were significantly higher than the means in other cohorts. These findings were as expected since the number of 1-copy duplications, albeit the huge reduction due to the parameter QC, were still relatively higher than the deletions. In addition, I also looked at comparing the results from these four cohorts, with previous findings from other studies (Shaikh *et al.*, 2009, Xu *et al.*, 2011, Kanduri *et al.*, 2013).

The analysis was extended to compare the lengths (means and medians in kb) and the frequencies (rare and singletons in %) of the inferred CNVs between the study cohorts in this chapter, as well as comparing our findings with the same publications as above (Kanduri *et al.*, 2013, Shaikh *et al.*, 2009, Xu *et al.*, 2011) (**Table 5- 6**) Rare CNVs were defined as the proportion of CNV calls (%) that had < 1% frequency within the same study cohort, whereas singletons were defined as the proportions of CNV calls (%) that occurred only in one person within the same cohort. **Table 5- 6** shows that among the study cohorts, the medians and means for CNV lengths in the NFBC1966 cohort were significantly different than the means and medians for CNV lengths in the other study cohorts ($p<0.05$). Additionally, I found that the proportions of rare CNVs (ranging from 7.77 to 15.05) among our study cohorts were significantly different than 39% that was reported by Xu *et al.* (2011) (Xu *et al.*, 2011), probably due to SNP contents of different arrays. Information on the percentage of rare CNVs

were not provided by the authors of the other two studies (Kanduri *et al.*, 2013, Shaikh *et al.*, 2009).

| Cohorts/ Studies | Sample size & population | Deletions, % (P-value) | | | Duplications, % (P-value) | | | Del-to-Dup ratio (P-value) | Mean #CNV, n (P-value) |
|---|---|---|---|---|---|---|---|---|---|
| | | All | Homo-zygous | Hemi-zygous | All | 1-copy | 2-copy | | |
| NFBC1966 | 4,905 healthy North Finnish adults | 71.54 (>0.1) | 14.89 (>0.1) | 56.65 (>0.1) | 28.46 (>0.1) | 28.33 (>0.1) | 0.13 (>0.1) | 2.51 (>0.1) | 5.29 (>0.1) |
| MERGED NFBC86 | 3,227 healthy North Finnish adolescents | 52.99 (>0.1) | 5.85 (>0.1) | 47.14 (>0.1) | 47.01 (>0.1) | 46.92 (>0.1) | 0.09 (>0.1) | 1.13 (>0.1) | 11.68 (>0.1) |
| NutriTech | 66 healthy UK adults | 69.38 (>0.1) | 3.1 (>0.1) | 66.28 (>0.1) | 30.62 (>0.1) | 30.43 (>0.1) | 0.19 (>0.1) | 2.27 (>0.1) | 7.82 (>0.1) |
| PMMO | 106 morbidly obese UK patients | **2.98 (<0.05)** | **0.16 (<0.05)** | **2.81 (<0.05)** | **97.02 (<0.05)** | **95.21 (<0.05)** | **1.81 (<0.05)** | **0.03 (<0.05)** | **191.38 (<0.05)** |
| Kanduri *et al.* (2013) | 286 healthy Finnish adults | 70.7 (0.271) | 12.4 (0.138) | NA | 29.3 (0.271) | NA | 0.05 (0.314) | 2.4 (0.207) | 20 (0.511) |
| Xu *et al.* (2011) | 8,148 healthy South East Asian adults | 68.03 (0.324) | 9.63 (0.337) | 58.4 (0.358) | 32.0 (0.324) | 31.83 (0.322) | 0.17 (0.425) | 2 (0.433) | 16 (0.467) |
| Shaikh *et al.* (2009) | 2,026 healthy Caucasians & African-Americans | NA | 6.1 (0.977) | 81.2 (0.073) | 12.7 (0.097) | NA | NA | **6.86 (0.003)** | 26.9 (0.595) |

**Table 5- 5:** Statistical results for one-sample t-test (two-tailed) on the proportions of each CNV type (%) across 4 different study cohorts (NFBC1966, NFBC1986, NutriTech, PMMO cohorts). The results were compared to the findings by other studies (Kanduri *et al.*, 2013, Shaikh *et al.*, 2009, Xu *et al.*, 2011). P-value < 0.05 was considered statistically significant (bold). CNV = copy number variant, #CNV = number of CNVs per person, NA = Data not available.

| Cohorts/ Studies | Cohort/Study description | Mean CNV length, kb (P-value) | | | Median CNV length, kb (P-value) | | Rare CNVs, % (P-value) | Singletons, % (P-value) |
|---|---|---|---|---|---|---|---|---|
| | | All | Deletion | Duplication | Deletion | Duplication | | |
| NFBC1966 | 4,905 healthy North Finnish adults | **176.2 (0.03)** | **100.9 (0.03)** | **225 (0.04)** | **73.17 (0.04)** | **105.5 (0.04)** | 7.77 (0.18) | 18.61 (0.40) |
| MERGED NFBC86 | 3,227 healthy North Finnish adolescents | 57.68 (0.35) | 36.26 (0.21) | 81.81 (0.36) | 12.97 (0.24) | 47.61 (0.67) | 8.94 (0.40) | 15.3 (0.30) |
| NutriTech | 66 healthy UK adults | 67.03 (0.49) | 43.07 (0.35) | 121.32 (0.42) | 24.45 (0.58) | 51.71 (0.82) | 10.3 (0.90) | 65.5 (0.05) |
| PMMO | 106 morbidly obese UK patients | 57.9 (0.35) | 57.05 (0.89) | 57.9 (0.18) | 20.88 (0.44) | 19.7 (0.14) | 15.05 (0.07) | 21.6 (0.52) |
| Kanduri *et al.* (2013) | 286 healthy Finnish adults | 52.39 (0.287) | 37.98 (0.238) | 87.32 (0.422) | NA | NA | NA | NA |
| Xu *et al.* (2011) | 8,148 healthy South East Asian adults | NA | NA | NA | 52.9 (0.238) | 80.23 (0.271) | **39 (0.0004)** | 6.33 (0.136) |
| Shaikh *et al.* (2009) | 2,026 healthy Caucasians & African-Americans | 38.35 (0.174) | 20.12 (0.074) | 125.1 (0.929) | 4.98 (0.134) | 44.76 (0.571) | NA | 22.2 (0.545) |

**Table 5- 6:** Statistical results for one-sample t-test (two-tailed) on the mean and median lengths of CNV types (kb), proportions of rare CNVs and singletons (%) across 4 different study cohorts (NFBC1966, NFBC1986, NutriTech, PMMO cohorts). The results were compared to the findings by other studies (Kanduri *et al.*, 2013, Shaikh *et al.*, 2009, Xu *et al.*, 2011). P-value < 0.05 was considered statistically significant (bold). CNV = copy number variant, NA = Data not available.

# 5.6. Discussions

## 5.6.1. Results overview

This chapter describes the results of CNV calling and filtering processes by the PennCNV tool for the NFBC1966, NFBC1986, NutriTech and PMMO cohorts. For each study cohort, the differences between the proportions of CNV calls before and after the filters had been applied, were depicted in **Figure 5- 2**, **Figure 5- 5**, **Figure 5- 6**, **Figure 5- 9** and **Figure 5- 11** respectively. The percentages of individuals who failed to survive the CNV filters ranged from 5.3% (PMMO) to 13.9% (NFBC1986) across the four cohorts ($p>0.06$). Furthermore, I found no significant difference on the proportions of CNV calls being excluded (range 58.4% - 82.4%) across the study cohorts ($p>0.06$).

The ratio for the number of CNVs between the two batches of NFBC1986 cohort was 1:3. A detailed inspection on the amounts of CNV calls predicted by the PennCNV tool, stratified by CNV types, revealed that the second batch (NFBC1986B) had a relatively higher proportion of 1-copy duplications (~84.6%) than the first batch (~28% in NFBC1986A). In this case, there was not expected to be a difference in the type of CNVs in these two batches, so this may be a technical artefact. Thus, the internal QCs were essential to clean up the CNV catalogue further, in which I have successfully reduced the false positive 1-copy duplications in the second batch of the NFBC1986 cohort by 91.5% (12-fold).

### 5.6.2. Comparison of CNV characteristics with other studies

The general characteristics of CNV calls were initially compared between the four study cohorts. In this study, I observed significant differences in the proportions of deletions and duplications in the PMMO cohort compared to the means of deletions and duplications from other study cohorts. This was as expected because the proportion of 1-copy duplications predicted in earlier sections was 97% of the total CNVs called for the PMMO cohort, which was 32-times higher than the deletion calls. Although the exact reason for this large difference was unknown, it is important to note that the array chip used to genotype 21 patients from the first batch of the PMMO cohort lacked any exonic probes compared to the ones used for the NFBC1986 cohort. Thus, smaller deletions in between the breakpoints of large duplications might not get detected accurately by the PennCNV algorithm. Furthermore, the merging of the CNV calls for the first and second batch of the PMMO cohort, and the subsequent merging of overlapping CNVs could also affect the distribution of deletions and duplications in the cohort. Nonetheless, more attention will be given to look into the validity of the duplication calls in my future work in order to increase the confidence of the CNV calls in this cohort. This can be done by looking at a small number of common duplications in the CNV catalogue of the PMMO cohort.

By looking at the lengths of different CNV types, I observed that the NFBC1966 cohort contained significantly larger CNVs overall, compared to the CNV lengths predicted in the other cohorts. This can be argued that compared to the other cohorts, the old array chip used to genotype the NFBC1966 cohort not only lacked any exonic probes, but it also had a relatively smaller genome-wide probe coverage (~300,000 probes as compared to >900,000 probes in Illumina's Human OmniExpress chips used in other cohorts). This may lead to low accuracy and sensitivity in the classification of the CNV calls. For example, two actual

duplications flanking a deletion might be called as one large duplication due to the lack of probes in-between the two duplications. The next step that can be done to address this issue is comparing the frequencies of several well-characterised common CNVs in the NFBC1966 cohort with the frequencies reported in other studies.

CNV characteristics were subsequently compared to the previous CNV studies on a Finnish population (n=286) (Kanduri *et al.*, 2013), a mixed population consisted of Caucasians and African-Americans (n=2,026) (Shaikh *et al.*, 2009), and a South East Asian population consisted of Chinese, Malays and Indians (n=8,148) (Xu *et al.*, 2011). The general characteristics of our CNV calls were largely consistent with those in the three studies, which used SNP genotyping arrays from Illumina. However, the deletion-to-duplication ratios in our study cohorts were significantly different to the ratio reported by Shaikh *et al.* (2009) (Shaikh *et al.*, 2009) (ratio=6.86; p=0.003; a mixed population that included Americans with African ancestry). There are a number of possibilities to explain the difference observed; 1) the author looked at duplications as a whole, without distinguishing between 1-copy and 2-copy duplications, and 2) the study excluded all autosomal duplications that had mean LRR < 0.25, which might have missed some 1-copy duplications with weak LRR signals from the probes. Therefore, the proportions of duplications reported could be underestimated, hence contributed to the significantly higher deletion-to-duplication ratio than those reported in our study cohorts. It is also important to note that a much higher genetic diversity exists among the Africans compared to the Caucasians due to ancient migration patterns and population bottlenecks (Campbell and Tishkoff, 2008). The presence of substructures within the mixed population could also be the reason that drives the differences observed in the proportions of deletions and duplications.

The results on the medians and means of CNV lengths were similar to those reported in the three studies. The significant difference (p=0.0004) between ours and Xu *et al.* (Xu *et al.*, 2011) in terms of the proportions of rare CNVs, can be attributed to the differences of genetic background between the populations under study. Our study cohorts were pre-dominantly Europeans (Finnish and British), with a large proportion of individuals being Caucasians, while the study populations by Xu *et.al*. (2011), were mainly South East Asians (Xu *et al.*, 2011). The author also suggested that there is an apparent difference in the CNV landscape between the Asians and the Caucasians, whilst evidently showing lesser efficiency in tagging common CNVs (by Illumina SNP arrays) in Asian populations, compared to the Caucasians (Xu *et al.*, 2011). This might indicate distinctions in the patterns of linkage disequilibrium (LD) between the two populations. Since the LD patterns are different (Xu *et al.*, 2011), it might in turn affect the frequencies of several CNV regions within the population, hence contributing towards increased (or decreased) proportions of rare CNVs within that population.

Overall, the characteristics of the inferred CNVs in this chapter can be compared to the general populations with European and Asian ancestries to a certain extent. However, I recommend taking the results for the PMMO cohort with extra caution, since the participants were morbidly obese patients selected for bariatric surgery, rather than 'healthy' individuals in the general population.

### 5.6.3. PennCNV tool as a preferred CNV calling tool

The work for this chapter used the PennCNV tool to call for CNVs based on the signal intensity information from the genotyped datasets. The PennCNV tool was preferred to other CNV calling tools, such as QuantiSNP (Colella *et al.*, 2007), dChip (Li *et al.*, 2008), Birdsuite (Korn *et al.*, 2008) or GADA (Pique-Regi *et al.*, 2008), simply due to several reasons: 1) it is the

most widely used CNV calling algorithm (~955 citations) compared to the QuantiSNP (~435 citations) (Mace *et al.*, 2016), 2) it offers more functionality and ease of format inter-changeability for downstream analyses (Winchester *et al.*, 2009), 3) its speed in processing a large number of samples is a key advantage over the other tools (Mace *et al.*, 2016), and 4) personally, I am familiar with perl-based scripting, which made it easier for me to operate the tool. Despite these reasons, false positive discoveries tend to occur regardless of which CNV calling algorithm that has been used. Therefore, post-processing steps such as the parameter filtering implemented in this chapter were important in reducing the false error rate. In addition, a study that evaluated the CNV calls predicted by different CNV calling algorithms (PennCNV, Birdsuite, and dChip) have shown that PennCNV has the highest sensitivity and specificity by showing less bias, high success rate in calling the CNVs and high reproducibility (87%) (Zhang *et al.*, 2014).

### 5.6.4. Hidden Markov Model to predict copy number changes

The PennCNV tool implements the Hidden Markov Model (HMM) as the basis of its algorithm calculation in predicting the copy number states from the LRR and BAF values (Wang *et al.*, 2007). The distance between consecutive SNPs, population frequency of B allele (PFB), along with the LRR and BAF values were used to calculate the likelihood of copy number variants between different copy number states (Wang *et al.*, 2007). These information on each SNP were incorporated into a hidden Markov Model (HMM) (Wang *et al.*, 2007). Briefly, HMM is an algorithm that can be used to model a situation that has limited internal states, in which changes of each internal state can generate several observations. For example, a situation where the signal information for SNP Z has a 1.5-fold increase compared to its immediate previous SNP Y, shows a change of internal states between the neighbouring SNPs.

The change of states from SNP Y to SNP Z can be due to an increase of either allele. Since the changes of internal states are invisible to us, the changes are considered 'hidden'. The 'current state' (SNP Z with 1.5-fold increase of signal) is assumed to only be dependent on the immediate previous state (SNP Y with a distinct signal), which itself is dependent on its own immediate previous state, and so on. This situation is assumed to follow a Markov process. Thereafter, the algorithm can predict the likelihood of a copy number for a specific stretch of DNA spanned by a number of SNPs. The potential of incorporating these signal information from high-density SNP data into a HMM was modelled by Wang, K., et. al. (2007) to develop a kilobase-resolution CNV calling tool for SNP genotyping data, called the PennCNV tool (Wang *et al.*, 2007).

# CHAPTER 6. Multi-Allelic CNVs in Human Amylase Gene Region

## 6.1. Introduction

This chapter describes the association analyses of multi-allelic copy number variants in human amylase gene cluster region with various obesity traits (BMI, WC, WHR), various body fat contents (pancreatic, total, subcutaneous, visceral), and dietary starch intakes. This work was conducted to investigate whether the association results between salivary amylase (*AMY1*) copy numbers and adiposity previously reported by Falchi *et. al.* (2014) (Falchi *et al.*, 2014), could be replicated in a relatively smaller, but well-characterised overweight-to-obese participants of the NutriTech cohort.

The human amylase region consists of salivary (*AMY1*) and pancreatic (*AMY2A* and *AMY2B*) α-amylase genes on chromosome 1 (Usher *et al.*, 2015). **Figure 6- 1** depicts the locations of different amylase genes within the amylase gene cluster region in humans. Its copy number has been understood to have undergone expansion throughout human evolution (Perry *et al.*, 2006) and this variation in the gene copy number has been highlighted in several genome-wide association studies (GWAS) (Iafrate *et al.*, 2004, Sudmant *et al.*, 2010). The amylase gene region has a complex, highly variable structure that include copy number deletions, duplications and inversions (Groot *et al.*, 1989). All amylase genes encode enzymatic proteins responsible for starch digestion and conversion into sugar. Due to the involvements in energy metabolism, these amylase genes may have different roles in modifying the body composition, thereby altering weight and adiposity in humans. Each gene varies in copy number; 2 to 17 copies for *AMY1*, 0 to 8 copies for *AMY2A*, and 2 to 6 copies for *AMY2B* (Groot *et al.*, 1991, Perry *et al.*, 2007, Usher *et al.*, 2015). The human genome reference assembly (GRCh37/hg19) shows three repeats of salivary *AMY1* gene with >99.9%

DNA sequence similarity between them (**Figure 6- 1**) (Kent WJ SC, 2002). Both pancreatic amylase (*AMY2A*, *AMY2B*) genes are located at the telomeric end of the region, each of which has >93% of its DNA sequence identical to salivary *AMY1* gene (Carpenter *et al.*, 2015, Kent WJ SC, 2002).

Recent work using quantitative PCR (qPCR) methods to determine copy number showed that reduced *AMY1* copy number is significantly associated with increased adiposity and having low *AMY1* copy number predisposes to common and childhood-onset obesity (Falchi *et al.*, 2014, Mejia-Benitez *et al.*, 2015, Viljakainen *et al.*, 2015). Falchi, *et. al.* (2014) first reported the association findings by conducting a family-based association analysis on 179 Swedish families to identify the associative relationship between signal intensity data of the probes on genotyping arrays (i.e. LRR and BAF) and adiposity-related transcriptomic data (Falchi *et al.*, 2014). The Swedish families were ascertained for obesity between the siblings (Falchi *et al.*, 2014). Among the probes that were on the DNA arrays used in that study, only one probe mapped to a CNV spanning the genomic region containing human salivary and pancreatic amylase genes (Falchi *et al.*, 2014). The author discovered significant inverse associations of the probe's signal intensity data with BMI and fat mass in the Swedish families (Falchi *et al.*, 2014), which later was replicated by another probe's signal intensity data in 972 participants of the TwinsUK project (Hysi *et al.*, 2010). Due to the presence of different amylase genes within the amylase cluster, the origin of the associated signals was distinguished by estimating the copy numbers for the salivary *AMY1* (ranged from 2 to 14 copies) and pancreatic *AMY2A* and *AMY2B* genes using qPCR methods. By using the estimated *AMY1* copy numbers, the author subsequently confirmed the associative relationships with both BMI and fat mass (Falchi *et al.*, 2014). Since then, other studies have also reported

significant associations of the salivary *AMY1* copy numbers with BMI and obesity (Marcovecchio *et al.*, 2016, Mejia-Benitez *et al.*, 2015).

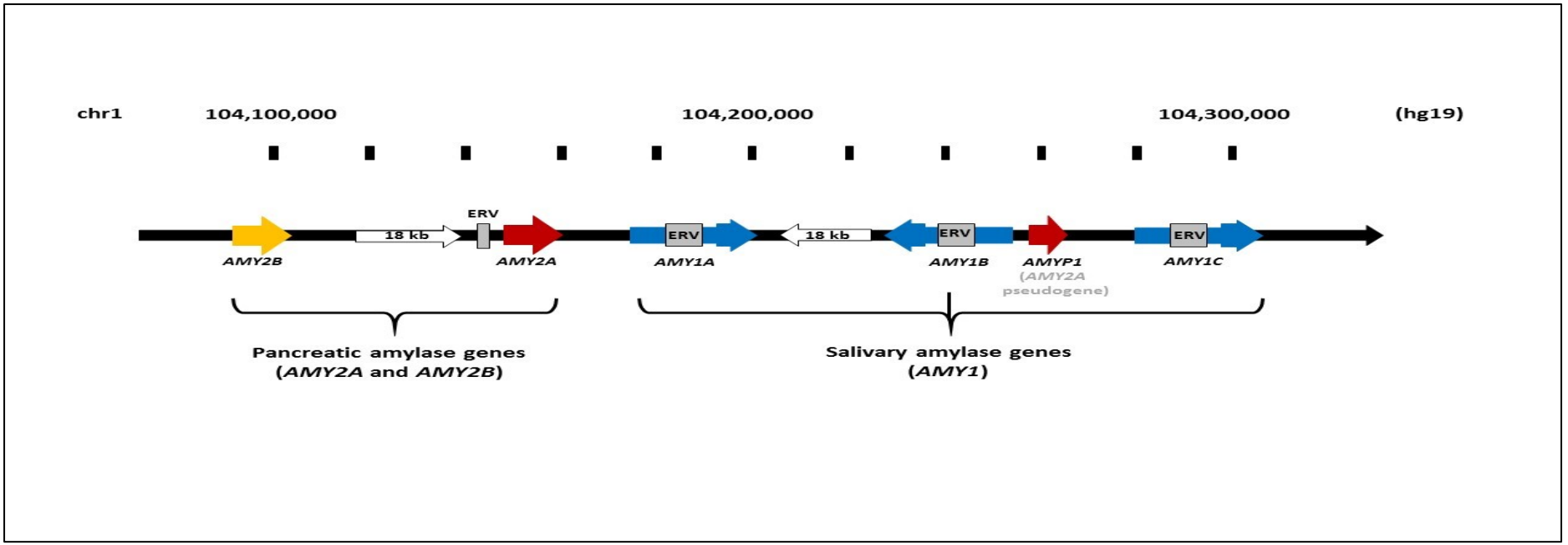


**Figure 6- 1:** The locations and orientations of the salivary *AMY1* gene and pancreatic *AMY2A* and *AMY2B* genes on chromosome 1 (hg19). The general overview of the amylase gene cluster region is that it consists of 3 copies of *AMY1* gene (blue arrow boxes denoted as *AMY1A*, *AMY1B*, and *AMY1C*), 1 copy of each pancreatic *AMY2* gene (yellow and red arrow boxes denoted as *AMY2A* and *AMY2B*), and 1 copy of AMY2A pseudogene (red arrow box denoted as *AMYP1*). There are two near-identical regions of 18-kb in length (shown as the white arrow box) that have high sequence identity with multiple parts in other loci. An endogeneous retrovirus (ERV) sequence located within each copy of *AMY1* gene has a high sequence similarity with another ERV locus on chromosome 12, which is the foundation of the PRT-based methods developed by Armour, J. A., *et. al.* (2007) (Armour *et al.*, 2007).

However, another study with a larger sample size (n=~3,500 individuals) has found contradictory results; Usher *et al.* (2015), via the use of higher-resolution molecular and computational methods, reported no association between *AMY1* copy number and BMI (Usher *et al.*, 2015). Furthermore, technical limitations and miscalibrations in CNV genotyping have been pointed out as a possible reason for these contrasting observations in analysing *AMY1* copy number (Carpenter *et al.*, 2015, Usher *et al.*, 2015). Carpenter *et al.* (2015) reported that qPCR measurements have poor clustering property and suggested that qPCR-based genotyping for *AMY1* CNVs may not be sufficiently reliable to distinguish single repeat unit differences in such a high polymorphic region (Carpenter *et al.*, 2015). In addition, they observed a significant numerical correlation between *AMY1* and *AMY2A* copy numbers ($r = 0.24$, $P = 3.6 \times 10^{-6}$) in the European population (Carpenter *et al.*, 2015). Given the high DNA sequence similarity between these two amylase genes, any structural variations in *AMY2A* could result in a weak *AMY1* copy number association. In order to resolve the fundamental structure in human amylase gene region, a combination of methods based on paralogue ratio tests (PRTs) of dispersed repeats were developed to provide a reliable, high-resolution assay capable of distinguishing repeat differences, even as low as single repeat unit differences for *AMY1*, *AMY2A* and *AMY2B* copy number analysis (Carpenter *et al.*, 2015).

It is known that salivary amylase is central to starch digestion in humans, but the effects of different *AMY1* gene dosages on the dietary starch intake have not been extensively studied. While the evidence remains scarce, dietary starch intake has been reported to modulate the *AMY1* copy number-body weight relationship in different dietary-starch groups (Rukh *et al.*, 2017). The study revealed that among individuals with high-starch diets, their body weights tend to increase with increasing copies of *AMY1* gene. However, in the low-starch diet group, having higher *AMY1* copies may reduce their body weights (Rukh *et al.*,

2017). This shows the importance of starch intake in influencing the body composition in humans.

The opposing findings have led to two different conclusions regarding the contribution and influence of the salivary *AMY1* gene on body weight and adiposity, which have made this issue as highly controversial. Therefore, the work in this chapter were carried out to explore this premise by using arguably the most sensitive assay for *AMY1* copy number genotyping developed by Armour J. A., *et. al.* (2007) (Armour *et al.*, 2007) in our highly characterised cohort, which is small but has an extensive, highly reproducible phenotyping. Due to the availability of dietary starch consumption in this cohort, this study was extended to explore the possibility of dietary starch intake in modifying the relationship between *AMY1* copy numbers and body weight, as reported previously (Rukh *et al.*, 2017).

## 6.2. Aim of this chapter

This chapter investigated the association of *AMY1* gene copy numbers with obesity traits (BMI, waist circumference, waist-to-hip ratio, adiposity contents) in middle-aged individuals. By using Paralogue Ratio Test (PRT), this chapter aimed to elucidate the effects of human *AMY1* copy numbers on obesity traits in a specific human intervention cohort (NutriTech) consisting overweight-to-obese individuals with a tight BMI range (25.0-35.6 kg/m$^2$). The investigation was extended to identify any effect of interaction between *AMY1* copy numbers and starch intake, on the obesity traits under study. The aim was addressed by the following questions:

1) *To ensure the validity of genetic analysis: does the study cohort represent a homogenous population of overweight-to-obesity individuals?*

I answered this question by carrying out multidimensional scaling (MDS) and constructing a plot of the first two dimension components, and anchored by the HapMap population data. This step was conducted to ensure the validity of genetic analysis.

2) *To ensure that the underlying assumptions of statistical tests are not violated: Do the distributions of the obesity-related phenotypes follow a normal distribution?*

I answered this question by comparing the distributions of BMI, WC, WHR, and various adiposity contents (in terms of body weight percentage) with a Normal distribution.

3) *To identify important potential confounding factors in the analysis: are the means of each phenotype affected by sex or age?*

I answered this question by comparing the means of BMI, WC, WHR, and various adiposity contents (in terms of body weight percentage) according to sex and age. The comparisons were performed via independent t-tests (for sex) and Analysis of Variance (ANOVA) test (for age).

4) *Are AMY1 copy numbers significantly associated with any obesity-related phenotypes?*

I answered this question by conducting different sets of linear regression tests to assess the associative relationships of *AMY1* copy numbers (as a continuous variable and as tertiles) with BMI, WC, WHR, and various adiposity contents (in terms of body weight percentage). Sex and age were used as covariates.

5) *Are AMY1 copy numbers significantly associated with starch intake?*

I answered this question by conducting different sets of linear regression tests to assess if the associative relationships of *AMY1* copy numbers (as tertiles) with BMI, WC, WHR, and various adiposity contents (in terms of body weight percentage) were modified by the dietary starch intake (as tertiles). Sex and age were used as covariates.

## 6.3. General methods used in this chapter

| Section | Method name (Described in Chapter 3) | Chapter 4 | Chapter 5 | Chapter 6 | Chapter 7 | Chapter 8 |
|---|---|---|---|---|---|---|
| **3.2** | Biological sample collection | ✓ | ✓ | ✓ | ✓ | ✓ |
| **3.3** | Genomic DNA extraction | ✓ | ✓ | ✓ | ✓ | ✓ |
| **3.4** | Obesity-related phenotypic measurements | ✓ | ✓ | ✓ | ✓ | ✓ |
| **3.5** | SNP genotyping | ✓ | ✓ | ✓ | ✓ | ✓ |
| **3.6** | Quality controls | ✓ | ✓ | ✓ | ✓ | ✓ |
| **3.7** | Population structure adjustments | ✓ | ✓ | ✓ | ✓ | ✓ |
| **3.8** | Principal component analysis | | | | | ✓ |
| **3.9** | CNV calling | | ✓ | | ✓ | ✓ |
| **3.10** | CNV quality controls | | ✓ | | ✓ | ✓ |

## 6.4. Specific methods

Specific methods were conducted to address the aim of this study chapter. The specific methods are described below.

### 6.4.1. Signal intensity correlation with BMI Using SNPs in Human Amylase Region

The Log R Ratio (LRR) and B allele frequency (BAF) information from each genotyped probe of each sample were extracted from GenomeStudio software. As the LRR values did not correlate to the number of copies in a linear way (LRR of -1/0/0.58 corresponds to copy number 1/2/3), the LRR values were inversely-logged to obtain R Ratio (RR) values. RR values are linearly correlated with the number of copies (RR of 0.5/1/1.5 corresponds to copy number 1/2/3). A total of 35 SNPs on the genotyping array were located along a 500-bp region (hg19; chr1:103,900,000-104,400,000) containing *AMY1*, *AMY2A* and *AMY2B* genes. Linear regression tests were conducted to identify the effect sizes (as β-coefficients) of those SNPs, and if any of the SNPs' signal intensity ratio were significantly associated with BMI within that 500-bp region. Sex, age, BAF information per SNP, and LRR variance per SNP were used as covariates in the regression analyses.

This method was done to replicate the method used by Falchi *et al.* that identified significant *AMY1* copy number association with BMI based on the probes on the human amylase region (Falchi *et al.*, 2014).

### 6.4.2. CNV genotyping using Multiplex Paralogue Ratio Tests

Initially, 72 NutriTech genomic DNA samples were sent to Armour J., *et al.* (University of Nottingham, UK) for CNV genotyping of each amylase gene using a combination of paralogue ratio tests (PRTs) from dispersed repeats and microsatellite analysis (Armour *et al.*, 2007, Carpenter *et al.*, 2015, Walker *et al.*, 2009). Briefly, an amylase-specific PRT was developed based on an endogenous retroviral (ERV) sequence located in *AMY1*, but not in *AMY2A* and *AMY2B* genes (Carpenter *et al.*, 2015, Samuelson *et al.*, 1996). As the same ERV sequence was also located in a specific region on chromosome 12 (as reference copy number), the ERV-based primer simultaneously amplified the *AMY1* ERV product together with a two-copy ERV product from chromosome 12. In addition, a tetranucleotide (TATC) microsatellite within the *AMY1* repeat units was identified and the microsatellite peaks were observed for their numbers and signal strengths. They were used to aid and match *AMY1* copy number from PRTs (Carpenter *et al.*, 2015). The results of PRTs and the *AMY1* microsatellite assay were combined to generate the most likely copy numbers based on Maximum Likelihood analysis. The copy numbers were calibrated using 16 three-generation CEPH pedigrees and validated by fibre-FISH and read depth from 1000 Genomes Project (Carpenter *et al.*, 2015). A detailed methodology was described by Armour J., *et al.* (Armour *et al.*, 2007).

### 6.4.3. Statistical analyses

The means for dietary starch intake (g/day) were adjusted with the daily energy intake (kcal/day). Both energy-adjusted starch intake and *AMY1* copy numbers were separated into tertiles. Continuous variables such as *AMY1* copy numbers, age, BMI, waist circumference, waist-to-hip ratio (WHR), fat percentages, daily energy intake, daily macronutrient intakes (protein, carbohydrates, fat) were compared between their means across the starch intake tertiles, using a general linear regression model adjusted for age and sex. The distribution of sex as a categorical variable, was compared by using a chi-square test.

Obesity traits investigated in the present study were BMI, waist circumference, WHR, pancreatic fat content, total abdominal fat (%), subcutaneous fat (%) and visceral fat (%). The mean values of these obesity traits were adjusted for age and sex, and then were compared using a general linear regression model across *AMY1* copy number tertiles.

The association analyses of *AMY1* copy numbers as a continuous variable, with obesity traits, were conducted according to the tiles of energy-adjusted starch intake. The general linear regression tests were adjusted for age and sex. Another general linear regression model was tested for the association analyses of starch intake tertiles as a continuous variable, with obesity traits (adjusted for age, sex and daily energy intake).

One-way Analysis of Variance (ANOVA) was used to compare the means of *AMY1* copy number. A rank-based nonparametric test (Kruskal-Wallis H test) was used to determine if there are statistically significant differences between the copy numbers on the obesity traits. The effect of interaction was also tested between *AMY1* copy number and energy-adjusted starch intake (tertiles) on obesity traits (continuous) in a custom linear model that included a multiplicative variable (as the variable for interaction) consisting of both predictors, adjusted for age and sex.

All means and 95% confidence intervals (CI) for obesity traits were calculated (adjusted for age, sex and daily energy intake) across the energy-adjusted starch intake and *AMY1* copy number tertiles by using a general linear regression model. A P-value < 0.05 was considered statistically significant. Post-hoc tests were implemented to determine if there were significant mean differences between the tertiles of starch intake, as well as between *AMY1* copy number tertiles, on obesity traits. All statistical analyses were carried out using IBM SPSS Statistics for Windows Version 24.0 (Released 2016; IBM Corp., Armonk, New York. USA).

## 6.5. Results

### 6.5.1. Quality controls & population structure adjustments

I would like to remind readers that the NutriTech cohort is a dietary intervention cohort consisted of overweight and obese participants (n=74), from which half of them undertook a 12-week period of strict dietary change, various tests (i.e. Oral Glucose Tolerance Test, Mixed Meal Test) and physical activities, while the remaining half had standard diets for the whole duration of the 12-week period (a detailed description of this cohort can be found in **Chapter 2.3 NutriTech cohort**).

One NutriTech participant (n=1) was excluded due to poor DNA quality. Another sample was detected to have a high inbreeding coefficient and was excluded as well (Refer to **Chapter 4.5.4 NutriTech cohort**). According to **Figure 4- 17**, a homogeneous clustering of the NutriTech participants was clearly shown within the European continent based on 171,523 common SNPs shared across HapMap Phase 3 population data. Additionally, 4 samples who had incomplete data in terms of sex and BMI were also excluded. In total, 6 from 74 samples (8.1%) were removed prior to any statistical analyses. The final NutriTech cohort analysis, thus, consisted of 36 participants (of which 50% males) in the diet intervention group (Group 1) and 32 participants (of which 51.6% males) in the control group (Group 2).

### 6.5.2. Distributions of phenotypes and amylase copy numbers

This step was carried out to ensure that the assumptions of the statistical tests in downstream analyses were not violated, and thus was not an investigation of association in itself. Otherwise, the multiple-testing problem could be further magnified.

*AMY1 copy numbers* – **Figure 6- 2** shows the overall distribution of estimated *AMY1* copy numbers in the NutriTech cohort. The residuals, which were the differences between observed and expected copy numbers (if they were to follow a normal distribution), were standardised across all copy numbers. The standardised values were used for the histogram and P-P plot, as shown **Figure 6- 2**. We observed the distributions to be significantly different than normal ($p<3.3 \times 10^{-6}$).

*Obesity-related phenotypes*– Standardised residuals were calculated from each obesity trait and adiposity measurement to observe whether the distribution of such residuals follow a normal distribution. **Figure 6- 3** shows histograms and normal P-P plots of standardised residuals for BMI, WC and WHR, at baseline, for 68 NutriTech participants. The mean residuals (standardised) for all three obesity traits were zero. The standardised residuals for BMI ranged from -5.18 to 6.74, WC ranged from -19.11 to 22.34, and -0.13 to 0.11 for WHR. Since these residuals were approximately normally distributed, the obesity traits were assumed to follow a normal distribution.

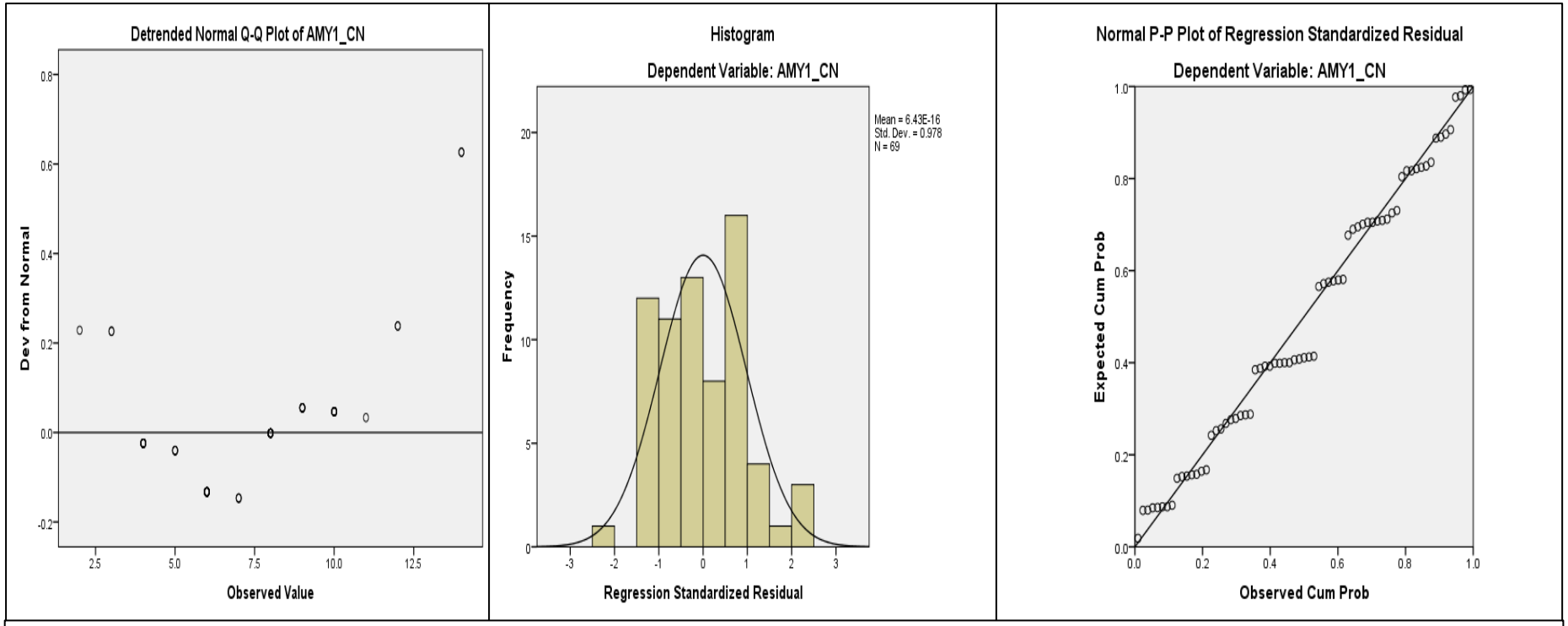


**Figure 6- 2:** The overall distribution of the salivary amylase (*AMY1*) copy numbers in the NutriTech cohort (n=68). The Normal Q-Q plot (left) of *AMY1* copy numbers displays the deviation of observed copy numbers from the normal line. The difference between the observed and expected normal copy numbers were defined as residuals. The standardised residuals were used in the histogram (centre) and plotted in the PP plot (right) to observe if, at least, the residuals were normally distributed. We can see from the P-P plot that the distribution of the estimated AMY1 copy numbers did not follow a normal distribution.

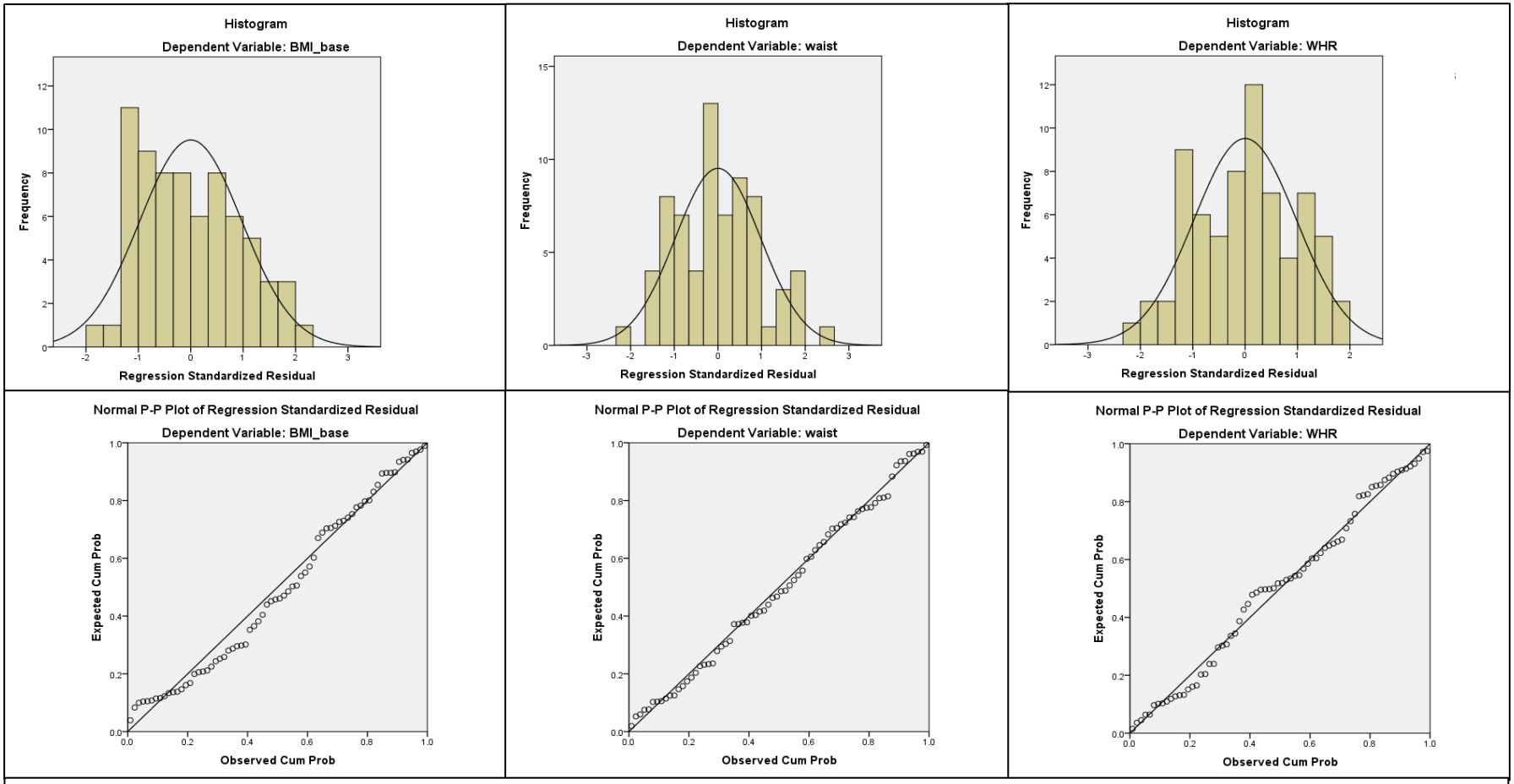


**Figure 6- 3**: The distributions of standardised residuals of BMI, waist circumference, and WHR, in the NutriTech cohort (n=68). BMI_base = body mass index at baseline; waist = waist circumference at baseline; WHR = waist-to-hip ratio at baseline.

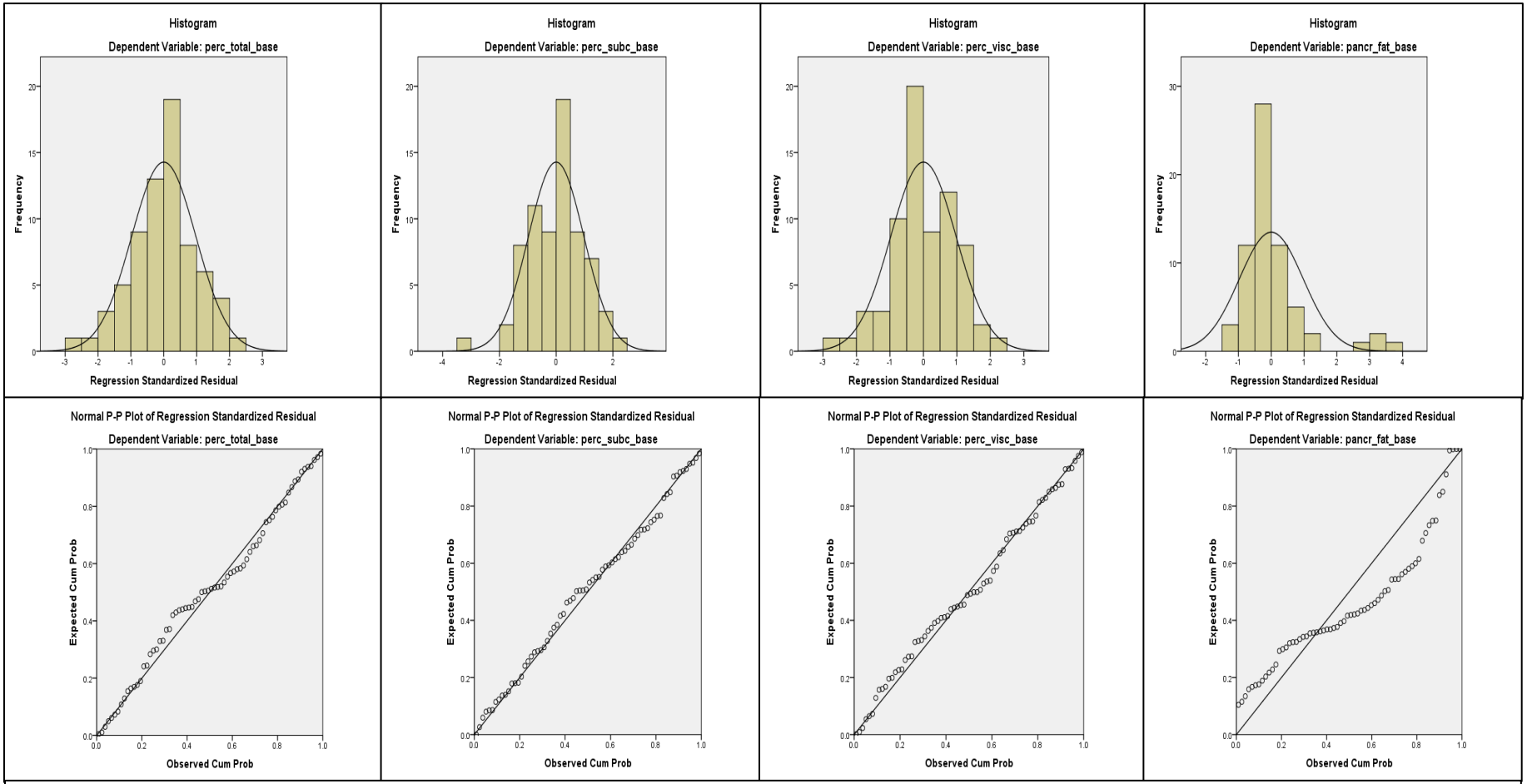


**Figure 6- 4:** The distributions of standardised residuals of total abdominal fat (perc_total_base), subcutaneous fat (perc_subc_base), visceral fat (perc_visc_base) and pancreatic fat (pancr_fat_base) at baseline, in the NutriTech cohort (n=68).

*Adiposity measurements* – **Figure 6- 4** displays the overall distributions of standardised residuals for each adiposity measurement in terms of histograms and normal P-P plots. Similar to obesity traits, the means of standardised residuals for all adiposity measurements were zero, with residuals ranged from -15.14 to 12.50 for total fat, -16.58 to 10.73 for subcutaneous fat, -4.23 to 3.31 for visceral fat, and -5.18 to 6.74 for pancreatic fat. The standardised residuals were observed to be approximately normally distributed along the normal line/curve. Therefore, all adiposity measurements were assumed to follow a normal distribution.

### 6.5.3. Sex, age and group effects

Independent t-tests (for sex) and ANOVA (for age) were conducted to compare the means for all obesity-related phenotypes (i.e. BMI, WC, WHR, pancreatic fat, total fat %, visceral fat %, and subcutaneous fat %). As described previously in **Chapter 2.3 NutriTech cohort,** the participants in the NutriTech cohort were classified into two groups; Group 1 (dietary intervention) and Group 2 (controls). To determine whether differences in group classification could have a significant effect on the differences in these phenotypes, t-tests were also carried out on the obesity-related phenotypes, adiposity measurements, and starch intake.

*Obesity-related phenotypes* – As expected, sex difference had a significant effect on the baseline means for waist circumference ($p=0.001$; Mean difference=7.86; [3.44, 12.27] 95% CI) and WHR ($p=2.26 \times 10^{-10}$; Mean difference=0.014; [0.076, 0.132] 95% CI), but not on mean BMI ($p=0.608$). Age difference had no significant effect on BMI ($p=0.05$) or any fat measurements ($p>0.5$) in this cohort.

*Adiposity (in body weight %)* – As expected, we found that there was a significant difference in total abdominal fat % ($p=2.99 \times 10^{-16}$; Mean difference=-14.90; [-17.67, -12.12] 95% CI), subcutaneous abdominal fat % ($p=2.90 \times 10^{-21}$; Mean difference=-16.29; [-18.66, -13.92] 95% CI), and visceral fat % (p=0.001; Mean difference=1.27; [0.57, 1.97] 95% CI), but not on pancreatic fat % (p=0.097), when compared between males and females.

*Between-Group* – Since we only used the trait information at baseline in this chapter, we did not expect to have any significant differences between the two groups who were later subjected to different dietary conditions, in BMI, waist circumference, WHR, as well as in adiposity measurements. Independent t-tests revealed no significant differences in starch intake, any obesity traits, or any fat measurement (p>0.1) between Group 1 and Group 2. The non-significant results were the same when the tests between the two Groups were stratified by sex (p>0.05). Thus, group classification was not used as a covariate in subsequent association analyses.

*AMY1 copy numbers* – Sex, age and group classifications were tested for any significant effect on the estimated *AMY1* copy numbers. Independent t-tests revealed no significant difference on the estimated *AMY1* copy numbers between males and females (p=0.151; mean difference= -0.901; [-2.142, 0.339] 95% CI]). In addition, we observed no significant differences in the estimated *AMY1* copy numbers between the two groups (p=0.073; mean difference= 0.633; [-1.190, 1.335] 95% CI]). An ANOVA between age and *AMY1* copy numbers suggested that age had no significant effect on the *AMY1* copy numbers estimated in the NutriTech cohort (p=0.202).

*Between-phenotypes correlation analysis:* At baseline, BMI was significantly correlated with waist circumference, WHR, total and subcutaneous fat ($p<0.05$; Pearson's coefficient (r) > 0.280), but not with pancreatic fat content and visceral fat %. When tested according to different sexes, the significant correlations were still observed in males ($p<0.05$). In addition, pancreatic fat was also significantly correlated with BMI in males ($p=0.005$; $r=0.475$). In females, BMI was significantly correlated with waist circumference ($p=9.77 \times 10^{-9}$), total fat ($p=3.15 \times 10^{-7}$), and subcutaneous fat ($p=4.46 \times 10^{-8}$), but not with WHR.

### 6.5.4. Association analysis of AMY1 copy numbers with obesity-related phenotypes, adiposity contents and starch intakes

The summary characteristics of the study cohort according to the energy-adjusted starch intake tertiles and *AMY1* copy number tertiles are shown in **Table 6- 2** and **Table 6- 3** respectively.

*Starch intake (tertiles) association* (**Table 6- 2**)*:* Across the starch intake tertiles, no significant differences were observed in age, *AMY1* copy number (as a continuous variable), BMI, WC, WHR, or any of the fat content %. Interestingly, I found that on average, the participants in the high starch intake tertile had numerically lower BMI, lower total fat % and lower subcutaneous fat %, than the participants in the lower tertiles, but with non-significant P values (**Table 6- 2**). Independent sample t-tests were subsequently performed to identify any significant differences in the means of BMI, total fat % and subcutaneous fat %, between the participants in the lowest and the highest starch intake tertiles. I found that both total ($p=8.88 \times 10^{-5}$) and subcutaneous ($p=0.0002$) fat % were significantly different for participants

| Variables | Starch intake | | | P-value |
|---|---|---|---|---|
| | Low (n=22) | Medium (n=22) | High (n=22) | |
| Energy-adjusted starch intake, g/day | 74.474 (52.5-86.6) | 98.95 (86.84-111) | 143.68 (113.66-211.88) | |
| Females, n | 15 (68.18%) | 12 (54.55%) | 5 (22.73%) | **0.008** |
| Age, y | 59.41 (57.53,61.28) | 59.64 (57.73,61.54) | 58.36 (56.43,60.3) | 0.398 |
| Estimated *AMY1* copy number, n | 7.43 ± 2.87 | 6.91 ± 2.22 | 7.09 ± 2.89 | 0.839 |
| | | | | |
| BMI, $kg/m^2$ | 29.31 (28.17,30.46) | 28.59 (27.34,29.84) | 29.21 (27.83,30.59) | 0.641 |
| Waist circumference (WC), cm | 98.17(93.57,102.77) | 95.37 (91.25,99.49) | 98.93 (94.35,103.51) | 0.307 |
| Waist-to-hip ratio (WHR) | 0.99 (0.87,0.95) | 0.9 (0.86,0.94) | 0.93 (0.9,0.96) | 0.315 |
| Pancreatic fat, fat:water+fat ratio | 5.20 (3.24,7.16) | 6.59 (3.05,10.13) | 5.02 (2.78,7.26) | 0.553 |
| Total fat, % | 41.72 (39.22,44.22) | 37.36 (32.92,41.8) | 31.25 (27.04,35.47) | 0.081 |
| Subcutaneous fat, % | 33.08 (30.13,36.03) | 29.26 (25.01,33.52) | 22.94 (18.67,27.2) | 0.116 |
| Visceral fat, % | 4.81 (4.01,5.61) | 4.5 (3.84,5.16) | 4.52 (3.81,5.23) | 0.18 |
| | | | | |
| Energy intake, kcal/day | 1376.64 (1304.6,1448.68) | 1798.24 (1743.57,1852.9) | 2568.61 (2345.43,2791.79) | **<0.0001** |
| Protein intake, g/day | 60.27 (53.51,67.03) | 70.09 (65.96,74.22) | 97.56 (88.43,106.69) | **<0.0001** |
| Fat intake, g/day | 54.78 (50.15,59.4) | 70.88 (66.12,75.64) | 104.7 (92.88,116.53) | **<0.0001** |
| Carbohydrate intake, mg/day | 153.55 (140.6,166.5) | 204.95 (187.75,222.14) | 300.05 (258.05,342.04) | **<0.0001** |

**Table 6- 2:** The main characteristics of NutriTech participants according to energy-adjusted starch intake tertiles at baseline (Week 4). The values shown are means (95% CIs) for continuous variables, means ± SDs for *AMY1* copy number, n (%) for categorical variables such as sex, and means (ranges) for energy-adjusted starch intake. All values were adjusted for age and sex (except for 'Female' variable where it was adjusted for age only). P-value < 0.05 were considered significant (bold).

| Variables | *AMY1* copy numbers | | | P-value |
|---|---|---|---|---|
| | Low (n=36) | Medium (n=12) | High (n=20) | |
| Estimated *AMY1* copy number | 5.11 ± 1.24 | 7.62 ± 0.65 | 10.35 ± 1.53 | |
| Females, n | 15 (41.6%) | 5 (38.5%) | 14 (70%) | 0.087 |
| Age, y | 59 (57.7,60.3) | 60.92 (58.74,63.11) | 58.35 (55.92,60.78) | 0.271 |
| Energy-adjusted starch intake, g/day | 113.42 (68.96-204.37) | 100.35 (52.5-211.88) | 97.08 (55.06-168.93) | 0.380 |
| | | | | |
| BMI, $kg/m^2$ | 29.12 (28.1,30.15) | 30.12 (28.46,31.78) | 28.36 (27.14,29.58) | 0.186 |
| Waist circumference, cm | 98.42 (94.92,101.92) | 101.45 (94.84,108.07) | 93.39 (89.85,96.92) | 0.332 |
| Waist-to-hip ratio (WHR) | 0.93 (0.9,0.96) | 0.93 (0.89,0.98) | 0.87 (0.84,0.9) | 0.200 |
| Pancreatic fat, fat:water+fat ratio | 5.93 (3.54,8.31) | 8.29 (3.53,13.05) | 3.74 (2.42,5.06) | 0.335 |
| Total fat, % | 36.57 (33.55,39.59) | 36.13 (29.71,42.54) | 38.48 (33.75,43.22) | 0.307 |
| Subcutaneous fat, % | 27.88 (24.68,31.08) | 27.63 (21.34,33.92) | 30.84 (26.36,35.32) | 0.325 |
| Visceral fat, % | 4.77 (4.23,5.31) | 4.91 (3.86,5.95) | 4.14 (3.44,4.85) | 0.833 |
| | | | | |
| Energy intake, kcal/day | 2047.37 (1850.46,2244.28) | 1822.38 (1320.98,2323.78) | 1766 (1531.54,2000.46) | 0.380 |
| Protein intake, g/day | 78.61 (72.27,84.95) | 71.37 (50.9,91.84) | 73.78 (62.38,85.19) | 0.486 |
| Fat intake, g/day | 82.51 (73.32,91.69) | 65.41 (46.31,84.5) | 73.79 (60.87,86.7) | 0.183 |
| Carbohydrate intake, mg/day | 238.43 (206.97,269.88) | 218.74 (149.45,288.02) | 193.03 (164.76,221.29) | 0.413 |

**Table 6- 3:** The main characteristics of NutriTech participants according to *AMY1* copy number tertiles at baseline (Week 4). The values shown are means (95% CIs) for continuous variables, means ± SDs for *AMY1* copy number, n (%) for categorical variables such as sex, and means (ranges) for energy-adjusted starch intake. All values were adjusted for age and sex (except for 'Female' variable where it was adjusted for age only). P-value < 0.05 were considered significant (bold).

in the highest starch intake tertile, as compared to the ones in the lowest tertile. For BMI, no significant difference was observed.

Interestingly, the number of females were significantly reduced with increasing starch intakes (**Table 6- 2**), where there were significantly more females in the lowest starch-intake group compared to the highest starch-intake group (p=0.008). The means of starch, energy, protein, fat and carbohydrate intakes were subsequently compared between males and females (using independent t-tests), revealing significant mean differences in those intakes between males and females (p<0.02). Furthermore, the mean energy-adjusted starch intake in males was 118.2 g/day ± 39.3 g/day (n=34; range=52.5 to 211.8 g/day), which is significantly higher than in females at p = 0.002 (n=32; 92.4 g/day ± 20.6 g/day, range= 55.1 to 161.2 g/day). Given the gender differences in fat contents, this could explain the lower fats generally observed in higher starch eaters. Since there were significantly higher intakes of macronutrients in males compared to females, it is expected that the participants in the higher starch intake tertile (containing a higher proportion of males) had, on average, significantly higher energy, protein, fat and carbohydrate intakes compared to the participants in lower starch intake tertiles, which contained a lower proportion of males (p<0.0001). This shows that males tend to consume more starch and macronutrients than females.

*AMY1 copy number (tertiles) associations* (**Table 6- 3**)*:* Across *AMY1* copy number tertiles at baseline, I did not find any significant differences in age, sex, obesity-related phenotypes and macronutrient intakes. Nevertheless, it is interesting to know that participants with low *AMY1* copy numbers had, on average, the highest starch intake, as compared to the participants with high *AMY1* copy numbers **(Table 6- 3).** Subsequently I then compared, using independent

sample t-tests, the means of starch intakes between the lowest and the highest *AMY1* copy number tertiles. However, no significant difference was observed in the means of energy-adjusted starch intakes between the lowest and the highest *AMY1* copy number tertiles (p=0.064).

*AMY1 copy number (continuous) associations:* **Figure 6- 5** and **Figure 6- 6** show the mean values of obesity-related phenotypes for each *AMY1* copy number (ranging from 2 to 14 copies) that were genotyped. More importantly, I did not find any significant association between any *AMY1* copy number and BMI (p=0.186), WHR (p=0.2), or any of the fat measurements (p>0.3) (**Table 6- 3**)**.** Moreover, linear regression tests between continuous *AMY1* copy numbers and BMI at baseline (adjusted for sex and age) revealed a non-significant association (p=0.444).

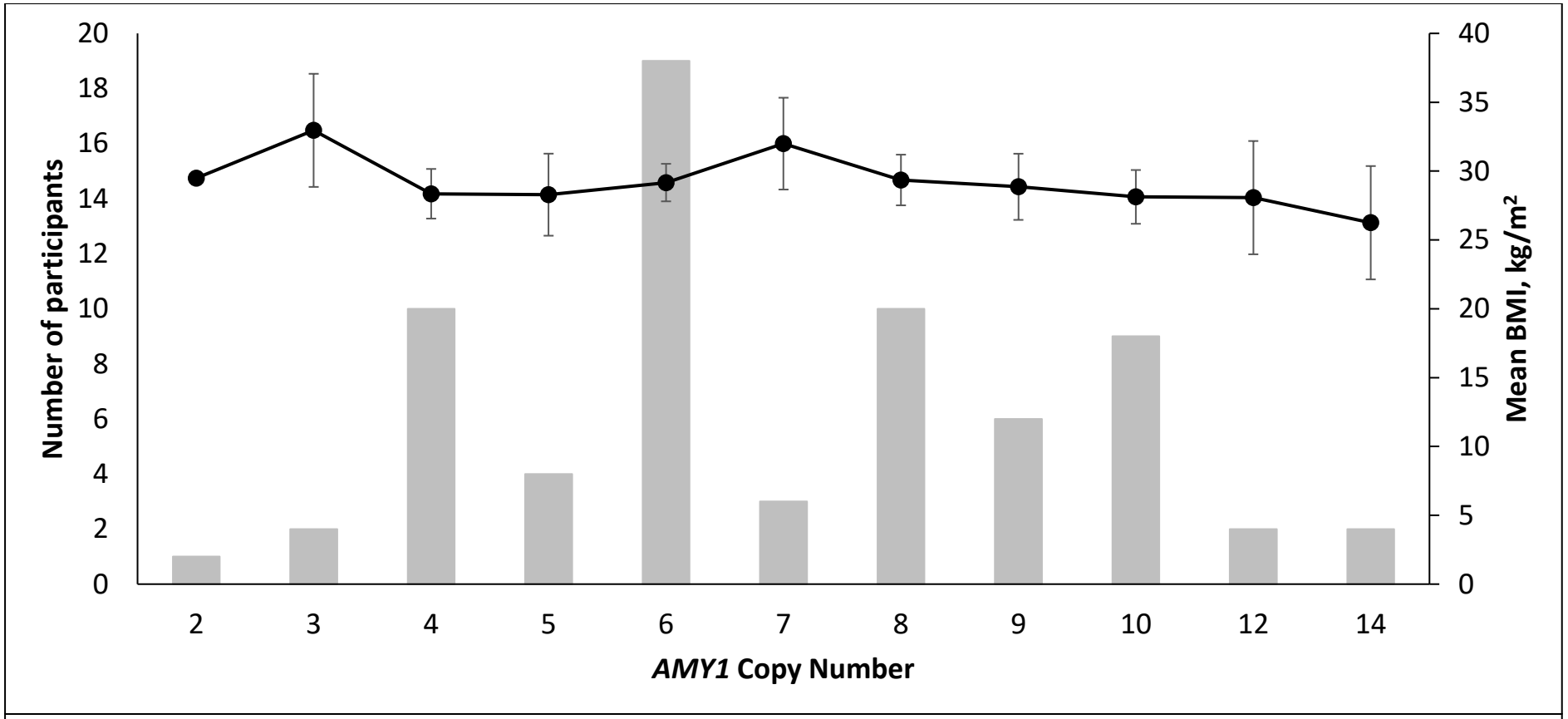


**Figure 6- 5:** The mean BMI of participants from the NutriTech cohort (n=68) according to *AMY1* copy numbers. The grey bars represent the total number of participants (y-axis) for each estimated *AMY1* copy number (x-axis). The black dots and lines represent the mean BMI of the participants (y-axis) for each estimated *AMY1* copy number (x-axis). *AMY1* = salivary amylase gene. Error bars = 95% confidence intervals of means (CI).

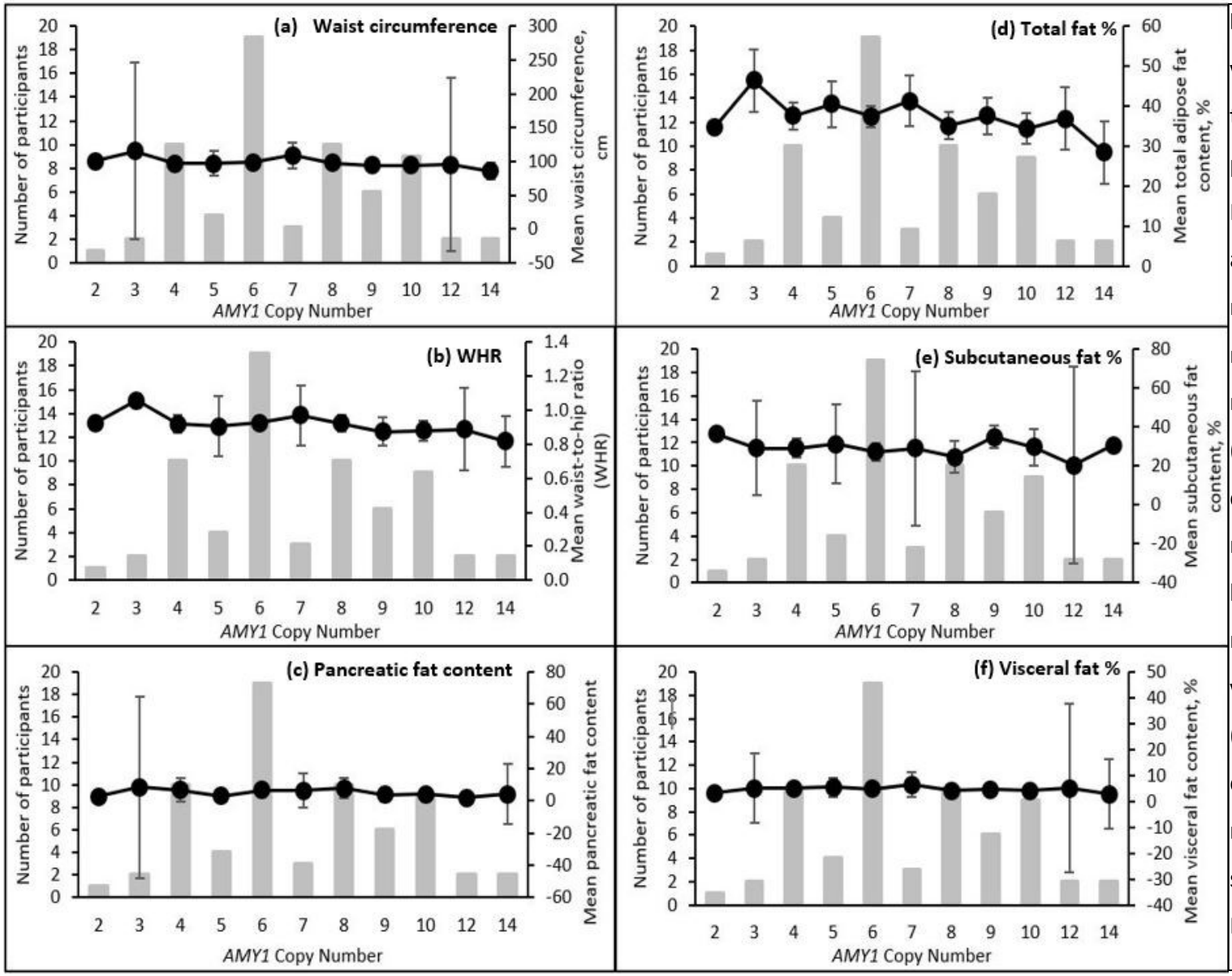


**Figure 6- 6:** The mean values of other obesity traits (**a-f**) for participants from the NutriTech cohort (n=68) according to *AMY1* copy number. The grey bars represent the total number of participants (y-axis) for each estimated *AMY1* copy number (x-axis). The black dots and lines represent the mean values of the participants (y-axis) for each estimated *AMY1* copy number (x-axis). *AMY1* = salivary amylase gene. Error bars = 95% CI.

### 6.5.5. Assessing the accuracy of PRT-based *AMY1* copy numbers

To assess how well the PRT-estimated *AMY1* copy numbers match the copy numbers estimated from signal intensity data, I carried out bivariate correlation tests between the PRT-estimated *AMY1* copy numbers and the LRR or RR for a SNP (rs425970) overlapping the salivary *AMY1* gene on chromosome 1. Generally for a 'normal' 2-copy gene, the RR values should be correlated with the number of copies of a SNP located within the 'normal' gene. **Figure 6- 7** shows the scatter plots of PRT-based estimated copy numbers for the *AMY1* gene against the LRR and RR of rs425970 SNP for 68 participants from the NutriTech cohort. The correlation analysis revealed that the PRT-estimated copy numbers were significantly highly correlated with the LRR ($p=1.96 \times 10^{-11}$, $r=0.701$) and mean RR ($p=1.70 \times 10^{-11}$, $r=0.703$) of rs425970. This shows that the *AMY1* copy numbers estimated from the PRT-based methods match well with the signal intensity information from the DNA arrays used in this study, further highlighting the accurate estimation of copy numbers using the PRT-based methods.

### 6.5.6. Association analysis between signal intensity data and BMI

The investigation was extended by analysing the signal intensity ratio of 35 SNPs that were located along the ~500-bp region containing all known human amylase genes (salivary *AMY1*, pancreatic *AMY2A* and pancreatic *AMY2B* genes). The associative relationship between the signal intensity data and BMI was assessed by performing a linear regression test in a SNP-by-SNP approach. Among the 35 SNPs in the amylase gene cluster region, only 23 of them were located on either *AMY1*, *AMY2A* or *AMY2B* genes. As expected, the overall effect size distribution for the 23 SNPs was insignificant, and no significant association detected between any of the SNPs with BMI **(Figure 6- 8)**.

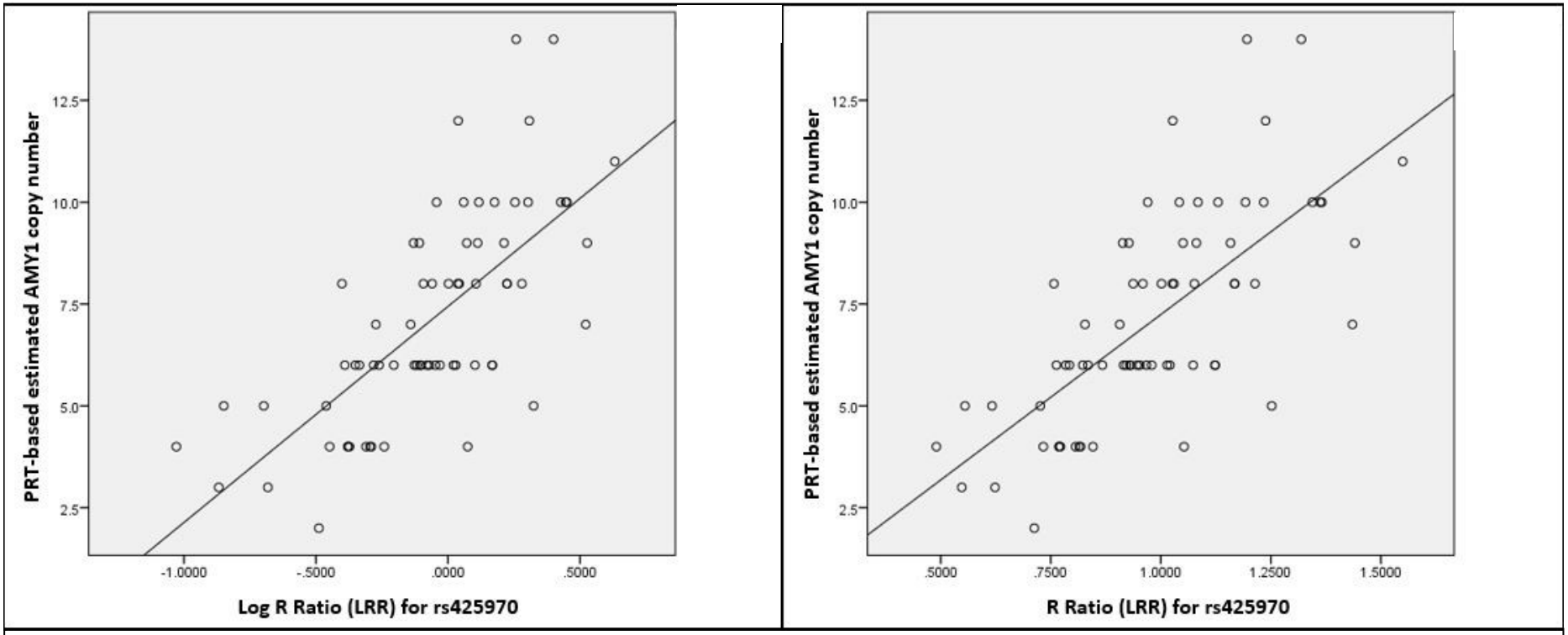


**Figure 6- 7:** Scatter plots of the PRT-based estimated AMY1 copy numbers and signal intensity data of rs425970 SNP (left panel: Log R Ratio; right panel: R Ratio) for the participants in the NutriTech cohort (n=68). Subsequent correlation analysis revealed significant correlations of the AMY1 copy numbers estimated from PRT-based methods with both signal intensity data.

### 6.5.7. *AMY1* and *AMY2A* copy number correlations

Earlier in the **Introduction** section, I have stated that the DNA sequences of *AMY1* and *AMY2A* genes have been reported to be highly similar(Carpenter *et al.*, 2015), emphasising the importance of accurate CNV genotyping of these genes. Thus, the distributions of copy numbers between *AMY1* and *AMY2A* genes were assessed, including the even-to-odd copy number relationship of the two genes. **Figure 6- 9** shows the distributions of PRT-estimated copy numbers for *AMY1* and *AMY2A* genes. Overall, the majority of NutriTech participants (76.8%) have even numbers of *AMY1* diploid copies, consistent with most amylase haplotypes reported in other studies (Armour *et al.*, 2007, Carpenter *et al.*, 2015, Rukh *et al.*, 2017, Yong *et al.*, 2016). By assessing the distributions of copy numbers in both genes, I found that if the participants do not have any variation in their *AMY2A* copies (assuming the normal 2 copies), the total *AMY1* copy numbers will be even-numbered. This shows that the copy numbers of *AMY1* and *AMY2A* are highly correlated ($P=2.92 \times 10^{-25}$), as reported previously (Carpenter *et al.*, 2015, Yong *et al.*, 2016).

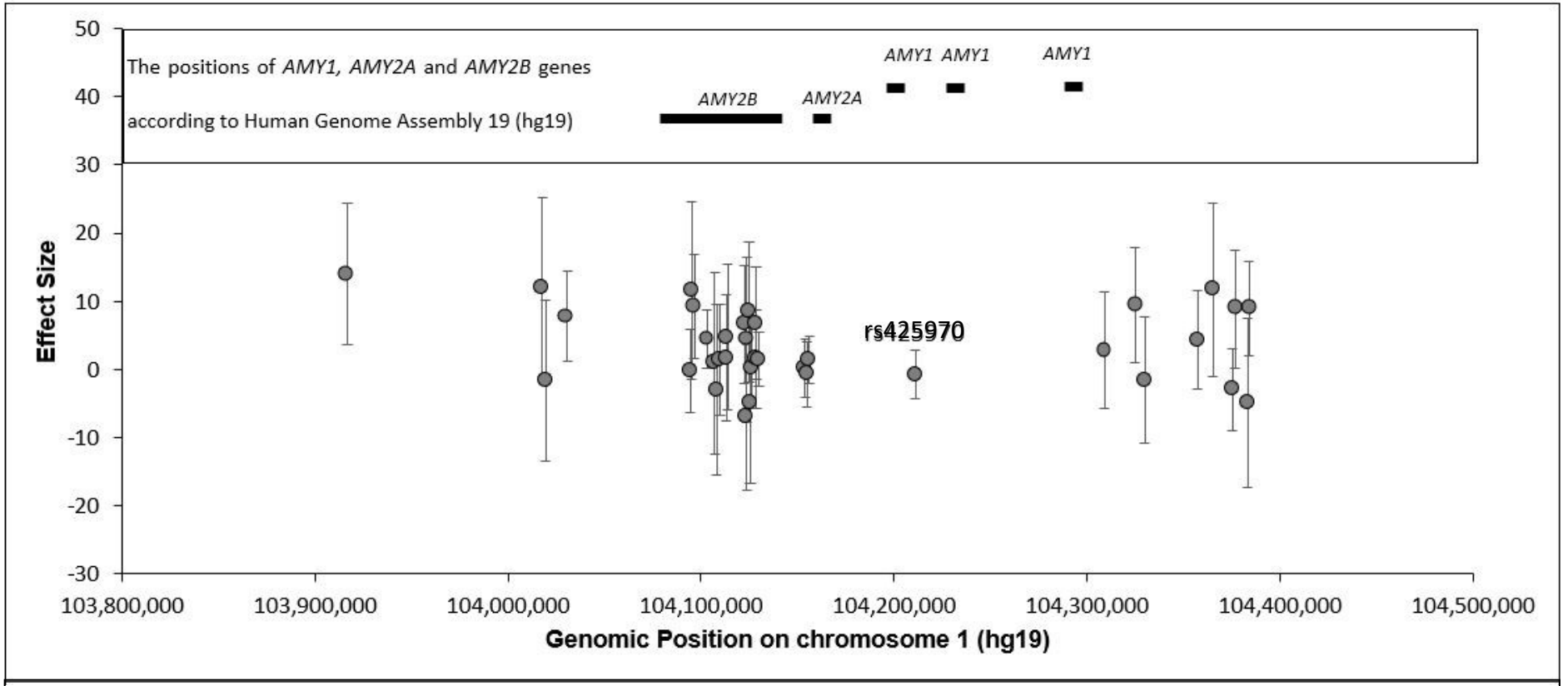


**Figure 6- 8:** Scatter plot of the individual effect size of each SNP with BMI, resulted from the linear regression tests of signal intensity ratio (R Ratio) for all 35 SNPs located within the 500-bp region containing *AMY1*, *AMY2A* and *AMY2B* genes on chromosome 1. No significant association was detected between any of the SNPs with BMI, even before any multiple testing correction. Error bars = 95% confidence intervals of the means (CI).

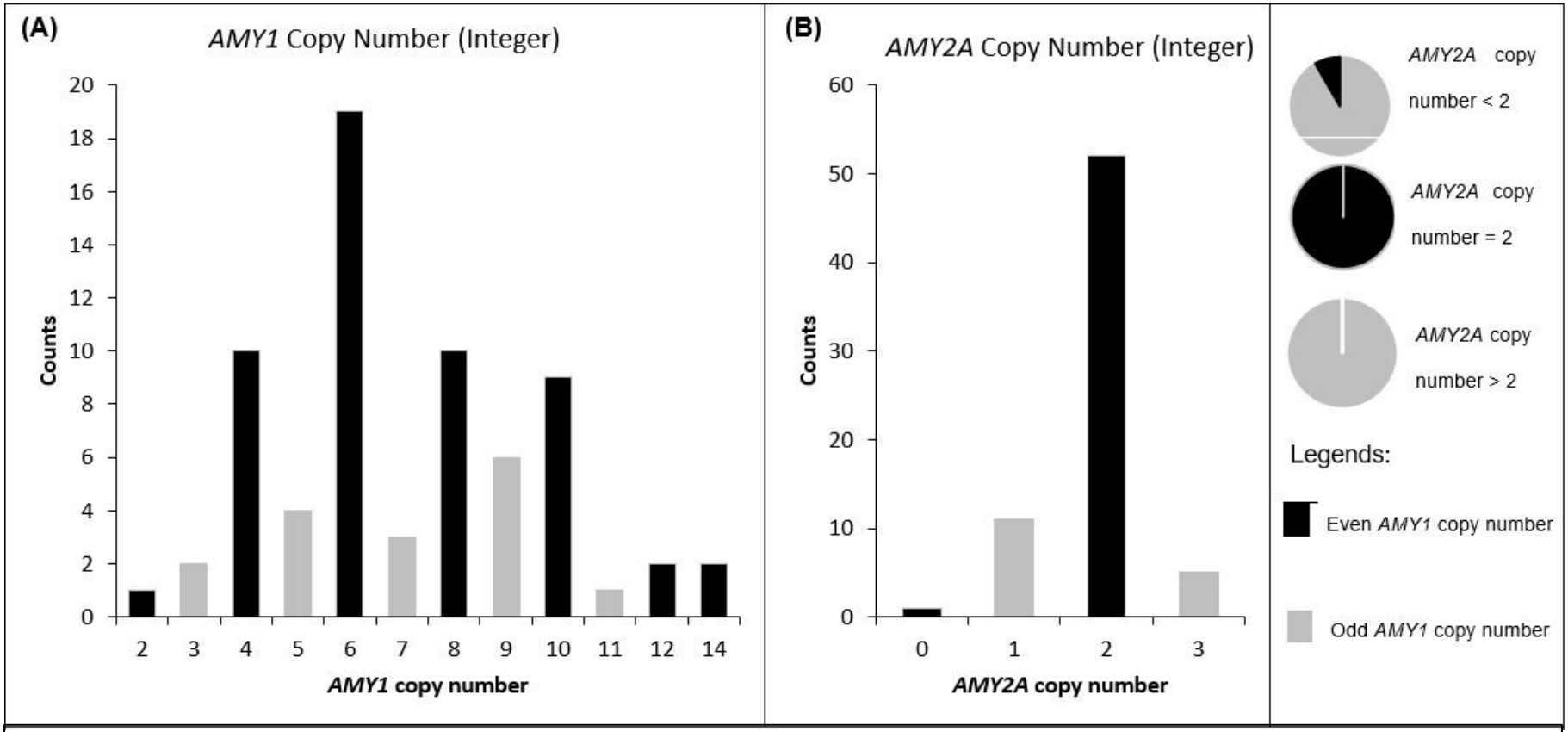


**Figure 6- 9:** The estimated PRT-based copy numbers of **(A)** *AMY1* gene, and **(B)** *AMY2A* gene, for 68 unrelated samples from the NutriTech cohort, with pie charts showing the proportion of even-to-odd *AMY1* copy numbers in relation to *AMY2A* copy numbers. These demonstrated high significant correlation between even *AMY1* copy numbers with *AMY2A* copy number of 2 ($p = 2.92 \times 10^{-25}$). PRT = Paralogue Ratio Test (Armour *et al.*, 2007).

# 6.6. Discussions

## 6.6.1. Attempted replication of previously reported association of BMI

This chapter investigates the relationship between the variation in *AMY1* copy numbers (measured by PRT-based methods), and obesity-related phenotypes (BMI, WC, WHR) as well as various fat measurements in a small group of middle-aged, overweight-to-obese participants from the NutriTech cohort. Initial analysis shows that there was, as expected, a significant difference between the sexes in the means for the fat measurements, but not for BMI. Thus, sex along with age were used as covariates where appropriate, while downstream analyses were done separately according to sex.

The discovery of the association between the salivary *AMY1* copy numbers and adiposity in 179 Swedish families was based on the relationship between the signal intensity data (LRR and BAF) of one probe from DNA arrays and adiposity (BMI and fat mass) (Falchi *et al.*, 2014). Furthermore, by estimating the *AMY1* copy numbers using qPCR methods, they even reported that *AMY1* copy number has a stronger effect on BMI, than do the strongest obesity-associated GWAS-identified SNPs reported to date; rs1121980 and rs9939609 in *FTO* gene ($P < 8 \times 10^{-8}$) (Frayling *et al.*, 2007, Hinney *et al.*, 2007). Falchi *et. al.* (2014) suggested that having lower *AMY1* copy numbers would explain the increase in BMI (BMI change per copy = -0.15, $P = 6.93 \times 10^{-10}$), and in obesity risk (odds ratio per copy = 1.19, $P = 1.46 \times 10^{-10}$) in adults (Falchi *et al.*, 2014).

Here, I have demonstrated a non-normal distribution of PRT-based estimated *AMY1* copy numbers, with differences in the number of even-to-odd copies of *AMY1* gene. The *AMY1*

copy number distribution pattern is similar, and in agreement with the patterns reported by Carpenter, *et al.* (2015), Usher, *et al.* (2015), and Yong, *et al.* (2016) (Carpenter *et al.*, 2015, Usher *et al.*, 2015, Yong *et al.*, 2016), but contradicts with the approximately normal distribution reported by Falchi, *et al.* (2014) (Falchi *et al.*, 2014). To be able to make a comparable assessment, I attempted to replicate the findings by Falchi M., *et. al.* (2014) (Falchi *et al.*, 2014) by first conducting linear regression tests between the LRR, BAF and RR of all SNPs (n=23) located within the amylase gene cluster region – no significant association was observed for all the 23 SNPs with BMI. Additionally, the same assessment was carried out in the Northern Finnish population, in which no significant association was detected. Then the *AMY1* copy numbers estimated from the PRT-based methods developed by Armour J. A., *et. al*. (2007) (Armour *et al.*, 2007) were linearly regressed with BMI, WC, WHR and various adiposity content measurements to assess if there is any association as reported. As expected, I did not find any significant associations between *AMY1* copy numbers and the obesity-related phenotypes, either in females, males, or in the whole cohort. Similar negative associations with BMI were also reported by Carpenter, *et al.* (2015), Usher, *et al.* (2015), and Yong, *et al.* (2016) (Carpenter *et al.*, 2015, Usher *et al.*, 2015, Yong *et al.*, 2016).

In this chapter, the *AMY1* copy number investigation was extended to identify its relationship with various body fat measurements; total abdominal fat %, subcutaneous abdominal fat %, visceral fat % and pancreatic fat content. Overall, I have demonstrated that females had significantly lower fat contents than males. However, no significant association of *AMY1* copy number with any of the fat measurements was found either in males, females, or in the whole cohort. The *AMY1* copy number is highly correlated with *AMY1* gene expression in saliva (Yang *et al.*, 2015). Other than Falchi *et. al.* (2014), a study conducted by Viljakainen, *et al.* (2015) demonstrated a significant correlation between *AMY1* copy number

and total body fat % in females (r = 0.512, P = 0.013) (Viljakainen *et al.*, 2015), although the methodological issue in their standard dPCR-based genotyping could be the underlying issue in estimating their *AMY1* copy numbers.

Several studies have reported significant associations between *AMY1* copy number and BMI in different populations (Falchi *et al.*, 2014, Marcovecchio *et al.*, 2016, Mejia-Benitez *et al.*, 2015, Viljakainen *et al.*, 2015). The positive association found by Falchi, *et al.* (2014) was replicated in Mexican children where they also observed significant inverse relationship between *AMY1* copy numbers and BMI ($p=4.25 \times 10^{-6}$) (Mejia-Benitez *et al.*, 2015). Mejia-Benitez, *et al.* (2015) carried out their investigation based on *AMY1* genotyping without any indication of enzyme digestion prior to the amplification process (Mejia-Benitez *et al.*, 2015). This could underestimate the copy numbers altogether, due to high co-segregating nature that could happen between the tandem 27kb-repeats of *AMY1* gene (Yong *et al.*, 2016). A recent study conducted on Italian children demonstrated a similar, yet lower significant association of *AMY1* copy numbers with BMI in boys (BMI z-score p=0.04) (Marcovecchio *et al.*, 2016). This study used conventional qPCR to quantify *AMY1* copy number, potentially providing false-positive associations due to systematic genotyping assay biases and miscalibration (Carpenter *et al.*, 2015). Furthermore, these studies were heavily confounded either by the majority of samples causing a shift in *AMY1* copy number distribution, or by the controls who had high and inconsistent *AMY1* copy numbers as outliers (Falchi *et al.*, 2014, Mejia-Benitez *et al.*, 2015, Yong *et al.*, 2016). For instance, Falchi, *et al.* (2014) designed their qPCR-based *AMY1* genotyping around two HapMap samples as controls; NA18972 and NA18596. The *AMY1* copy number for NA18972 is inconsistent across different platforms used in several studies, as they have found it to have either 14, 16, or 18 copies (Carpenter *et*

*al.*, 2015, Falchi *et al.*, 2014, Yong *et al.*, 2016). On the other hand, NA18596 has been known to contain mosaicism – a common finding in cell-lines (Yong *et al.*, 2016).

In order to address the unreliability of qPCR assays in detecting even-numbered *AMY1* copy number, Yong, *et al.* (2016) replaced both controls with NA10472, and used it to calibrate the copy number of a common, known region in NA10851 (a standard control used for microarray) (Yong *et al.*, 2016). Their assay demonstrated six *AMY1* gene copies for both samples, consistent with previous reports (Perry *et al.*, 2007, Usher *et al.*, 2015, Yong *et al.*, 2016). This shows that the results of Falchi, *et al.* (2014) may have been, in part, undermined by both qPCR unreliability in detecting clusters, and inconsistency of the controls used. As extensively discussed by Carpenter *et al.* (2015), qPCR calibration is essential to avoid failure detection of even-numbered copy number clusters, since the qPCR assay has a high risk of miscalibrations (Carpenter *et al.*, 2015). Furthermore, the qPCR assay is less reliable in detecting single unit repeat differences in multi-allelic CNV regions.

### 6.6.2. CNV Genotyping of amylase gene region

In this study, the genomic DNA were sent to the University of Nottingham, for *AMY1* CNV genotyping using a combination of paralogue ratio tests (PRTs) and microsatellite analysis, along with fiber-FISH and sequence read depth, developed by Armour A. J., *et. al.* (2007) to measure *AMY1* copy numbers (Carpenter *et al.*, 2015, Armour *et al.*, 2007). Briefly, the combined methods took advantage of the presence of an endogenous retrovirus (ERV) sequence locus on chromosome 12 to be used as a 2-copy calibration for *AMY1* copy number on chromosome 1. The ERV locus is also found upstream of *AMY1* gene. By amplifying both ERV loci on different chromosomes, the calibration provided us with a high-resolution genotyping of *AMY1* copy number without any influence from either *AMY2A* or *AMY2B* copy

numbers. By using the PRT-based methods, they managed to address the *AMY1-AMY2A* high correlation issue prior to validation using read-depth data and fibre-FISH (Carpenter *et al.*, 2015). Along with having similar even-to-odd differential distribution for *AMY1* copy number, I also demonstrated a high significant association between even-numbered *AMY1* copy numbers and having 2 copies of the *AMY2A* gene, consistent with reports by Usher, *et al.* (2015) and Yong, *et al.* (2016) (Usher *et al.*, 2015, Yong *et al.*, 2016).

The use of digital PCR (dPCR) has been suggested to provide a sensitive measurement for highly polymorphic CNVs (Falchi *et al.*, 2014, Mejia-Benitez *et al.*, 2015, Whale *et al.*, 2012). Assays using dPCR systematically count single molecules that are amplified, hence quantifying the gene dosage in a linear nature (Baker, 2012, Hindson *et al.*, 2013, Weaver *et al.*, 2010, Yong *et al.*, 2016). Viljakainen *et al.* (2015), using a standard dPCR assay, reported *AMY1* copy number correlation with BMI, suggesting that low *AMY1* copy number is associated with early-onset obesity in Finnish adolescent females (n<150, p=0.025) (Viljakainen *et al.*, 2015). However, human *AMY1* region CNVs are rather complicated to resolve due to the presence of pancreatic *AMY2A* and *AMY2B* genes at the telomeric end of the human amylase gene cluster. The high sequence identity between *AMY1* and *AMY2A* (93.2%), and *AMY2B* (93.6%) might have contributed towards increased inaccuracy in quantifying *AMY1* copy number (Carpenter *et al.*, 2015). Nonetheless, Usher *et al.* (2015), using a modified dPCR, reported no association of *AMY1* copy number with BMI and obesity in a well=powered study (N~3,500) (Usher *et al.*, 2015). To remove copy number bias in replicating cells, they used a control probe that targeted a region just outside the amylase region, instead of using the standard control probe targeting the *RPP30* gene. This provided a higher resolution of CNV typing in highly polymorphic region such as the amylase region. Thus, I believe that the use of a standard dPCR assay (and small sample size) is still insufficient to provide the necessary

sensitivity for measurement of *AMY1* copy number, and careful attention must be given to validating the accuracy of CNV typing outputs.

### 6.6.3. Looking at starch consumption

I did not find any significant association between *AMY1* copy numbers (as a continuous variable, as tertiles) and starch intake. However, it is worth mentioning that participants in the lowest tertile of AMY1 copy numbers, on average, had the highest starch intake. Arguments have been previously made regarding the relationship between human *AMY1* copy number and dietary starch intake, with Perry *et. al.* (2007) suggesting that higher *AMY1* copy numbers were more prevalent in populations with high-starch diets compared to populations with low-starch diets (Perry *et al.*, 2007). There is a belief that the number of *AMY1* gene copies has been expanded throughout human history. Chimpanzees, which are on low-starch diets, consistently have two *AMY1* gene copies in their genome (Yong *et al.*, 2016). These studies postulated that having high dietary intake of starch may be related to higher *AMY1* copy numbers, or vice versa. However, comparisons between populations with different dietary starch intakes were not conducted in this chapter due to unavailability of a study cohort with contrasting starch diets.

In this study, I observed that participants in the lowest *AMY1* copy number tertile had, on average, the highest dietary starch intake. This could be due to several possibilities; 1) the reduced amount of salivary amylase in our mouth contribute to the reduced sensation of starch digestion in the oral cavity and thus creates a perception that alters the behaviour to consume more starchy foodstuffs, or 2) the reduced amount of salivary amylase digests consumed starch food at a much slower rate, thereby prolonging the whole digestive process while slowing down the release of sugar from the consumed starch food. Such perceived

deprivation could contribute towards increasing the intake of dietary starch in the long run. However, further investigations must be conducted to address the two possibilities.

A previous study has shown that dietary starch intake could modify the overall relationship between *AMY1* copy number and BMI (Rukh *et al.*, 2017). Along with the non-significant association of *AMY1* copy numbers on BMI, the author found that increasing *AMY1* copy number significantly decreases the BMI in people with low-starch diets, whereas in people with high-starch diets, that was not the case (Rukh *et al.*, 2017). The study revealed that the BMI has the tendency to be increased with increasing *AMY1* copy number in people with high-starch diets (Rukh *et al.*, 2017). However, in NutriTech participants with high starch diets, increasing *AMY1* copy numbers significantly correlated with decreasing BMI, WC and WHR ($r>-0.4$, $p<0.02$). However, we did not find any significant correlation of *AMY1* copy number with any of the obesity-related phenotypes and fat measurements in participants with low dietary starch intake. Thus, in our dataset, different dietary starch consumption behaviours resulted in a different *AMY1* copy number-BMI relationship. Therefore, our results support the notion that dietary starch intake is important in modifying the relationship between *AMY1* copy number and BMI (Rukh *et al.*, 2017).

Overall, *AMY1* copy number is not associated with obesity-related phenotypes (primarily BMI and WHR), nor with various body fat measurements (total fat, subcutaneous fat, visceral fat, and pancreatic fat) in the NutriTech cohort. I also postulate that the differences in the dietary starch intake may play a significant role in modifying the relationship between salivary amylase copy numbers and BMI.

# CHAPTER 7. Rare Deletions in The Chromosome 16p11.2 Region

## 7.1. Introduction

This chapter describes a comprehensive study identifying previously reported, recurrent obesity-associated deletions at two non-overlapping genomic locations within the 16p11.2 region. The deletions were statistically investigated to identify their associations with two obesity-related phenotypes, BMI and WHR, as well as their effects on the risks of obesity. The investigation was carried out in three types of study cohorts; a birth cohort of North Finland Birth Cohort 1966 (NFBC1966), an obesity cohort of Personalised Medicine for Morbid Obesity (PMMO), and a general population cohort (UK Biobank; UKBB).

The first deletion variant investigated in this study was a heterozygous deletion of ~600 kilobases (kb) in length (minimum length ~520 kb), spanning the genomic region located at 29.6 megabase (Mb) to 31.2 Mb position on the short arm of chromosome 16 (hg19). The putative deletion region is located between two segmental duplication spots (i.e. between breakpoint 4 (BP4) and breakpoint 5 (BP5); OMIM #611913) of the 16p11.2 chromosomal region, and contains more than 30 genes. The putative region of the ~600-kb variant is depicted in **Figure 7- 1** (red shade). In this study, the ~600 kb deletion variant is referred to as the proximal variant because it is located closer to the centromere of chromosome 16, as compared to the other deletion variant (distal), which will be described later in this section. The proximal deletion variant was initially identified as a copy number variant (CNV) associated with autism (Kumar *et al.*, 2008, Marshall *et al.*, 2008, Weiss *et al.*, 2008) and mental retardation (Bijlsma *et al.*, 2009), but further studies, including those carried out by members of Professor Blakemore's research group, have shown that it also had significant

associative relationships with BMI, and obesity in general populations (Maillard *et al.*, 2016, Walters *et al.*, 2010, Yu *et al.*, 2011).

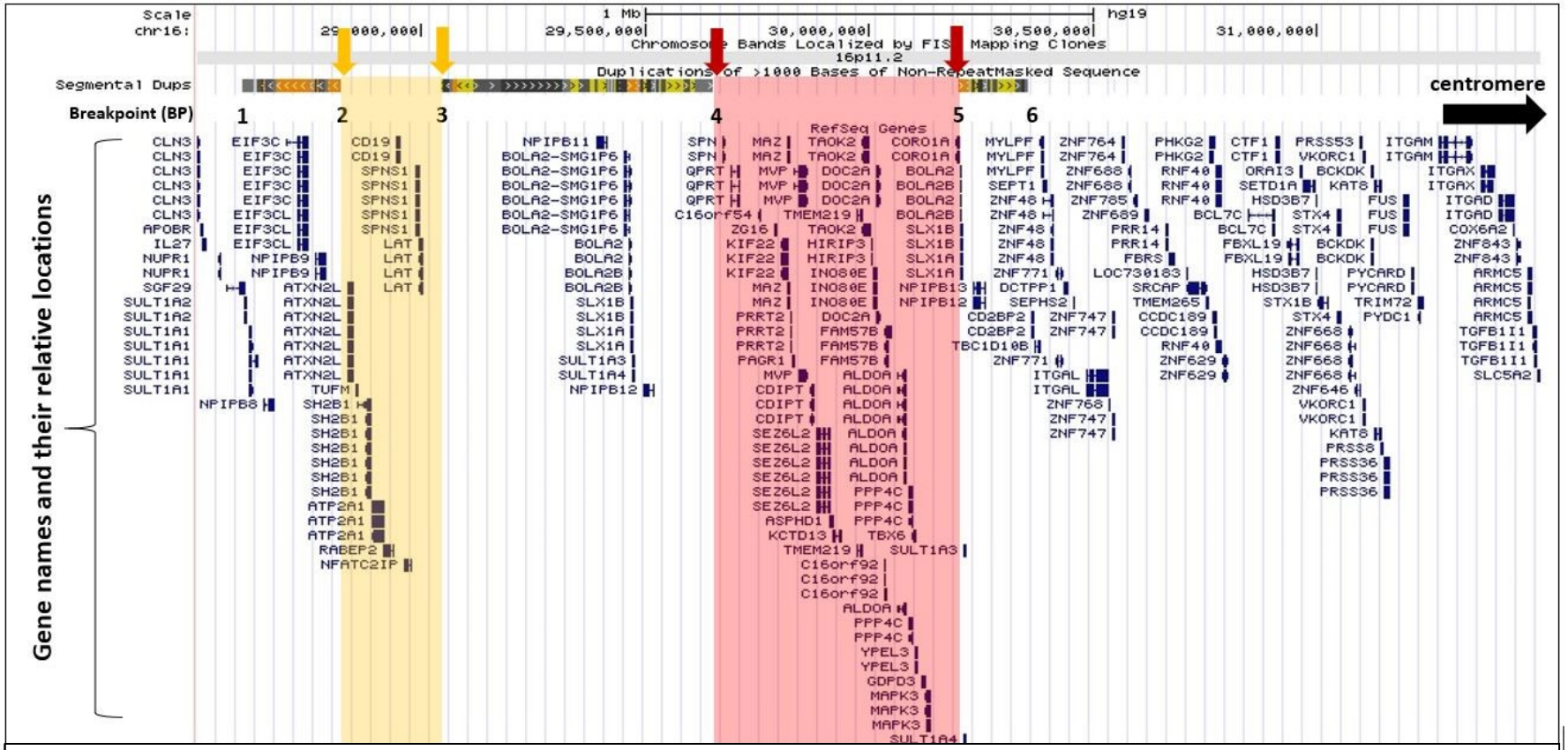


**Figure 7- 1:** The genomic locations of proximal and distal 16p11.2 deletion variants on the 16p11.2 region according to the human genome assembly 19 (GRCh37/hg19). The proximal deletion region (~600 kb) is shaded in red, located between breakpoints 4 and 5 (red arrows). The distal deletion region (~200 kb) is shaded in yellow, located between breakpoints 2 and 3 (yellow arrows). Segmental Dups = segmental duplications. Figure was adapted from Database of Genomic Variants (DGV) website.

The contribution of the recurrent proximal 16p11.2 deletion to disease susceptibility was reported by Sebat, *et. al.* (2008), who noted that heterozygous deletion increases the risk for autism spectrum disorder (ASD). Similar findings were subsequently reported by other groups (Kumar *et al.*, 2008, Marshall *et al.*, 2008, Weiss *et al.*, 2008), providing a strong evidence of its significant association with ASD. Other than ASD, the proximal 16p11.2 deletion has also been associated with developmental delay (DD) (Fernandez *et al.*, 2010, Shinawi *et al.*, 2010), language difficulties, intellectual disabilities (ID) and macrocephaly (large head) (Rosenfeld *et al.*, 2010).

Interestingly, there is wider evidence of strong correlations between developmental/cognitive disabilities in general and the prevalence of obesity (Boeka and Lokken, 2008, Chen *et al.*, 2010). A recent meta-analysis revealed that obesity is associated with ASD (Zheng *et al.*, 2017), adding strength to the notion that obesity could be a co-morbidity of developmental/cognitive disabilities, or vice versa. Walters, *et. al.* (2010) reported the proximal 16p11.2 deletion as a significant risk factor for obesity and is significantly associated with BMI (Walters *et al.*, 2010). The author also reported that the significant relationship was still true among obese patients without ASD, DD or ID (Walters *et al.*, 2010). In addition, the ~600-kb heterozygous deletion was observed to be frequently co-segregated with severe early-onset obesity (Bochukova *et al.*, 2010, Walters *et al.*, 2010).

The second deletion variant investigated in this study was a relatively smaller heterozygous deletion (~200 kb) located at 28.7 Mb to 29.1 Mb of the same 16p11.2 region (hg19). The distal deletion variant is located between BP2 and BP3, upstream of the proximal deletion region, as shown in **Figure 7- 1** (yellow shade). This distal deletion region encompasses several genes, including *SH2B1* gene. The *SH2B1* gene encodes the Src homology 2B (SH2B) adaptor protein 1 that binds various protein kinases (including leptin and

insulin receptors), thereby mediating the regulation of glucose homeostasis, energy balance and insulin sensitivity (Ren *et al.*, 2007). The *SH2B1* protein has been shown to positively regulate the leptin-melanocortin pathway, which is one of the important pathways controlling food intake behaviour (Aerts *et al.*, 2015). In mice, deletion of the *sh2b1* gene resulted in several metabolic conditions, including hyperphagia (excessive eating behaviour), obesity and insulin resistance (Ren *et al.*, 2007). This has led researchers to believe that the *SH2B1* gene might have a significant role in the regulation of body weight and the development of diabetes, via the leptin signalling (Aerts *et al.*, 2015, Ren *et al.*, 2007).

The contribution of the distal, *SH2B1*-containing 16p11.2 deletion variant to obesity was first reported by Bochukova, *et. al.* (2010) (Bochukova *et al.*, 2010): they found that the ~200-kb deletion was significantly enriched in morbidly obese children (n=300), with and without DD (Bochukova *et al.*, 2010). Another study that was conducted on >20,000 patients with abnormal phenotypes (i.e. DD and ID), also revealed significant enrichment of the distal 16p11.2 deletion in the obese patients (Bachmann-Gagescu *et al.*, 2010). In contrast to the proximal 16p11.2 deletion, studies on the *SH2B1*-containing deletion in obesity are relatively scarce. To date, there are still no reports on the association of the distal ~200-kb deletion with obesity-related anthropometric traits, such as BMI and WHR. Additionally, there are rare instances of patients with larger deletions covering both the distal and proximal 16p11.2 deletion regions, as discovered by Bochukova, *et. al.* (2010) (Bochukova *et al.*, 2010).

Overall, the above-mentioned reports have paved the way for further research investigating the relationships between the 16p11.2 deletions and obesity, as well as other obesity-related phenotypes (i.e. BMI and WHR). In this study, I explored the relationships of both 16p11.2 deletion variants with BMI and WHR, as well as investigating the overall

contribution of the same deletions on the risk of obesity in different types of study cohorts; a birth cohort, an obesity-ascertained patient cohort, and a general population cohort.

## 7.2. Aim of this chapter

The overall aim of this chapter is to identify the presence of proximal and distal 16p11.2 heterozygous deletions using a combination of methods, including interrogation of CNV calls predicted by a CNV calling algorithm tool (from **Chapter 5**), visual inspection of signal intensity plots, and run of homozygosity (ROH) analysis. The aim was addressed by the following questions:

1. *Are there any predicted deletions within the 16p11.2 region (copy number<2) in the clean CNV catalogue of each study cohort involved in this chapter?*

I answered this question by interrogating the CNV catalogues for the NFBC1966 and PMMO cohorts (compiled from **Chapter 5**). Any individuals with predicted deletions overlapping the distal (28.7-29.1 Mb) and proximal (29.6-31.2 Mb) 16p11.2 region, were flagged as potential carriers.

For the UKBB, the large size of the cohort necessitated a different strategy: the chromosome 16p11.2 region was first screened for a run of homozygosity (ROH) using PLINK tool. This could indicate either true homozygosity (assuming that there is a strongly conserved haplotype structure in this region), or hemizygosity resulting from a deletion. All individuals identified in this manner were subjected to visual inspection of intensity plots as described below.

2. *Are the predicted deletions identified from the CNV catalogues truly present in the individuals flagged as potential carriers?*

I answered this question by visually inspecting the Log R Ratio (LRR) and B Allele Frequency (BAF) scatter plots spanning both putative deleted areas of the 16p11.2 region.

3. *Upon confirmation of the presence of the deletion variants, do any of the deletion variants associate with obesity-related phenotypes (i.e. BMI and WHR)?*

I answered this question by evaluating the relationships between each 16p11.2 deletion variant and the obesity-related phenotypes using statistical comparisons and regression analyses.

4. *Does the 16p11.2 deletion variant have a significant effect on the risks of obesity and morbid obesity?*

I answered this question by designing a case/control association analysis that comprised of PMMO cohort as obese patient population (cases) and the UKBB cohort as a general control population. I assessed the associative relationships between each 16p11.2 deletion variant and the risk of obesity/morbid obesity using Fisher's exact tests.

## 7.3. General methods used in this chapter

| Section | Method name (Described in Chapter 3) | Chapter 4 | Chapter 5 | Chapter 6 | Chapter 7 | Chapter 8 |
|---|---|---|---|---|---|---|
| **3.2** | Biological sample collection | ✓ | ✓ | ✓ | ✓ | ✓ |
| **3.3** | Genomic DNA extraction | ✓ | ✓ | ✓ | ✓ | ✓ |
| **3.4** | Obesity-related phenotypic measurements | ✓ | ✓ | ✓ | ✓ | ✓ |
| **3.5** | SNP genotyping | ✓ | ✓ | ✓ | ✓ | ✓ |
| **3.6** | Quality controls | ✓ | ✓ | ✓ | ✓ | ✓ |
| **3.7** | Population structure adjustments | ✓ | ✓ | ✓ | ✓ | ✓ |
| **3.8** | Principal component analysis | | | | | ✓ |
| **3.9** | CNV calling | | ✓ | | ✓ | ✓ |
| **3.10** | CNV quality controls | | ✓ | | ✓ | ✓ |

## 7.4. Specific methods

Specific methods were conducted to address the aim of this study chapter. The specific methods are described below.

### 7.4.1. Identification of 16p11.2 deletions from the NFBC CNV catalogues

A cleaned CNV catalogue was prepared for the NFBC1966 and PMMO cohorts (refer to **Chapter 5**). Each catalogue contained high-quality CNV calls that were predicted by the PennCNV (May 2014 version) tool (Wang *et al.*, 2007). Each catalogue was scanned for any deletion calls (copy number < 2) that spanned either the proximal or the distal 16p11.2 region (hg19; chr16:28,000,000-31,000,000). All individuals of such deletion calls were flagged as potential 16p11.2 deletion carriers.

### 7.4.2. Run of homozygosity (ROH) scans on putative 16p11.2 copy number regions

Runs of homozygosity (ROH) were conducted in PLINK tool Version 1.07(Purcell *et al.*, 2007), under the assumption that if a heterozygous deletion is truly present in an individual, all of the consecutive SNPs within the putative deletion regions would be homozygous. Separate ROH tests were performed for the proximal (~600 kb) and distal (~200 kb) 16p11.2 deletion variants in each study cohort. Since both deletion variants have been previously reported as rare (Bochukova *et al.*, 2010, Walters *et al.*, 2010), ROH scan was useful to filter out a majority of individuals in a large cohort, such as in the UKBB cohort. Prior to scanning, the `CleanedCohort` PLINK-binary files were filtered to only include the SNPs on the 16p11.2

region (`CleanedCohort_chr16p11.2`). The command lines for ROH scans were then executed in PLINK, as below:

For proximal 16p11.2 deletion region (~600 kb);

```
./plink  --bfile CleanedCohort_chr16p11.2 –homozyg --homozyg-kb 500 --out
CleanedCohort_16p11.2_proximal_ROH
```

For distal 16p11.2 deletion region (~200 kb);

```
./plink  --bfile CleanedCohort_chr16p11.2 –homozyg --homozyg-kb 200 --out
CleanedCohort_16p11.2_distal_ROH
```

### 7.4.3. Selecting potential 16p11.2 deletion carriers from ROH scans

The output files (.hom) from previous method, contained all homozygosity stretches that were found in individuals along the 16p11.2 region. These homozygosity stretches were checked further to remove all stretches that did not overlap either distal (hg19; chr16:28,500,000-29,100,000) or proximal (hg19; chr16:29,600,000-31,200,000) deletion regions. All individuals who survived the filtering process, were selected as potential 16p11.2 deletion carriers.

### 7.4.4. Signal intensity plots for the 16p11.2 region

For NFBC1966, and PMMO cohorts, 'For-CNV Table' files extracted from the GenomeStudio software (refer **Chapter 3.5.2 Processing of genotyped data**) were used to plot the distributions of Log R ratio (LRR) and B Allele Frequency (BAF) data for all the SNPs located on the 16p11.2 region, spanning both putative deletion regions (GRCh37/hg19; chr16: 28,000,000-31,000,000). The plotting was conducted by running a perl script (available upon

request), which was developed by Robin Walters (The University of Oxford, UK). Briefly, the script uses the R tool (version 3.4.2) (R Core, 2013) to scan all the SNPs within a user-specified region (chr:start bp:end bp format), takes the LRR values for those SNPs and creates a scatter plot in R, with SNP genomic positions on the x-axis. The process was the same for BAF plots.

For the UKBB cohort, the LRR and BAF data were provided in two separate tables. In addition, the individual and SNP lists were also provided to us in separate files. Thus, it was not possible to use the same script to plot the data. Therefore, the perl script was modified by the same developer to change its parameter scanning functions so that it can read different files to obtain different information at once. Further optimisation work and tests were conducted to ensure that the modified perl script functioned properly. Once confirmed, I split each information file (individual, SNP, LRR, BAF files) into 10 smaller files to increase efficiency, reduce the computing power as well as reduce the memory used by the modified script. The columns in LRR and BAF files corresponded to the individuals listed in the individual file, while the rows corresponded to the SNPs listed in the SNP file. The scanning region was the same as in the previous cohorts, spanning both putative 16p11.2 deletion regions (GRCh37/hg19; chr16: 28,000,000-31,000,000).

The output for each study cohort was in a pdf format, with each page containing an LRR plot on top of a BAF plot of the same x-axis (genomic positions). Each page represents each individual from the same study cohort.

### 7.4.5. Visual inspections of signal intensity plots of potential 16p11.2 deletion carriers

The LRR and BAF plots for all potential 16p11.2 deletion carriers were visually inspected for any aberrations of signal intensities (LRR) and allelic signals (BAF) for consecutive SNPs

that covered either deletion regions. If a heterozygous deletion is present, the LRRs for the SNPs within the putative deletion region would be lower than the normalised values (scattered around zero) along the x-axis, and the BAFs for the same SNPs would scatter either around 0 or 1 (which represent homozygosity), but not around 0.5 (which represents heterozygosity). However, if the BAF plot showed likewise, but the LRR plot showed no evidence of a heterozygous deletion, it might suggest that those consecutive SNPs have homozygous genotypes.

### 7.4.6. Statistical analyses

*Obesity classifications:* Normal, obese, and morbidly obese individuals were classified according to the BMI thresholds in adults; BMI<30 kg/m$^2$ (normal), 30 kg/m$^2$≤BMI<40 kg/m$^2$ (obese), BMI≥40 kg/m$^2$.

*Obesity-related phenotype distributions:* The distributions of the residuals of each continuous independent variable such as waist-to-hip ratio (WHR) and body mass index (BMI), were compared to a normal distribution by visual inspection of frequency histograms and normal P-P plots.

*Sex, age and batch effects:* The means of those continuous variables were subsequently compared according to age, sex and batch, using Analysis of Variance (ANOVA) tests. The effects of age, sex and batch differences on obesity status (BMI≥30 kg/m$^2$), were compared by using Pearson's chi-square tests.

*Independence of observation*: Durbin-Watson tests were carried out between each obesity trait (i.e. BMI, WHR) and 16p11.2 deletion status to determine whether the dependent and independent variables were independently observed from one another and to check for presence of auto correlation. Durbin-Watson statistics < 1 was considered as positively auto

correlated, while Durbin-Watson statistics > 3 was considered as negatively auto correlated. A Durbin-Watson statistics around 1.9 to 2 was considered as no significant evidence of auto correlation present.

*Obesity-related phenotype correlations:* Bivariate correlation analyses were performed on all obesity-related phenotypes under study (i.e. WHR and BMI), obesity status, 16p11.2 deletion carrier status, sex, age, and genotyping batch.

*Obesity-related phenotype associations:* Independence of observation was determined by calculating Durbin-Watson statistics, in which a value of 2 would mean that the two variables were independently observed from each other. The means of each obesity trait were compared between the deletion carriers and the non-carriers by using independent t-tests. The association analysis of each 16p11.2 deletion (carrier status as a categorical independent variable), with obesity traits (as continuous dependent variables), were conducted by performing general linear regression tests, adjusted for age, sex and batch. If only a single deletion carrier was present, one-sample independent t-tests were required to compare the obesity traits of the deletion carrier (i.e. WHR and BMI) with the means of such obesity traits from obesity sub-populations within the study cohorts (general, obese, morbidly obese sub-populations). All means, 95% confidence intervals (CI), and standard errors of the means (SEM) were calculated for each test. P-values (adjusted for linear regressions, unadjusted for independent t-tests) < 0.05 were considered as statistically significant. Multiple testing correction was not performed due to insufficient number of groups (carrier status; either have the deletion or do not have the deletion).

*Obesity status associations*: Due to the low frequency of both 16p11.2 deletion variants, the statistical analyses were also carried out using Fisher's exact tests, as recommended by Boeka *et. al.* (2008) (Boeka and Lokken, 2008) for the analyses of rare variants in extreme

phenotypes. Fisher's exact tests were conducted by using contingency tables for the number of deletion carriers versus obesity status in the NFBC1966 and UKBB cohorts. All patients in the PMMO cohort were obese, thus Fisher's exact test was not performed in this cohort. To make an overall analysis of the risk of obesity caused by the presence of each deletion variant, the PMMO (as an obesity patient population) and UKBB (as a general population) cohorts were pooled according to the obesity status and the deletion carrier status. These two cohorts were recruited in the UK, and predominantly of British ancestry (refer to **Chapter 2**). The enrichment of each 16p11.2 deletion variant in the patient population (cases) in comparison to the general population (controls) was investigated. Contingency tables were created for the number of carriers of each 16p11.2 heterozygous deletion (as 0 and 1) versus the obesity status (as 0 and 1) in the pooled dataset. In view of analysing the significance of having each 16p11.2 deletion variant in extreme obesity (morbid obesity defined by BMI>40 $kg/m^2$), contingency tables were also constructed for the number of deletion carriers versus morbid obesity status (as 0 and 1) in the pooled dataset. Odds ratios and 95% CIs were calculated for the tests described above.

All statistical analyses were performed using IBM SPSS Statistics for Windows Version 24.0 (Released 2016; IBM Corp., Armonk, New York. USA) and Rstudio version 1.1.419 (RStudio, Inc., Boston, MA, USA) (RStudio, 2015).

# 7.5. Results

## 7.5.1. Distributions of obesity-related traits

To avoid any violation on the assumptions for the statistical tests in downstream analyses, the overall distributions of BMI and WHR were assessed via visualising the histograms and P-P plots. This step is important to reduce false discoveries and limit any multiple testing problem that could arise due to the violation.

**(a) NFBC1966 cohort**

**Figure 7- 2** shows the distribution of the standardised residuals of BMI and WHR as compared to the normal distribution curve/line in the NFBC1966 cohort. The mean residuals for both phenotypes were zero (BMI residual range=-9.93 to 30.18, WHR residual range=-0.37 to 0.89). The residuals for both phenotypes were normally distributed in this cohort. The residuals, which represent the 'errors' between observed and expected values from the normal distribution, were largely contained close to the zero. Thus, BMI and WHR were treated as normally distributed phenotypes.

**(b) PMMO cohort**

The distribution of standardised residuals for BMI in this obesity cohort were shown in **Figure 7- 3**. The mean residual for BMI was zero (range= -17.80 to 31.13). Both phenotypes had approximately normal distributions. However, the data on WC, hip measurement or WHR were not available for the genotyped patients from the PMMO cohort. The P-P plot (**Figure 7- 3**) depicts the clustering of the standardised residuals along the normal line.

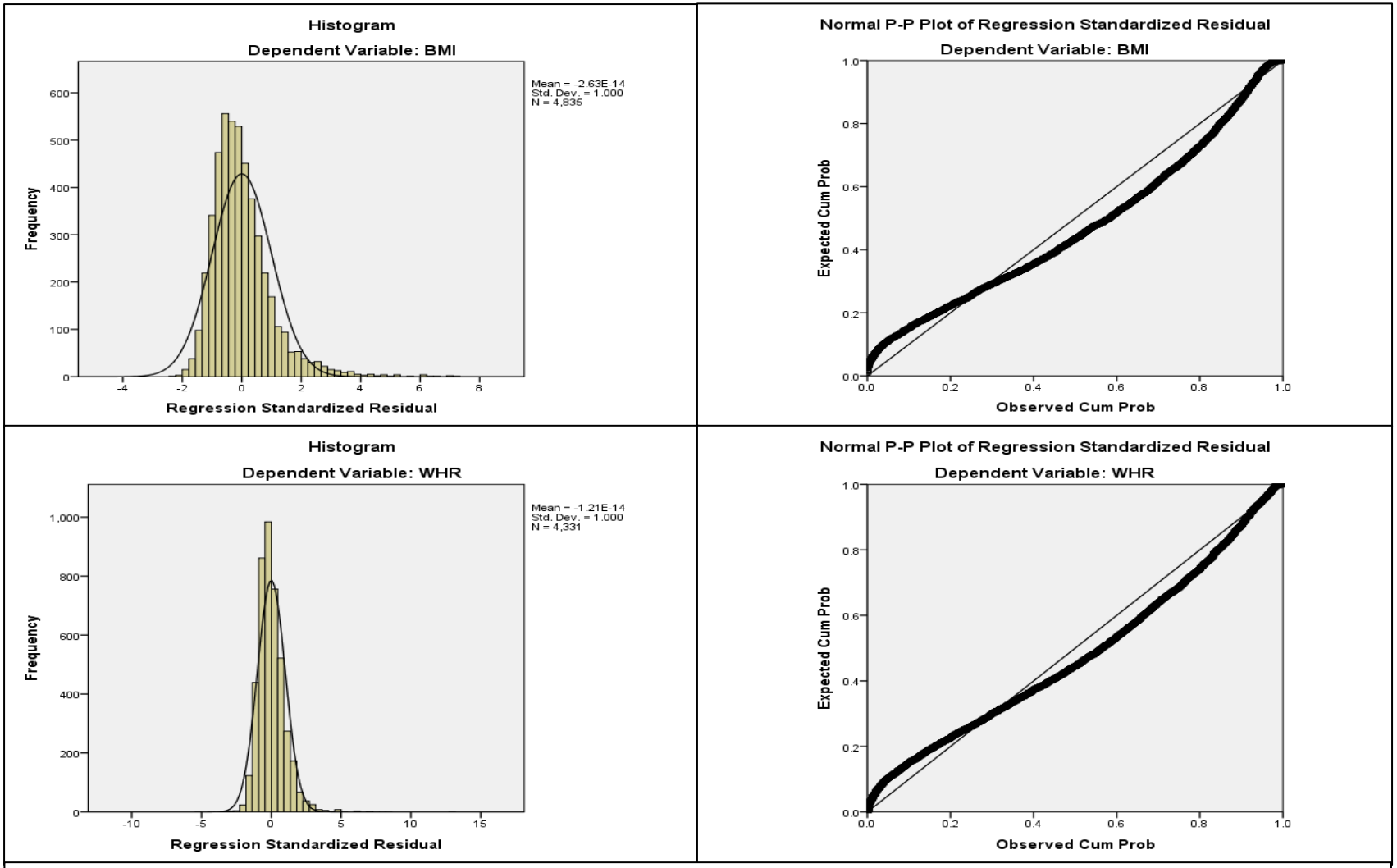


**Figure 7- 2:** The distributions of standardised residuals of obesity traits, BMI (n=4,835) and WHR (n=4,331), in the NFBC1966 cohort.

BMI = body mass index, WHR = waist-to-hip ratio.

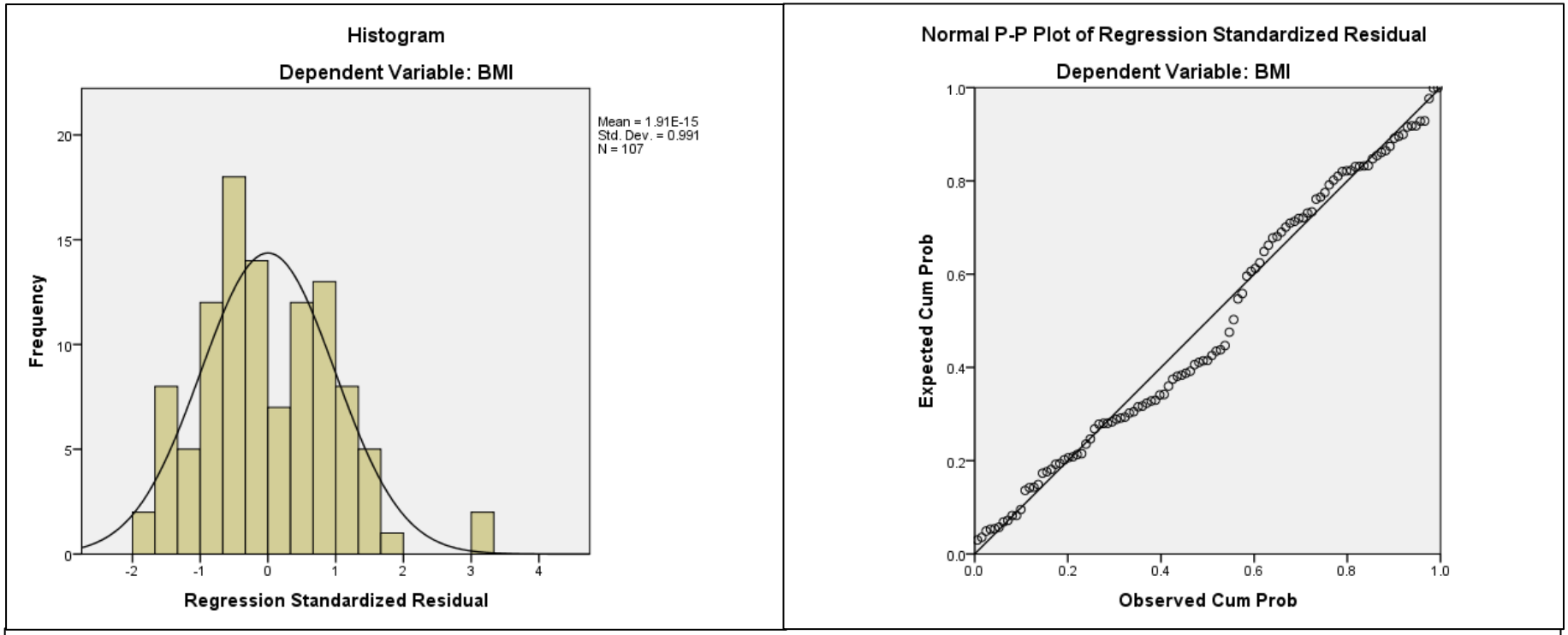


**Figure 7- 3:** The distributions of standardised residuals of obesity traits, BMI (n=107) in the PMMO cohort. BMI = body mass index, WHR = waist-to-hip ratio.

**(c) UKBB cohort**

**Figure 7- 4** displays the overall distribution of the standardised residuals of each obesity trait under study (BMI and WHR), in comparison to the normal distribution in the UKBB cohort. The BMI standardised residuals ranged from -15.28 to 47.81 (mean=0), while the WHR residuals ranged from -0.74 to 2.16 (mean=0). Both obesity-related phenotypes were normally distributed in the UKBB cohort.

### 7.5.2. Sex, age and batch effects

Comparison of means were carried out for each obesity-related phenotype, either by independent t-tests (for sex) or ANOVA (for age and batch). Chi-square tests were also performed for comparisons between obese and non-obese individuals. This step was carried out to assess the effects of age and batch differences on BMI and WHR, and whether males, on average, have significantly different BMI or WHR than females.

**(a) NFBC1966 cohort**

The mean BMI and mean WHR were significantly different in different batches ($p=0.04$ and $p<0.001$ respectively). Males, on average had significantly higher BMI than in females ($p=1.54 \times 10^{-18}$; mean BMI=25.20 kg/m$^2$ in males, mean BMI=24.14 kg/m$^2$ in females). Similarly, the mean WHR for males (WHR=0.91) and females (WHR=0.81) were significantly different ($p<0.0001$). Chi-squared tests revealed that the prevalence of obesity among males was not significantly different than among the females ($p=0.169$). Additionally, a simple binary logistic regression was carried out to identify for any significant relationship between obesity status and genotyping batch. I found that genotyping batch was not a significant predictor for obesity in this cohort. Nonetheless, both variables were used as covariates in

regression analyses. Age was not investigated for this birth cohort since all participants were born in the same year.

**(b) PMMO cohort**

Independent t-tests revealed that on average, males and females had similar BMI (p=0.676). The mean BMI for male patients was 52.15 kg/$m^2$ (range=34.45-81.15) while the female patients had mean BMI of 51.37 kg/$m^2$ (range=35.46-82.63). I also found that the mean BMI was not significantly different between the two genotyping batches (p=0.737). Age was not analysed due to lack of data.

Chi-square tests were carried out to identify for any significant effect of sex and batch between the obese (n=96) and morbidly obese (n=11) patients within this cohort. I found that the prevalence of morbid obesity among obese males was not significantly different than in obese females (Pearson's Chi-squared p=0.214). Similarly, different genotyping batches did not have a significant effect on the obesity/morbid obesity status of the patients in the PMMO cohort (Pearson's Chi-squared p=0.515). Although non-significant, sex and batch were included as covariates in association analyses.

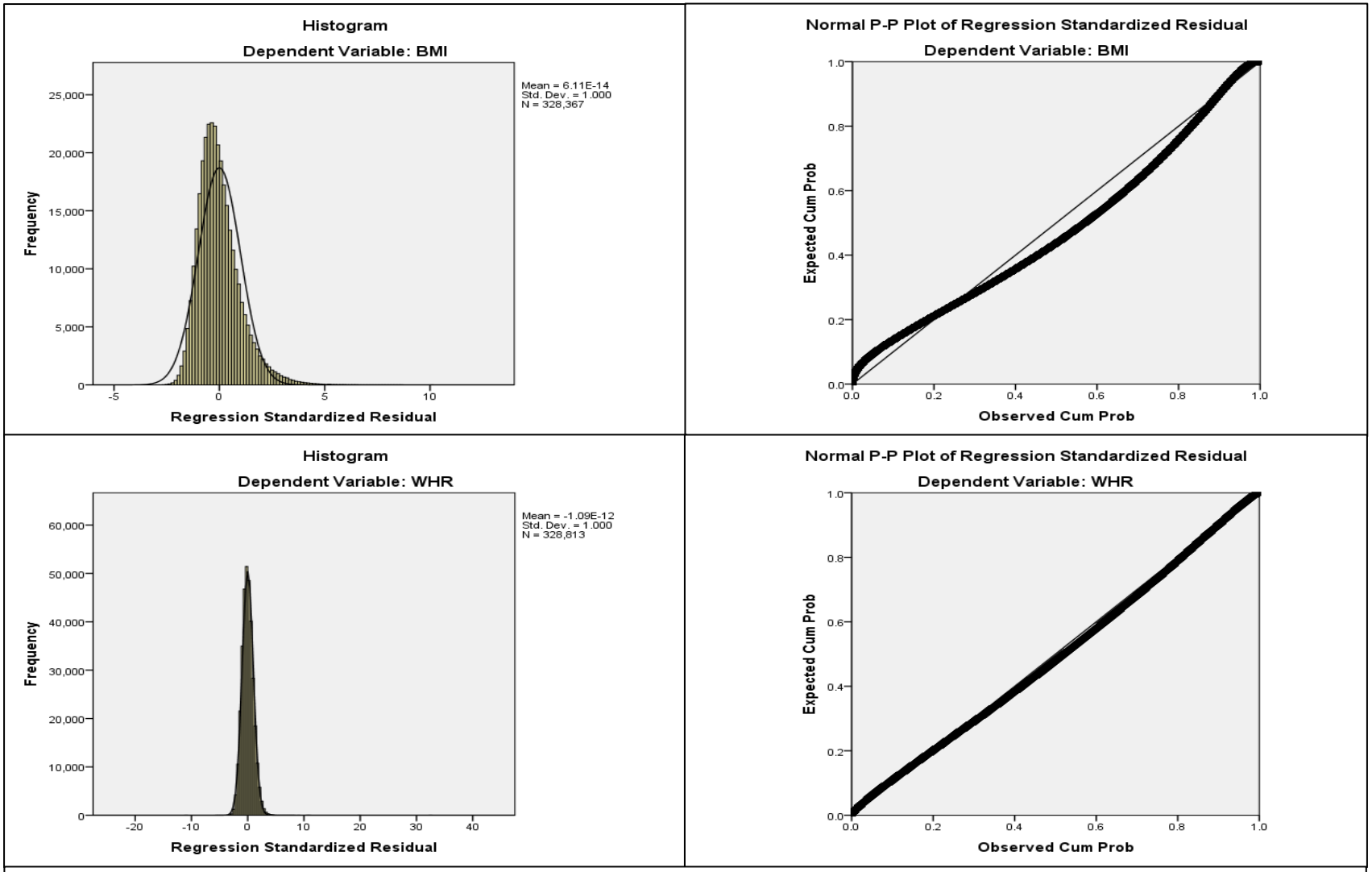


**Figure 7- 4:** The distributions of standardised residuals of obesity traits, BMI (n=328,367) and WHR (n=328,813), in the UK Biobank cohort. BMI = body mass index, WHR = waist-to-hip ratio.

**(c) UKBB cohort**

The mean BMI and mean WHR were significantly different in different batches ($p=3.63 \times 10^{-94}$ and $p=4.16 \times 10^{-211}$ respectively). On average, males (BMI=27.82 kg/m$^2$, WHR=0.94) had significantly higher BMI and WHR compared to females (BMI=27.01 kg/m$^2$, WHR=0.82), with $p < 1.00 \times 10^{-17}$. In terms of age difference, the mean BMI and mean WHR were significantly different ($p=7.98 \times 10^{-185}$ and $p<0.001$ respectively).

I also found that obesity was more prevalent in males compared to the females in this cohort (Odds ratio (female) =0.88; [0.874, 0.902] 95% CI; Pearson's Exact $p=3.81 \times 10^{-48}$). In addition, the prevalence of obesity was significantly different according to different batches (Odds ratio=1.002; $p=3.36 \times 10^{-35}$) and ages (Odds ratio=1.007; $p=1.55 \times 10^{-38}$).

### 7.5.3. Obesity-related phenotype correlation analysis

Bivariate correlation tests were carried out to assess if there were any co-relationships between BMI and WHR.

**(a) NFBC1966 cohort**

Both obesity-related phenotypes (BMI and WHR) were significantly correlated to each other (n=4,331; $p=1.79 \times 10^{-282}$; Pearson's correlation (r) = 0.508).

**(b) PMMO cohort**

This step was not performed in this cohort due to unavailability of WHR data.

**(c) UKBB cohort**

The BMI was significantly correlated with WHR (n=328,277; $p<0.001$; r=0.441).

### 7.5.4. Confirmation of proximal 16p11.2 deletion carriers (~600 kb)

**(a) NFBC1966 cohort**

The clean CNV catalogue generated from **Chapter 5** was interrogated to identify any proximal 16p11.2 deletion calls, spanning the putative proximal 16p11.2 deletion region (hg19: chr16:29,600,000-31,200,000) in the NFBC1966 cohort. A total of 3 individuals (~0.06%) were identified from the CNV catalogue. The predicted heterozygous deletions (chr16:29,563,365-30,085,308) of 521,944 kb in length had at least 1 SNP/13 kb coverage, with internal confidence scores > 40. The presences of these deletions were subsequently confirmed by visually inspecting the signal intensity plots of the potential carriers (**Figure 7-5**). At 31 years of age, one of the deletion carriers was morbidly obese (BMI =45.72 kg/m$^2$; WHR=1.04), while another carrier was obese (BMI=33.42 kg/m$^2$; WHR=0.98). However, the third carrier was found to be underweight, with BMI of 19.2 kg/m$^2$. The mean BMI for the ~600-kb deletion carriers was 32.78 kg/m$^2$ (mean WHR=0.93).

**(b) PMMO cohort**

Upon scanning the CNV catalogue generated in **Chapter 5** for the PMMO cohort, only one CNV call was identified as a putative heterozygous deletion of the proximal 16p11.2 region. The predicted heterozygous deletion (chr16:29,647,342-30,198,151) of 550,810 kb in length had at least one SNP in every 7 kb of its length, with an internal confidence score of 376.88. The putative deletion was confirmed by visually inspecting the signal intensity plot of the carrier (**Figure 7- 6**). The deletion carrier was morbidly obese (sex=female; BMI= 61.16 kg/m$^2$).

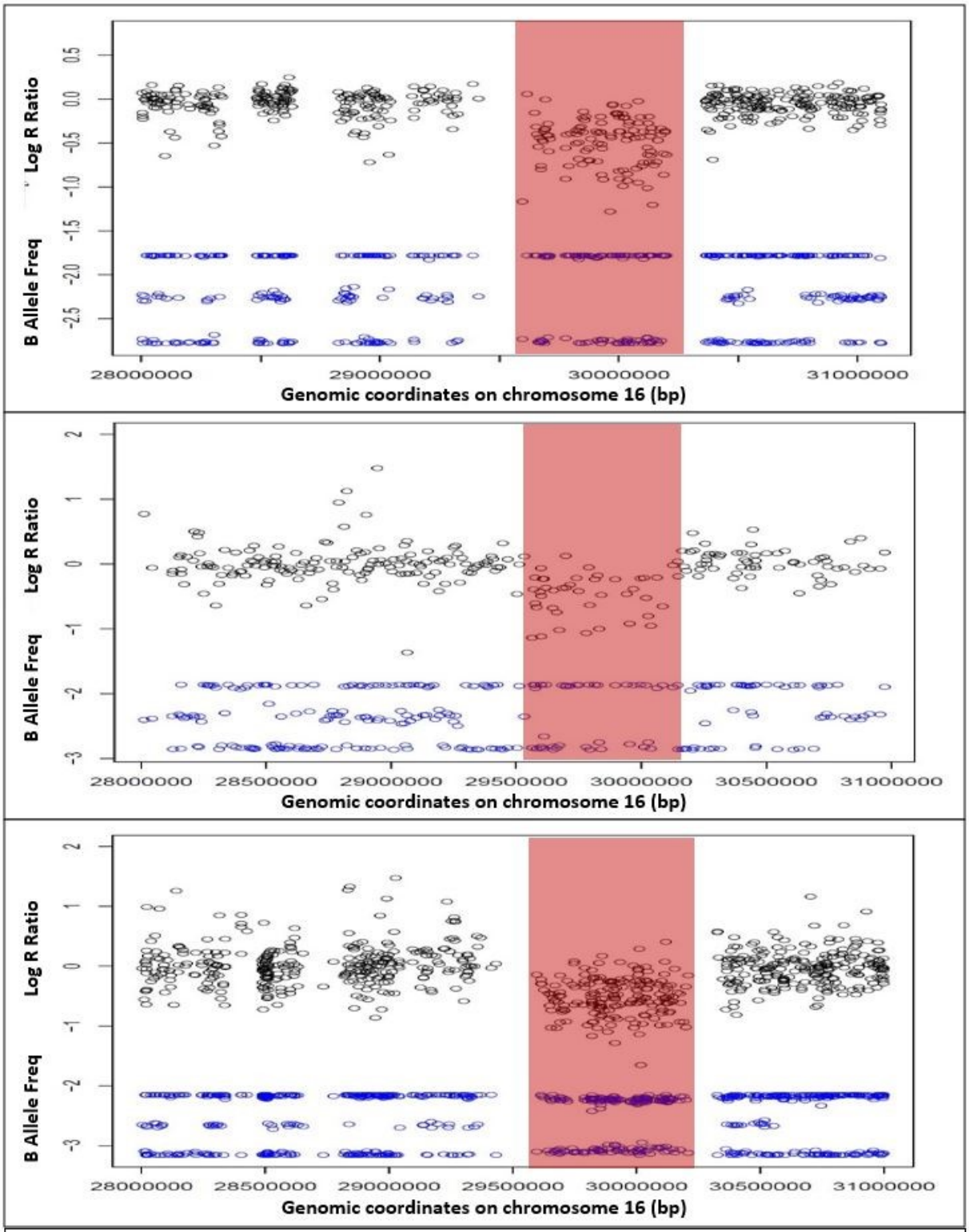


**Figure 7- 5:** Signal intensity plots of the proximal 16p11.2 deletion carriers in the NFBC1966 cohort (n=3). Each plot consists of Log R Ratio (LRR) values (top) and B Allele Freq (BAF) values (bottom), plotted against the genomic coordinates of SNPs on chromosome 16 (hg19). The red shades display the deletion (~600 kb) of the proximal 16p11.2 region.

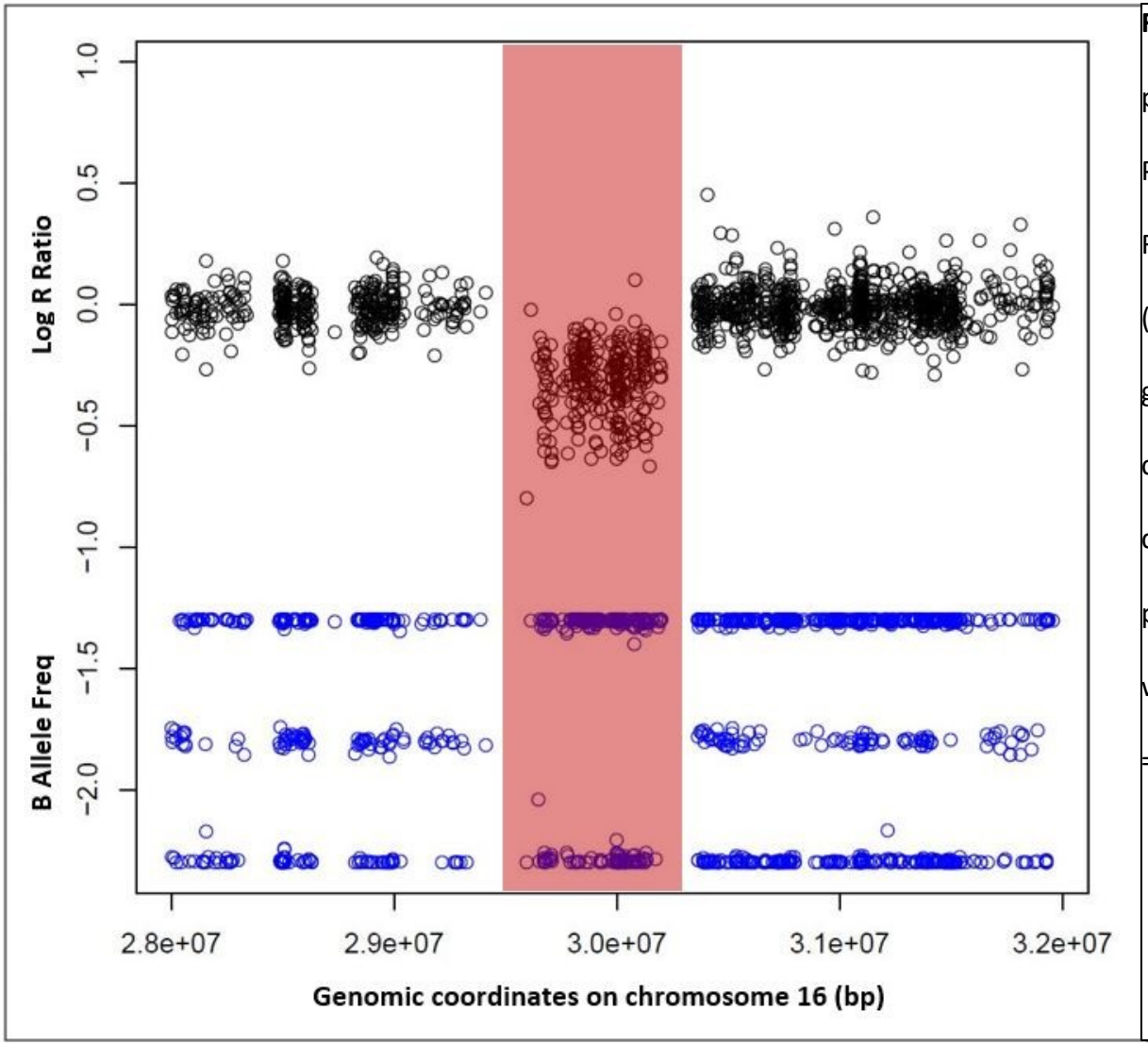


**Figure 7- 6:** Signal intensity plot of the proximal 16p11.2 deletion carrier in the PMMO cohort (n=1). The plot consists of Log R Ratio (LRR) values (top) and B Allele Freq (BAF) values (bottom), plotted against the genomic coordinates of SNPs on chromosome 16 (hg19). The red shade displays the deletion (~600 kb) of the proximal 16p11.2 region. The female carrier was morbidly obese (BMI=61.16 kg/$m^2$). BMI = body mass index.

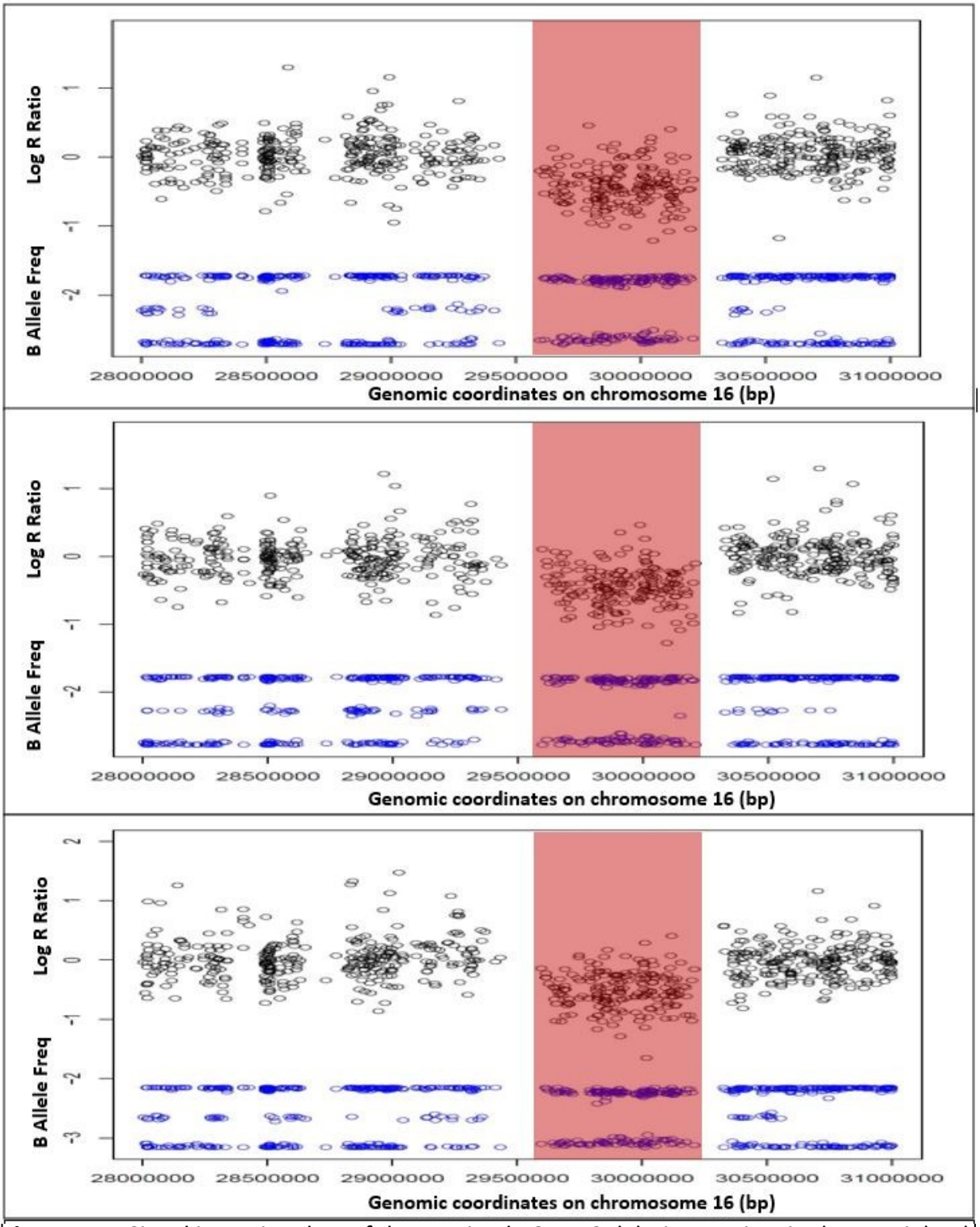


**Figure 7- 7:** Signal intensity plots of the proximal 16p11.2 deletion carriers in the UK Biobank cohort (n=3). Each plot (for each carrier) consists of Log R Ratio (LRR) values (top) and B Allele Freq (BAF) values (bottom), plotted against the genomic coordinates of SNPs on chromosome 16 (hg19). The red shades display the deletion (~600 kb) of the proximal 16p11.2 region.

**(c) UKBB cohort**

Based on the ROH, a total of 242 participants were identified to have a stretch of homozygosity of at least 500 kb in length, spanning the 16p11.2 region. The signal intensity plots for these participants were visually inspected to confirm the presence of a heterozygous 16p11.2 deletion of the proximal region. Only 3 of them (2 males, 1 female) were confirmed for the presence of such deletion (**Figure 7- 7**). The mean BMI and WHR of the deletion carriers were 39.7 kg/m$^2$ and 0.97 respectively, of which two of them were morbidly obese (BMI≥40 kg/m$^2$). The morbidly obese carriers consist of a female (BMI=46.45 kg/m$^2$; WHR=0.82) and a male participant (BMI=41.63 kg/m$^2$; WHR=0.96). The third carrier was a male who had a BMI of 31.05 kg/m$^2$ (WHR=1.11).

### 7.5.5. Confirmation of distal 16p11.2 deletion carriers (~200 kb)

**(a) NFBC1966 cohort**

The clean CNV catalogue generated from **Chapter 5** was scanned for any deletions located within the putative distal 16p11.2 deletion region (hg19: chr16:28,700,000-29,100,000) to identify whether any heterozygous deletion is present in the NFBC1966 cohort. A total of 4 in 4,905 (~0.08%) individuals from the general NFBC1966 population were identified from the CNV catalogue to have the distal 16p11.2 heterozygous deletion (~200 kb). The predicted heterozygous deletion (chr16:28,733,550-28,927,795) had a coverage of at least one SNP in every 8-kb base pairs, with internal confidence scores > 30. I confirmed the predicted heterozygous deletions by visually inspecting the signal intensity plots of all potential carriers (**Figure 7- 8**). At 31 years of age, 3 of the deletion carriers were overweight (BMI range=27.4 kg/m$^2$ to 29 kg/m$^2$), while only one person was obese (BMI=37.23kg/m$^2$; WHR=0.99).

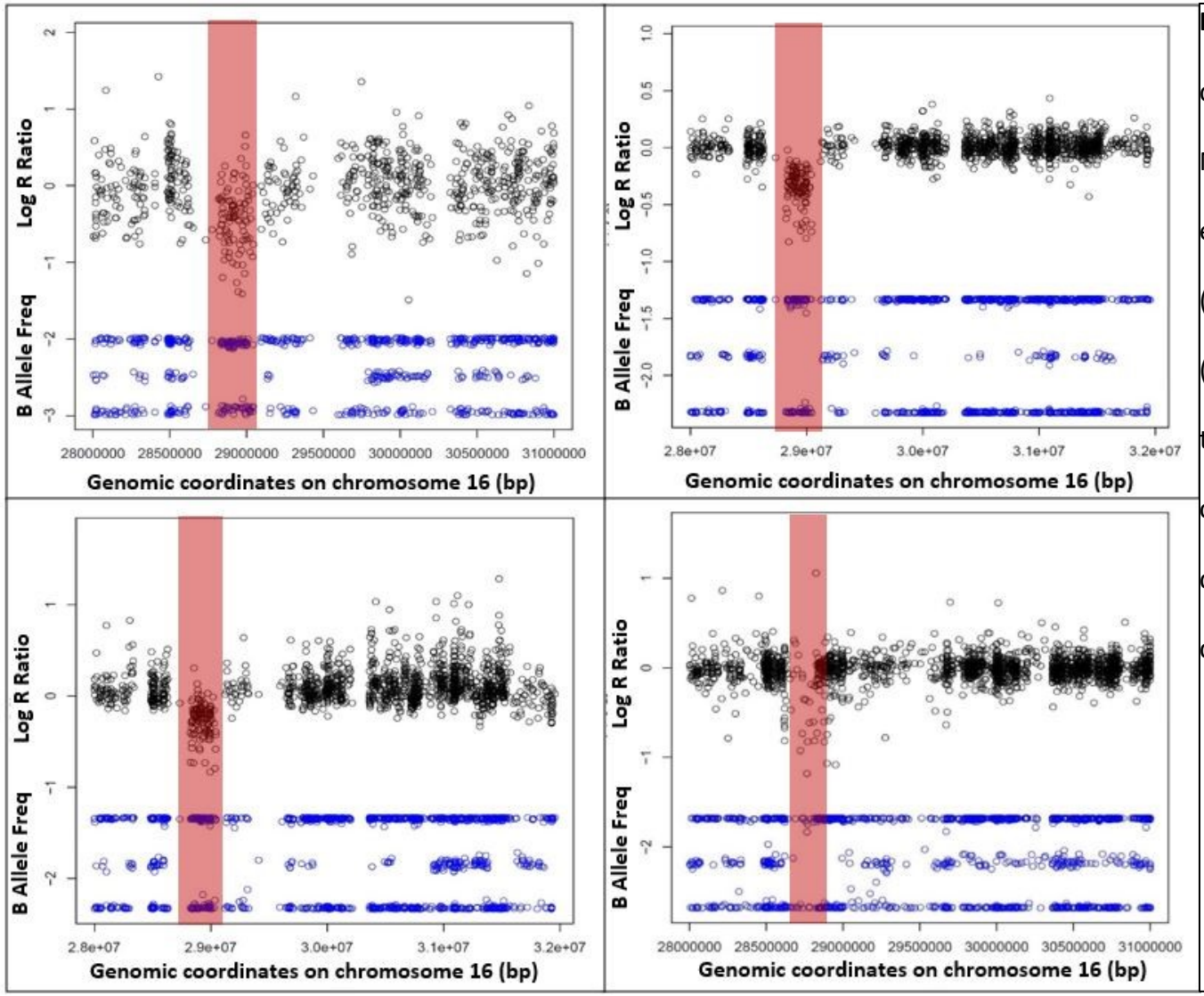


**Figure 7- 8:** Signal intensity plots of the distal 16p11.2 deletion carriers in the NFBC1966 cohort (n=4). Each plot (for each carrier) consists of Log R Ratio (LRR) values (top) and B Allele Freq (BAF) values (bottom), plotted against the genomic coordinates of SNPs on chromosome 16 (hg19). The red shades display the deletion (~200 kb) of the distal 16p11.2 region.

**(b) PMMO cohort**

The cleaned CNV catalogue for the PMMO cohort (refer to **Chapter 5**) was interrogated for any predicted heterozygous deletion calls that spanned the putative distal 16p11.2 region (hg19: chr16:28,700,000-29,100,000). I found one predicted distal 16p11.2 deletion (chr16:28,930,759-29,043,450) with coverage of at least one SNP per 2 kb in length (internal confidence score=33.757). The predicted deletion call was confirmed via visual inspection of signal intensity plots of the carrier, as shown in **Figure 7- 9**. The deletion carrier was obese at the time of recruitment (sex=male; BMI=38.44 kg/$m^2$).

**(c) UKBB cohort**

A total of 304 participants were initially identified to have a ~200 kb of homozygosity stretch within the 16p11.2 region. Upon visual inspection of each signal intensity plot of these participants, only 1 participant (female) was confirmed to have the heterozygous deletion for the distal 16p11.2 deletion region (**Figure 7- 10**). The BMI of the deletion carrier, who is morbidly obese, was 40.6 kg/$m^2$ (WHR=0.97).

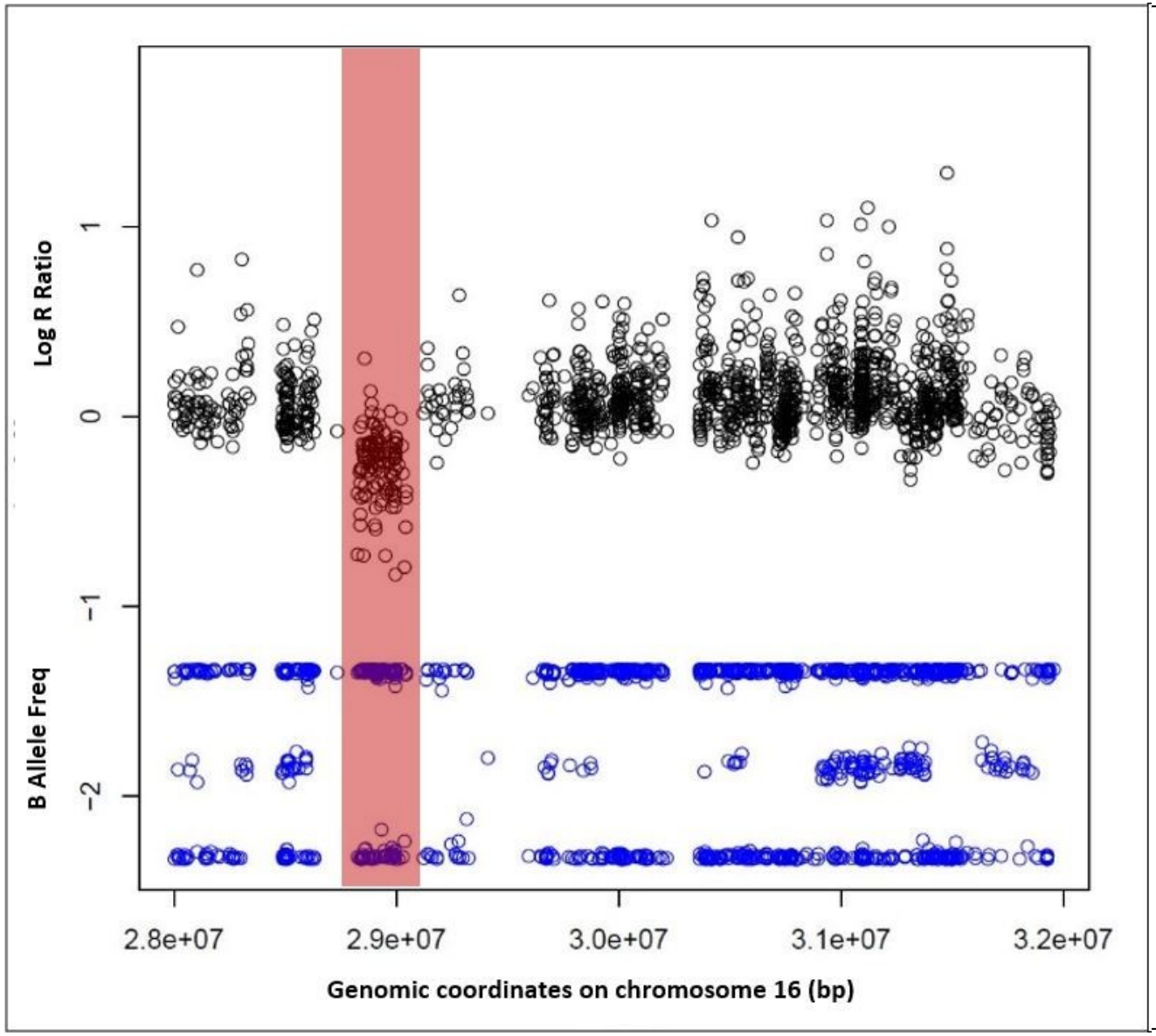


**Figure 7- 9:** Signal intensity plot of the distal 16p11.2 deletion carrier in the PMMO cohort (n=1). The plot consists of Log R Ratio (LRR) values (top) and B Allele Freq (BAF) values (bottom), plotted against the genomic coordinates of SNPs on chromosome 16 (hg19). The red shade displays the deletion (~200 kb) of the distal 16p11.2 region. The male carrier was obese (BMI=38.44 kg/m$^2$). BMI = body mass index; WHR = waist-to-hip ratio.

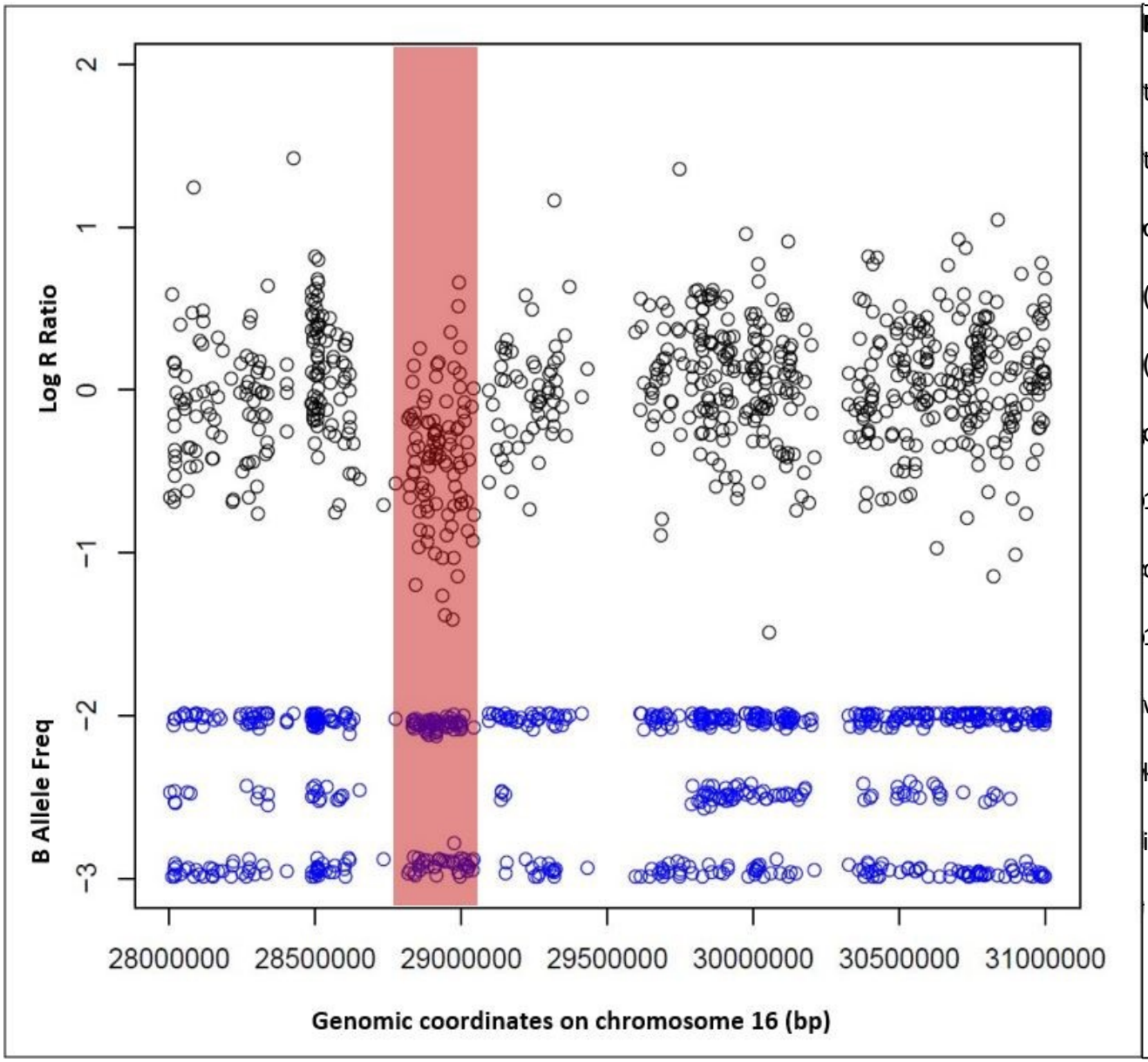


**Figure 7- 10:** Signal intensity plot of the distal 16p11.2 deletion carrier in the UK Biobank cohort (n=1). The plot consists of Log R Ratio (LRR) values (top) and B Allele Freq (BAF) values (bottom), plotted against the genomic coordinates of SNPs on chromosome 16 (hg19). The red shade displays the deletion (~200 kb) of the distal 16p11.2 region. The female carrier was morbidly obese, with BMI at 40.6 kg/m$^2$ (WHR=0.97). BMI = body mass index; WHR = waist-to-hip ratio.

### 7.5.6. Association analyses of proximal 16p11.2 deletion on obesity-related traits

**(a) NFBC1966 cohort**

Initial detections of the proximal 16p11.2 deletions by the PennCNV tool were subsequently confirmed by visual inspections of signal intensity plots for 3 out of 4,905 individuals (~0.06%; 2 females, 1 male). Independence of observation was determined for both obesity-related phenotypes (BMI and WHR) and deletion carrier status (Durbin-Watson statistics=2.00) to avoid violating the assumption of non-autocorrelation. Prior to regression tests, independent t-tests were performed to compare the means of obesity traits (i.e. BMI and WHR) between the carriers and non-carriers of the proximal 16p11.2 deletion. Linear regressions were performed for the association analysis between the deletion and both obesity traits (i.e. BMI and WHR). The regression tests were adjusted for sex and genotyping batch. **Table 7- 1** shows the statistical output for the linear regression between the proximal 16p11.2 deletion and BMI, as well as WHR.

*Body mass index* – Independent t-tests revealed that the mean BMI of the proximal 16p11.2 deletion carriers was significantly different than the mean BMI of the non-carriers ($p= 8.0 \times 10^{-4}$; mean BMI difference=-8.13; [-12.918, -3.343] 95% CI). Linear regression tests revealed that the ~600-kb deletion was associated with BMI (p=0.001; adjusted $r^2$=0.018) (**Table 7- 1**). On average, individuals with the heterozygous ~600-kb deletion, were predicted by the regression model to have an increased BMI of 8.313 $kg/m^2$ ([3.563, 13.063] 95% CI). The association of the heterozygous deletion and BMI stayed significant when the regression test was conducted in females only ($p=2.39 \times 10^{-6}$; adjusted $r^2$=0.009) (**Table 7- 1**).

| Sub-population (n deletion carriers) | Adjusted $r^2$ | SE of the Estimate | Beta coefficient | 95% CI for Beta coefficient | | P-value |
|---|---|---|---|---|---|---|
| | | | | Lower Bound | Upper Bound | |
| **Body mass index (BMI)** | | | | | | |
| All (n=3) | 0.018 | 4.195 | 8.313 | 3.563 | 13.063 | **0.001** |
| Obese (n=2) | 0.048 | 4.233518 | 4.879 | -1.048 | 10.805 | 0.106 |
| *morbidly obese (n=1) | - | - | - | - | - | - |
| *Males (n=1) | - | - | - | - | - | - |
| Females (n=2) | 0.009 | 4.661054 | 15.596 | 9.128 | 22.064 | **$2.39 \times 10^{-6}$** |
| **Waist-to-hip ratio (WHR)** | | | | | | |
| All (n=3) | 0.340 | 0.070 | 0.042 | 0.06 | 0.219 | **0.001** |
| Obese (n=2) | 0.275 | 0.074 | 0.108 | 0.004 | 0.212 | **0.042** |
| *morbidly obese (n=1) | - | - | - | - | - | - |
| *Males (n=1) | - | - | - | - | - | - |
| Females (n=2) | 0.005 | 0.079 | 0.075 | 0.089 | 0.310 | **$4.09 \times 10^{-4}$** |

**Table 7- 1:** Statistical outputs for association analyses of the proximal 16p11.2 deletion (~600 kb) on obesity-related phenotypes (BMI and WHR) in the NFBC1966 cohort (n=4,905). Linear regression tests were carried out for the association analyses, adjusted for sex and genotyping batch. P-value < 0.05 was considered as significant (bold). Obese = BMI > 30 kg/$m^2$, morbidly obese = BMI > 40 kg/$m^2$, * Linear regression test was not performed due to lack of carriers, CI = confidence interval, BMI = body mass index, WHR = waist-to-hip ratio, SE = standard error, adjusted $r^2$ = proportion of the variance explained by the regression model, beta coefficient = estimated effect size.

*Waist-to-hip ratio* – A comparison of means between the proximal deletion carriers and non-carriers indicated a significant difference on the means of WHR (p=0.014; mean WHR difference= 0.139; [0.06, 0.219] 95% CI). The linear regression tests between the proximal 16p11.2 deletion and WHR, showed a significant evidence of association in the whole cohort (p=0.001; adjusted $r^2$=0.34), as well as in females only (p=4.09 x $10^{-4}$; adjusted $r^2$=0.005) (**Table 7- 1**). In addition, I also observed a borderline significant association in the obese individuals (p=0.042).

**(b) PMMO cohort**

Upon confirmation of the proximal 16p11.2 heterozygous deletion in 1 out of 115 patients (~0.87% of obese; ~1.04% of morbidly obese), I looked at whether the BMI and the heterozygous deletion were independently observed via Durbin-Watson statistics. I found that BMI was independently observed from the presence of proximal 16p11.2 heterozygous deletion among the PMMO patients (Durbin-Watson statistics =1.91). Since only 1 female carrier was detected, statistical tests could not be performed to assess the associative relationship of the ~600-kb deletion and obesity-related phenotypes in this cohort. The proximal 16p11.2 deletion carrier was a morbidly obese female from the obesity cohort (PMMO).

**(c) UKBB cohort**

A combination of ROH and visual inspection of the signal intensity plots was used to confirm the presence of the ~600-kb heterozygous deletions in 3 from 329,419 (<0.05%) participants from the UK Biobank cohort (UKBB). Linear regression tests were conducted to find statistical associations between the proximal 16p11.2 deletion and obesity traits (i.e. BMI

and WHR), adjusted for sex, age and genotyping batch. Both obesity traits (BMI and WHR) and the presence of the BP4-BP5 deletion were independently observed from each other (Durbin-Watson statistics = 1.996). **Table 7- 2** shows the results of the statistical analyses conducted for the associations of proximal 16p11.2 deletion with BMI and WHR.

*Body mass index* – In the UKBB cohort, the proximal 16p11.2 was found to be significantly associated with BMI ($p=9.75 \times 10^{-6}$; adjusted $r^2=0.01$), in which having the heterozygous deletion would increase the BMI by 12.08 kg/m$^2$ ([6.73, 17.44] 95% CI) (**Table 7- 2**). The test was extended to observe the associative relationship according to sex. However, the regression analysis was not performed in females because only 1 female was confirmed to have the ~600-kb deletion. The sex-stratified analysis revealed that the association between the proximal deletion and BMI was still significant among males ($p=0.004$; adjusted $r^2=0.001$). In males, the presence of the ~600 kb deletion would increase the BMI by 8.52 kg/m$^2$ ([2.66, 14.37] 95% CI). The association analysis was subsequently carried out in different obesity sub-populations within the UKBB cohort. I observed a significant association of the proximal 16p11.2 deletion with BMI among the obese participants ($p=0.006$; adjusted $r^2=0.022$), but not in the morbidly obese sub-population ($p=0.868$) (**Table 7- 2**).

*Waist-to-hip ratio* – According to **Table 7- 2**, the presence of the proximal 16p11.2 deletion (~600 kb) was significantly associated with WHR in males only ($p=0.04$; adjusted $r^2=0.032$). The regression model predicted that on average, males who have the deletion would have an increased WHR of 0.093 unit ([0.004, 0.181] 95% CI). Furthermore, I could not find any significant association between the proximal 16p11.2 deletion and WHR in both obesity sub-populations within the UKBB cohort (i.e. obese, morbidly obese) (**Table 7- 2**).

| Sub-population (n deletion carriers) | Adjusted $r^2$ | SE of the Estimate | Beta coefficient | 95% CI for Beta coefficient | | P-value |
|---|---|---|---|---|---|---|
| | | | | Lower Bound | Upper Bound | |
| **Body mass index (BMI)** | | | | | | |
| All (n=3) | 0.010 | 4.732 | 12.084 | 6.729 | 17.440 | **$9.75 \times 10^{-6}$** |
| Obese (n=3) | 0.022 | 3.812 | 6.098 | 1.784 | 10.413 | **0.006** |
| morbidly obese (n=2) | 0.011 | 3.741 | 0.439 | -4.747 | 5.625 | 0.868 |
| Males (n=2) | 0.001 | 4.222 | 8.515 | 2.663 | 14.367 | **0.004** |
| *Females (n=1) | - | - | - | - | - | - |
| **Waist-to-hip ratio (WHR)** | | | | | | |
| All (n=3) | 0.452 | 0.066 | 0.065 | -0.011 | 0.140 | 0.093 |
| Obese (n=3) | 0.505 | 0.062 | 0.016 | -0.054 | 0.087 | 0.653 |
| morbidly obese (n=2) | 0.515 | 0.069 | -0.056 | -0.152 | 0.040 | 0.251 |
| Males (n=2) | 0.032 | 0.064 | 0.093 | 0.004 | 0.181 | **0.040** |
| *Females (n=1) | - | - | - | - | - | - |

**Table 7- 2:** Statistical results for the association analyses of the proximal 16p11.2 deletion (~600 kb) on obesity traits (BMI and WHR) in the UK Biobank cohort (n=329,419). Linear regression tests were carried out for the association analyses, adjusted for sex, age and batch. P-value < 0.05 was considered as significant (bold). Obese = BMI > 30 kg/$m^2$, morbidly obese = BMI > 40 kg/$m^2$, * Linear regression test was not performed due to lack of carriers, CI = confidence interval, BMI = body mass index, WHR = waist-to-hip ratio, SE = standard error, adjusted $r^2$ = proportion of the variance explained by the regression model, beta coefficient = estimated effect size.

### 7.5.7. Association analyses of distal 16p11.2 deletion on obesity-related phenotypes

**(a) NFBC1966 cohort**

The CNV calling process by the PennCNV tool have yielded 4 carriers of the distal 16p11.2 deletion variant (~0.08%) from 4,905 individuals from the NFBC1966 cohort. All predicted deletions were visually confirmed from the signal intensity plots. Both obesity-related phenotypes were found to be independently observed from the presence of such deletion (Durbin-Watson statistics ≥ 1.994). **Table 7- 3** shows the statistical results of the linear regression tests between proximal 16p11.2 deletion and WHR for the NFBC1966 cohort.

*Body mass index* – Independent t-tests for comparison of means revealed that the mean of BMI of deletion carriers was significantly different than in non-carriers (p=0.007; mean BMI difference= -5.726; [-9.875, -1.578] 95% CI). In relation to the association analysis, I found that the presence of distal 16p11.2 heterozygous deletion was significantly associated with BMI (p=0.0061; adjusted $r^2$=0.017). The model predicted that on average, individuals who had such a deletion would have their BMI increased by 5.77 kg/$m^2$ ([1.65, 9.89] 95% CI). When stratified by sex, I found a significant association amongst males only (p=0.002; adjusted $r^2$=0.003) (**Table 7- 3**). In the obese sub-population, no significant association was detected (p>0.05).

| Sub-population (n deletion carriers) | Adjusted $r^2$ | SE of the Estimate | Beta coefficient | 95% CI for Beta coefficient | | P-value |
|---|---|---|---|---|---|---|
| | | | | Lower Bound | Upper Bound | |
| **Body mass index (BMI)** | | | | | | |
| All (n=4) | 0.017 | 4.197 | 5.768 | 1.650 | 9.886 | **0.0061** |
| *obese (n=1) | - | - | - | - | - | - |
| *morbidly obese (n=0) | - | - | - | - | - | - |
| Males (n=2) | 0.003 | 3.612 | 7.848 | 2.835 | 12.861 | **0.002** |
| Females (n=2) | 0.001 | 4.681 | 3.673 | -2.823 | 10.169 | 0.268 |
| **Waist-to-hip ratio (WHR)** | | | | | | |
| All (n=4) | 0.339 | 0.0703 | -0.073 | 0.004 | 0.142 | **0.038** |
| *obese (n=1) | - | - | - | - | - | - |
| *morbidly obese (n=0) | - | - | - | - | - | - |
| Males (n=2) | 0.001 | 0.059 | -0.081 | -0.001 | 0.162 | 0.052 |
| Females (n=2) | 0.0004 | 0.0798 | -0.065 | -0.045 | 0.176 | 0.248 |

**Table 7- 3:** Statistical outputs for association analyses of distal 16p11.2 deletion (~200 kb) on obesity traits (BMI and WHR) in the NFBC1966 cohort. Linear regression tests were carried out for the association analyses, adjusted for sex and batch. P-value > 0.05 was considered as significant (bold). Obese = BMI > 30 kg/$m^2$, morbidly obese = BMI > 40 kg/$m^2$, * Linear regression test was not performed due to lack of carriers, BMI = body mass index, WHR = waist-to-hip ratio, SE = standard error, adjusted $r^2$ = proportion of the variance explained by the regression model, beta coefficient = estimated effect size.

*Waist-to-hip ratio* – Independent t-tests revealed that the mean of WHR for the deletion carriers was not significantly different from the mean of WHR of the non-carriers ($p>0.05$). The association tests revealed significant association of the distal 16p11.2 heterozygous deletion with WHR in the whole cohort, adjusted for sex and batch ($p=0.038$; adjusted $r^2=0.339$) (**Table 7- 3**). The presence of such deletion was predicted in the regression model to increase the WHR by 0.073 unit ([0.004, 0.142] 95% CI).

**(b) PMMO cohort**

One out of 115 patients (~0.87%) from the bariatric surgery cohort (consisting obese and morbidly obese patients) had the distal 16p11.2 deletion variant. Thus, the BMI of the male carrier (BMI=38.44 kg/$m^2$) was not statistically assessed due to insufficient carriers.

**(c) UKBB cohort**

One in 6,100 (~0.02%) morbidly obese participants (female, BMI=40.60 kg/$m^2$, WHR=0.97) was confirmed to have a heterozygous deletion of the distal 16p11.2 deletion. Due to the lack of cases to run a linear regression test, statistical tests were not performed to assess the associations of the distal 16p11.2 deletion with each obesity-related phenotypes in this cohort.

### 7.5.8. Risk of obesity in the general population

To investigate whether the 16p11.2 deletion variants increase the risks of obesity in the general population, I carried out Fisher's exact tests between the deletion carrier status (binary; carrier, non-carrier) and obesity status (binary; not obese, obese) for the NFBC1966 and UKBB cohorts. Two-by-two contingency tables were constructed for each 16p11.2 deletion variant prior to the Fisher's exact tests. This work was conducted in Rstudio version 1.1.419 (RStudio, Inc., Boston, MA, USA) (RStudio, 2015). **Table 7- 4** shows the statistical results of Fisher's exact tests for the effects of each 16p11.2 deletion on the risk of obesity in the NFBC1966 and UKBB cohorts.

*Proximal 16p11.2 heterozygous deletion* – The statistical results in **Table 7- 4** show that the ~600-kb deletion was significantly associated with obesity in both population cohorts. In the NFBC1966 cohort, the proximal 16p11.2 deletion carriers were 20.66 times more likely to be obese than the non-carriers (Fisher's exact p=0.022; [1.073, 1209.829] 95% CI). In the UKBB cohort, the significant association was higher than in the NFBC1966 cohort, with Fisher's exact p at 0.0134. By having the ~600-kb deletion, the carriers from the UKBB cohort were 12.66 times more likely to be obese than the non-carriers ([1.253, 621.287] 95% CI).

*Distal 16p11.2 heterozygous deletion* – According to the results shown in **Table 7- 4**, no significant association was detected between the *SH2B1*-containing deletion and obesity in either population cohorts (NFBC1966 and UKBB).

| Study cohort | Deletion carrier, n | Obese/Not obese, n | Odds Ratio (OR) | 95% Confidence Intervals | | p-value (Fisher's exact) |
|---|---|---|---|---|---|---|
| | | | | Upper | Lower | |
| **Proximal 16p11.2 heterozygous deletion (~600 kb)** | | | | | | |
| NFBC1966 (obesity) | 3 | 428/4,407 | 20.655 | 1.073 | 1209.829 | **0.0221** |
| UKBB (obesity) | 3 | 79,096/250,323 | 12.659 | 1.253 | 621.287 | **0.0134** |
| **Distal 16p11.2 heterozygous deletion (~200 kb)** | | | | | | |
| NFBC1966 (obesity) | 4 | 428/4,407 | 3.437 | 0.065 | 42.901 | 0.3099 |
| UKBB (obesity) | 1 | 79,096/250,323 | 6.33 | 0.33 | 372.525 | 0.1453 |

**Table 7- 4:** The effects of the 16p11.2 heterozygous deletion variants on the risk of obesity in two general population cohorts (NFBC1966 and UKBB). Fisher's exact $p < 0.05$ was considered as significant (bold). NFBC1966 = North Finland Birth Cohort 1966; UKBB = United Kingdom Biobank; n = number of individuals; OR = odds ratio.

### 7.5.9. Risk of obesity in a pooled case-control association analysis of PMMO and UKBB

I followed up the association analyses by combining the datasets from the PMMO and UKBB cohorts to compare the effects of each 16p11.2 deletion variant in an extreme phenotype group versus controls (refer **Chapter 7.4.6: Statistical analyses –** *Obesity status associations*). Both cohorts consisted of Caucasians. The case/control dataset was created to provide sufficient power for a robust statistical conclusion on the effects of the 16p11.2 deletion variants on the risks for obesity and morbid obesity. The case/control dataset was also used to identify if the 16p11.2 deletion variants were significantly enriched in obesity patient population (PMMO cohort) as compared to the general UK population (UKBB cohort). **Table 7- 5** shows the statistical findings for the associations of each 16p11.2 deletion variant between the obesity patient population and the general population. The odds ratios (OR) for the predisposition of obesity (and morbid obesity) were also presented.

*Proximal 16p11.2 heterozygous deletion* – The proximal 16p11.2 deletion was significantly enriched in the obesity patients as compared to the obese individuals in the general UK population (Fisher's exact p=0.0057; OR=237.28; [4.47, 3161.68] 95% CI). However, the significant enrichment was reduced to p=0.042 in morbidly obese patients (OR=34.95; [0.59, 675.17] 95% CI), as compared to the morbidly obese individuals in the control population. When I included the non-obese individuals in the control population, I observed an increased significance of proximal 16p11.2 deletion enrichment in obese (Fisher's exact p=0.0014) and morbidly obese patients (Fisher's exact p=0.0012) (**Table 7- 5**).

| Sub-population | n | OR | 95% Confidence Intervals | | p-value |
|---|---|---|---|---|---|
| | | | Upper | Lower | |
| **Proximal 16p11.2 heterozygous deletion** | | | | | |
| PMMO/general (all UKBB) | 114/329,419 | 975.638 | 18.354 | 16384 | **0.0014** |
| Morbidly obese PMMO/general (all UKBB) | 96/329,419 | 1142.226 | 21.793 | 16384 | **0.0012** |
| PMMO/obese UKBB | 114/80,148 | 237.284 | 4.465 | 3161.684 | **0.0057** |
| Morbidly obese PMMO/morbidly obese UKBB | 103/7,153 | 34.948 | 0.588 | 675.172 | **0.0419** |
| Obese PMMO+UKBB/non-obese (UKBB) | 79,203/249,271 | 15.734 | 1.761 | 741.334 | **0.0039** |
| morbidly obese (PMMO+UKBB)/not morbidly obese (PMMO + UKBB) | 6,197/322,277 | 155.484 | 12.527 | 7667.663 | **$2.65 \times 10^{-5}$** |
| **Distal 16p11.2 heterozygous deletion** | | | | | |
| PMMO/general (all UKBB) | 114/329,419 | 2790.872 | 36.825 | $4.50 \times 10^{15}$ | **0.0007** |
| Morbidly obese PMMO/general (all UKBB) | 96/329,419 | 1779.831 | 2.840 | $4.50 \times 10^{15}$ | **0.0009** |
| PMMO/obese UKBB | 114/80,148 | 714.110 | 8.958 | $4.50 \times 10^{15}$ | **0.0028** |
| Morbidly obese PMMO/morbidly obese UKBB | 103/7,153 | 34.288 | 0.578 | 662.286 | **0.0427** |
| Obese PMMO+UKBB/non-obese (UKBB) | 79,203/249,271 | 9.440 | 0.758 | 494.230 | **0.0459** |
| morbidly obese (PMMO+UKBB)/not morbidly obese (PMMO + UKBB) | 6,197/322,277 | 52.010 | 0.663 | 3923.569 | **0.0374** |

(Table continues on the next page)

(Table continuation from the previous page)

| Sub-population | n | OR | 95% Confidence Intervals | | p-value |
|---|---|---|---|---|---|
| **Either deletion variant** | | | | | |
| PMMO/general (all UKBB) | 114/329,419 | 1403.899 | 131.282 | 16384 | **$1.78 \times 10^{-6}$** |
| Morbidly obese PMMO/general (all UKBB) | 96/329,419 | 867.934 | 17.366 | 8192 | **0.0015** |
| Obese PMMO/obese UKBB | 114/80,148 | 356.084 | 31.989 | 2428.229 | **$2.99 \times 10^{-5}$** |
| Morbidly obese PMMO/morbidly obese UKBB | 103/7,153 | 21.341 | 0.403 | 268.346 | 0.0606 |
| Obese PMMO+UKBB/non-obese (UKBB) | 79,203/249,271 | 22.037 | 2.830 | 987.849 | **$2.99 \times 10^{-4}$** |
| morbidly obese (PMMO+UKBB)/not morbidly obese (PMMO + UKBB) | 6,197/322,277 | 104.321 | 14.912 | 1133.618 | **$1.84 \times 10^{-6}$** |

**Table 7- 5:** Statistical analyses (Fisher's exact tests) on the associations of the 16p11.2 heterozygous deletion variants in patient population (PMMO cohort) as compared to the general population (UKBB cohort), and the predispositions of 16p11.2 deletion carriers to obesity and morbid obesity. Fisher's exact $p < 0.05$ was considered as significant (bold). PMMO = Personalised Medicine for Morbid Obesity cohort; UKBB = United Kingdom Biobank; n = number of individuals tested; OR = odds ratio.

In the comparison between 1) obese and non-obese people, and 2) morbidly obese and non-morbidly obese people, I found that the ~600-kb heterozygous deletion was associated with obesity (Fisher's exact p=0.0039; OR=15.73; [1.76, 741.33] 95% CI) and morbid obesity (Fisher's exact p=2.65 x $10^{-5}$; OR=155.48; [12.53, 7667.66] 95% CI) (**Table 7- 5**).

*Distal 16p11.2 heterozygous deletion* – As for the ~200-kb deletion containing the *SH2B1* gene, I found a significant enrichment of the heterozygous deletion in the obese patients as compared to the obese individuals from the general population (Fisher's exact p=0.0028; OR=714.11; [8.96, 4.50 x $10^{15}$] 95% CI) (**Table 7- 5**). The significant enrichment of the distal 16p11.2 deletion was nominal in morbidly obese patients from the obesity population (Fisher's exact p=0.043; OR=34.29; [0.58, 662.29] 95% CI). Similar to the proximal 16p11.2 deletion, the distal ~200-kb deletion had an increased significance of enrichment in both obese (Fisher's exact p=0.0007) and morbidly obese patients (Fisher's exact p=0.0009), when the non-obese individuals were included in the control population (**Table 7- 5**). In an overall obesity versus non-obesity (and morbid obesity versus non-morbid obesity), I found that the distal 16p11.2 heterozygous deletion was not associated with obesity (p=0.05), but it was significantly associated with morbid obesity (p=0.038; OR=52.01; [0.66, 3923.57] 95% CI) (**Table 7- 5**).

*Either 16p11.2 deletion* – I also looked at the associations of having either the proximal or the distal 16p11.2 deletions (i.e. proximal/distal deletion), with the risk for obesity and morbid obesity. Analysis for having both deletion variants was irrelevant since none among the PMMO (bariatric patients) and the UKBB (general population) cohorts had both variants. Compared to the obese and morbidly obese individuals from the general UK population, I

found that the proximal/distal deletion was significantly enriched in the obese patients (Fisher's exact p=2.99 x $10^{-5}$; OR=356.08; [31.99, 2428.23] 95% CI), but not in morbidly obese patients (Fisher's exact p=0.061; OR=21.34; [0.40, 268.35] 95% CI) (**Table 7- 5**). Despite that, the significances were increased to p=1.78 x $10^{-6}$ (obese) and p=0.0015 (morbidly obese) when the general population group (controls) was expanded with the inclusion of non-obese individuals. In the comparison between obese and non-obese people from the pooled dataset, the proximal/distal deletion was significantly associated with obesity (Fisher's exact p= 2.99 x $10^{-4}$; OR=22.04; [2.83, 987.85] 05% CI) (**Table 7- 5**). I found the same conclusion in morbid obesity, with the association of proximal/distal deletion to be significant at p=1.84 x $10^{-6}$ (Fisher's exact; OR=104.32; [14.91, 1133.62] 95% CI).

# 7.6. Discussions

## 7.6.1. Results overview

The 16p11.2 chromosomal region contains an abundance of segmental duplications with high sequence similarity (>95%). These segmental duplications have been accounted for the high frequency of recurrent structural variations that contribute to the overall genomic instability within the 16p11.2 region (Bachmann-Gagescu *et al.*, 2010, Bochukova *et al.*, 2010, Walters *et al.*, 2010). Here I report an association study of two rare, recurrent deletion variants in the 16p11.2 region, with two obesity-related phenotypes (i.e. BMI and WHR) in 3 different study cohorts of adults; Northern Finland Birth Cohort 1966 (NFBC1966), Personalised Medicine for Morbid Obesity (PMMO), and the United Kingdom Biobank (UKBB). I also presented the association results between each 16p11.2 deletion variant, and the risks for obesity and morbid obesity in a pooled case/control dataset consisted of obesity patients as cases (PMMO) and general UK individuals as controls (UKBB).

In this study, I have presented a significant association between the proximal 16p11.2 deletion and BMI in both NFBC1966 and UKBB cohorts. In addition, I also found that in the UKBB cohort, the association between the ~600-kb deletion and BMI was significant among the obese participants (BMI≥30 kg/$m^2$). Importantly in the UKBB cohort, I replicated the known effect of the ~600-kb deletion on BMI near the genome-wide significance level (Beta coefficient=12.08 kg/$m^2$; p=9.75 x $10^{-6}$) (Jacquemont *et al.*, 2011, Mace *et al.*, 2017, Zufferey *et al.*, 2012). Despite reaching a similar conclusion in the NFBC1966 cohort, the significance was not as high, possibly because it has a smaller sample size than the UKBB cohort. Having said that, the adult NFBC1966 cohort had an observed statistical power of 92.9% to detect

true significant associations for both obesity traits. In the PMMO cohort, in which obesity was ascertained, the BMI of the morbidly obese patient (female) who has the proximal 16p11.2 deletion was not significantly different than the mean BMI from the rest of the cohort. This could be attributed to the small window for variation from a tight BMI range; 34.4 to 82.6 $kg/m^2$. In terms of the associative relationship between the proximal 16p11.2 deletion and WHR, I have replicated the association from a previous report (Beta coefficient=-0.04; p=2.3 x $10^{-7}$) (Mace *et al.*, 2017) albeit the lower significance of my findings in the NFBC1966 cohort (Beta coefficient=-0.14; p=6.0 x $10^{-4}$).

I have also shown that the smaller, distal 16p11.2 deletion (~200 kb) is significantly associated with BMI and WHR in the adult NFBC1966 cohort. The associative relationship could not be investigated in the PMMO and UKBB cohorts due to the paucity of deletion carriers in each of the two cohorts. Nonetheless, I did compare the BMI (and WHR) of the distal 16p11.2 deletion carriers with the mean BMI (and mean WHR) of the non-carriers in the two cohorts. I found that both obesity traits of the ~200-kb deletion carrier were significantly different than the means of the non-carriers in the UKBB cohort, but not in the PMMO cohort (BMI).

In my case/control association design, I have shown that the proximal 16p11.2 heterozygous deletion was significantly associated with obesity and morbid obesity. It is reasonable to suggest that, at least in the UK, adults who have the ~600-kb deletion variant (heterozygous) are predisposed to obesity and morbid obesity (**Table 7- 5**). My findings are consistent with previous associations reported by multiple studies (Bochukova *et al.*, 2010, Jacquemont *et al.*, 2011, Walters *et al.*, 2010), although the relatively lower significant results may be attributed to the possibility of underestimating the number of ~600-kb deletion carriers in the general UKBB cohort (discussed below). In addition, I have presented sufficient

evidence that the smaller, SH2B1-containing, distal 16p11.2 heterozygous deletion is associated with obesity and morbid obesity, when the control group was expanded to included non-obese individuals. The ~200-kb deletion was significantly enriched in obese patient population in comparison to the control population (all participants, Fisher's exact p=0.0007; obese-only participants, Fisher's exact p=0.003). My findings are comparable to the findings stated by Bachmann-Gagescu, *et. al.* (2010), in which they also reported a significant enrichment in patients at p=0.003 (Bachmann-Gagescu *et al.*, 2010).

### 7.6.2. The frequency of 16p11.2 deletion variants

Overall, this chapter presents the evidence of a significant relationship between the proximal 16p11.2 heterozygous deletion and BMI, in which the presence of such deletion in individuals increases the risk for obesity in the general population. I have confirmed the presence of the proximal 16p11.2 deletion variant in three individuals from the NFBC1966 (~0.5% of all obese individuals) and UKBB cohorts (~0.004% of all obese individuals), as well as in one obese patient from the PMMO cohort (~0.9% of all obese individuals).

The deletion frequencies in obese sub-populations from the NFBC1966 cohort are similar to the ones reported in previous studies, which are between ~0.5% and ~1% (Kumar *et al.*, 2008, Marshall *et al.*, 2008, Weiss *et al.*, 2008, Walters *et al.*, 2010). In fact, I have replicated the frequency reported by Walters *et. al.* (2010), since the author also included the NFBC1966 cohort among their population cohorts (Walters *et al.*, 2010). The prevalence of the proximal 16p11.2 heterozygous deletion among the obese patients (PMMO) was ~0.87%, although an overestimation could be the case due to obesity ascertainment in the patient recruitment. However, the frequency that I reported for the UKBB cohort could be an underestimation on several grounds; 1) individual screening of nearly 500,000 signal intensity plots was both time

consuming and not possible to be conducted in a limited time, 2) the participant recruitment for the UKBB cohort was not ascertained for any obesity status, unlike the study designs from the previous reports (Kumar *et al.*, 2008, Marshall *et al.*, 2008, Weiss *et al.*, 2008, Walters *et al.*, 2010) and it is not fully representative of the BMI or socio-economic status (SES) distribution of the UK general population (Castro *et al.*, 2015), 3) the UKBB could be subjected to sample selection bias because of its lower age restriction and individual consents – Since it is expected to be deleted for people with intellectual disability, language-related disability or severe psychiatric disease or severe autism, these people, in addition to the younger people, might not be recruited due to the lacking capacity for consent (Ge *et al.*, 2018). A total of 4 carriers (1 from NFBC1966; 2 from UKBB; 1 from PMMO) were morbidly obese.

I also have confirmed the presence of the relatively smaller, ~200-kb heterozygous deletion of the 16p11.2 region in the three cohorts. Overall, one individual was confirmed to have the distal 16p11.2 deletion from the PMMO and UKBB cohorts, while 4 NFBC1966 individuals were confirmed to carry the *SH2B1*-containing, heterozygous deletion. The deletion carrier from the PMMO cohort was obese at the time of recruitment (BMI=38.4 kg/m$^2$), while the carrier from the UKBB cohort was morbidly obese (BMI=40.6 kg/m$^2$). Out of the 4 carriers in the NFBC1966 cohort, only one of them was obese at 31 years of age. The other carriers were overweight.

In this study, the overall frequency of the ~200-kb deletion among the obese individuals was ≤0.2% (1 in 428 for NFBC1966; 1 in 79,096 for UKBB). Bochukova, *et. al.* (2010) has reported the prevalence of the distal 16p11.2 deletion was 0.41% (in severe early-onset obese patients) and 0.03% (in controls with European ancestry) respectively (Bochukova *et al.*, 2010). My case/control design revealed that the distal 16p11.2 deletion was significantly enriched in obese patients (cases) as compared to the general UK population (controls), in

which the prevalence was 0.8% and 0.0003% respectively. In comparison to the findings by Bochukova, *et. al.* (2010) (Bochukova *et al.*, 2010), I postulate that the prevalence in this study was over-estimated in the obese patients due to obesity ascertainment in the PMMO cohort. In addition, the obese patient with the ~200-kb heterozygous deletion was not morbidly obese; hence I could not make a comparable conclusion regarding the prevalence of the distal 16p11.2 deletion in morbid obesity. In contrast, the prevalence of the distal 16p11.2 deletion in the controls may have been under-estimated due to possibilities that have been discussed earlier in this section. Additionally, stratifying the participants from the control group into different classes of obesity revealed that the prevalence of the ~200-kb deletion among morbidly obese controls was 0.02% (1 in 6,100 participants), which is comparable to the prevalence reported by Bochukova, *et. al.* (2010).

### 7.6.3. Cognitive defects as a comorbidity of obesity

One of the proximal 16p11.2 deletion carriers (from PMMO) was identified to exhibit another comorbidity of obesity (psychiatric disorder), while the data on such comorbidity were not available for the other carriers at the time of writing for this thesis. This is consistent with previous reports on the strong relationship between obesity and cognitive/mental disorders (Kumar *et al.*, 2008, Walters *et al.*, 2010). A study on children with various chronic conditions shows that individuals with autism, learning difficulties, and attention deficit-hyperactivity disorder (ADHD) have increased risk of obesity (Chen *et al.*, 2010). In addition, extreme obesity patients are highly likely to display cognitive dysfunction (Boeka and Lokken, 2008). Despite that, the heterozygous ~600-kb deletion has a variable penetrance, since it has been frequently observed in the control populations (Bijlsma *et al.*, 2009, Kumar *et al.*, 2008, Weiss *et al.*, 2008), morbidly obese patients (Bochukova *et al.*, 2010, Jacquemont *et al.*, 2011,

Walters *et al.*, 2010), as well as in the obese patients without any cognitive defects (Bijlsma *et al.*, 2009, Fernandez *et al.*, 2010).

Maillard, *et. al.* (2015) took another step further in investigating the relationship between psychiatric disorders and the ~600-kb deletion by observing the variation in different parts of the brain (Maillard *et al.*, 2015). The author suggests that the proximal 16p11.2 deletion is associated with various changes in the brain anatomy. While the association was dosage-dependent, the significant findings were independent of autism spectrum disorder (ASD) and schizophrenia (SZ) diagnosis (Maillard *et al.*, 2015). In essence, the study postulates that people who have their proximal 16p11.2 region deleted, would have higher volumes of grey and white matter tissues, as compared to the people with no deletion of the same region (Maillard *et al.*, 2015). These anatomical parts of our brain are believed to be involved in various cognitive functioning, particularly social, reward and language perceptions (Maillard *et al.*, 2015). Although there is evidence to support the relationship between obesity and cognitive impairments, the fundamental mechanisms and interplay between the two phenotypes and the proximal 16p11.2 region remains elusive and further investigations are needed to address this matter.

The distal ~200-kb deletion region is located between breakpoint 2 and breakpoint 3 (BP2-BP3) within the 16p11.2 region. This region contains several genes, including the *SH2B1* gene that encodes Src homology 2B adaptor protein 1. This protein is known for its involvement in the regulation of leptin-melanocortin signalling (Aerts *et al.*, 2015), which is one of the main signalling pathways for food intake and energy homeostasis. The relevance of the protein encoded by the *SH2B1* gene in obesity aetiology was supported in a mice study by Ren, *et. al.* (2007), in which the author showed that mice (on a high fat diet) with a disrupted *Sh2b1* gene (i.e. heterozygous deletion), exhibited obesity phenotype and severe insulin resistance (Ren

*et al.*, 2007). In addition, there is an increasing evidence to support the notion that deletions of this *SH2B1*-containing region have significant implications on obesity from previous studies (Bachmann-Gagescu *et al.*, 2010, Bochukova *et al.*, 2010), as well as from our study. However, the obesity phenotype could possibly be expressed as a secondary phenotype in various neurodevelopmental and cognitive disabilities, such as in ASD, DD or ID. Bachmann-Gagescu, *et. al.* (2010) argued that their findings on the association of the distal 16p11.2 deletion with obesity, although in agreement with Bochukova, *et. al.* (2010) (Bochukova *et al.*, 2010), their patient population was ascertained for various abnormal phenotypes (Bachmann-Gagescu *et al.*, 2010). The author concluded that the ~200-kb heterozygous deletion increases the risk for obesity, along with DD and ID (Bachmann-Gagescu *et al.*, 2010). Despite the findings, the relationship between the *SH2B1*-containing 16p11.2 deletion and developmental/cognitive defects warrants further investigation to characterise the phenotypic expressions of such cognitive dysfunctions in non-obese patients.

### 7.6.4. Structural instability in 16p11.2 region and obesity

The proximal 16p11.2 deletion has been reported to be mostly inherited, as reported by Walters, *et. al.* (2010) (Walters *et al.*, 2010), but it can arise *de novo*, possibly due to non-allelic homologous recombination that frequently occurred between the flanking segmental duplications (>99% sequence similarity) (Bijlsma *et al.*, 2009, Fernandez *et al.*, 2010, Walters *et al.*, 2010). The recurrent *de novo* presence may be attributable to the absence of linkage disequilibrium between the ~600-kb deletion region and other nearby variants. However, I could not investigate whether that is the case since no parental information available for the three study cohorts involved in this chapter.

Walters, *et. al.* (2010) have postulated that the deletion could be a frequent causal variant for obesity, since it has been repeatedly seen to be the most common genetic cause of obesity other than *MC4R* point mutations (Vaisse *et al.*, 1998, Walters *et al.*, 2010, Yeo *et al.*, 1998). That author also has indicated that the haploinsufficiency of at least one gene (or several genes) within the ~600-kb region, is highly likely to affect molecular pathways involved in the obesity development (Walters *et al.*, 2010). As I described earlier, the distal 16p11.2 deletion has also been implicated in obesity in many studies (Bachmann-Gagescu *et al.*, 2010, Bochukova *et al.*, 2010), as well in this study. Bochukova, *et. al.* (2010) argued that haploinsufficiency in *SH2B1* gene may possibly be the main driver that could essentially influence the phenotypic expression of the observed phenotype (Bochukova *et al.*, 2010). Interestingly, the proximal 16p11.2 deletion region has been suggested to have a direct contact with the distal 16p11.2 deletion region, despite the two regions being ~800 kb apart (Loviglio *et al.*, 2017). Recent epigenetics study on long-range chromatin contacts shows a strong evidence that the proximal and distal 16p11.2 regions are involved reciprocally in a complex chromatin loop structure (Loviglio *et al.*, 2017). The looping causes these two regions to be in contact with each other, thereby affecting the regulations and the expressions of the genes located in the two regions in a coordinated manner (Loviglio *et al.*, 2017). They have also shown that the chromatin interaction concurrently altered the differential expressions of some of the genes. This could mean that any aberrations on the chromatin interaction could result in the manifestation of obesity phenotype (Loviglio *et al.*, 2017). Despite the promising observations, further investigations are needed to address specific questions regarding the functional aspects of the interaction and interplay in different modified models of such chromatin crosstalk. Nonetheless, this has shed light on the potential use of chromatin

contacts as a mechanism to investigate physical interactions of two loci that are significantly associated with the same phenotype.

# CHAPTER 8. Discovery of Rare CNVs in Morbid Obesity

## 8.1. Introduction

This chapter describes the identification of large rare CNVs potentially associated with human obesity in the European general population, as represented by the Northern Finnish Birth Cohorts of 1966 and 1986 (NFBC1966 and NFBC1986).

The discovery stage analysis was carried out using the cleaned CNV catalogue as a proxy-CNV list (refer to **Chapter 5.5.1 NFBC1966 cohort** and **Chapter 5.5.2 NFBC1986 cohort**) for CNV association tests with increasing permutations (10,000, 50,000 and 100,000) on BMI. The suggestive top hits that survived all the permutation tests were selected, and their presence was confirmed using visual inspection of signal intensity plots (LRR and BAF). If the number of confirmed CNV carriers was sufficient for meaningful comparison purposes, the suggestive top hits were statistically analysed using mean comparisons to identify their putative effects on BMI and WHR.

Low frequency CNV calls (rare and singletons) were also recorded to identify extremely rare CNVs potentially contributing to the obesity phenotype (i.e. a potentially damaging large deletion/duplication in one morbidly obese person) which might later be followed up in other populations.

Although there are common CNVs across the genome, rare CNVs have also been actively researched for phenotypic effects related to obesity. For example in **Chapter 7**, I presented the analysis of two non-overlapping, rare 16p11.2 deletion variants associated with BMI and obesity. These deletion variants which cause the 16p11.2 Deletion syndromes (proximal and distal) have been significantly associated with body weight and implicated in obesity (Bochukova *et al.*, 2010, Walters *et al.*, 2010).

Nonetheless, the phenotypic implications of other rare CNVs are still relatively under-explored in the general population, possibly due to those CNVs being singletons. For this reason, clinical databases of deletions and duplications, such as DECIPHER (DatabasE of genomiC varIation and Phenotype in Humans using Ensembl Resources) (Firth *et al.*, 2009) have been instigated to allow data-sharing across the world. In this way, the phenotype of rare CNVs may be validated as other carriers are identified and characterised for clinical characteristics. For example, only two isolated cases were reported to harbour the 17p13.3 telomeric duplication syndrome encompassing the tumour protein p53 (*TP53*) gene, while another duplication variant of the 17p13.3 telomeric duplication syndrome were more frequently reported. The former has been found in patients with ID and obesity (Belligni *et al.*, 2012, Kuroda *et al.*, 2014), while the latter has been previously implicated in ID (Bruno *et al.*, 2010). Similarly, the identification of possibly damaging, syndromic CNVs may increase our understanding in the aetiology of obesity.

## 8.2. Aims of this chapter

The first aim of the work described in this chapter was to identify the presence of large CNVs that are associated with obesity-related traits (i.e. BMI and WHR) in individuals from two birth cohorts; NFBC1966 and NFBC1986. Based on the findings, the next aim was to describe the syndromic CNVs that were identified in morbidly obese carriers. The aims were addressed by interrogating large CNV calls (length≥100kb) predicted by a CNV calling algorithm tool (as described in **Chapter 3.9 CNV calling** and **Chapter 3.10 CNV quality controls**), running permutation-based CNV association tests in the discovery stage to identify potential genomic imbalances that have significant effects on BMI. This was followed by visual inspection of signal intensity plots to confirm the presence of such CNVs, and comparisons of the mean values of obesity-related traits in the carriers to those in non-carriers within the same study cohort. The aims were achieved by addressing the following questions:

1. *Are there any predicted CNV calls (genome-wide) in the clean CNV catalogue, which are significantly associated with BMI under permutation-based tests?*

I answered this question by executing CNV association tests under 100,000 permutations on the large CNV calls for each study cohort. All significant CNV probes were cross-checked in the CNV catalogue to obtain the relative CNV coordinates that have been predicted in the work described in **Chapter 5.5.1 NFBC1966 cohort** and **Chapter 5.5.2 NFBC1986 cohort**.

2. *Are the predicted genomic imbalances identified from permutation-based tests truly present in the individuals flagged as potential carriers?*

I answered this question by visually inspecting the Log R Ratio (LRR) and B Allele Frequency (BAF) scatter plots spanning the predicted CNV coordinates. These plots were generated using a script kindly provided by Dr Robin Walters (Oxford). Individuals with a validated genomic imbalance were flagged as confirmed carriers.

3. *Do any of the confirmed large genomic imbalances associate with obesity-related phenotypes (i.e. BMI and WHR)?*

I answered this question by evaluating the relationships between each genomic imbalance and the obesity traits using statistical comparisons of means (i.e. independent t-tests and ANOVA). However, this step was highly dependent on the number of CNV carriers.

4. *Are there any known genomic syndromes caused by the CNVs identified in this study?*

I answered this question by consulting public online databases for relevant genomic syndromes and for other CNVs previously implicated in the same syndromes.

## 8.3. General methods used in this chapter

| Section | Method name (Described in Chapter 3) | Chapter 4 | Chapter 5 | Chapter 6 | Chapter 7 | Chapter 8 |
|---|---|---|---|---|---|---|
| **3.2** | Biological sample collection | ✓ | ✓ | ✓ | ✓ | ✓ |
| **3.3** | Genomic DNA extraction | ✓ | ✓ | ✓ | ✓ | ✓ |
| **3.4** | Obesity-related phenotypic measurements | ✓ | ✓ | ✓ | ✓ | ✓ |
| **3.5** | SNP genotyping | ✓ | ✓ | ✓ | ✓ | ✓ |
| **3.6** | Quality controls | ✓ | ✓ | ✓ | ✓ | ✓ |
| **3.7** | Population structure adjustments | ✓ | ✓ | ✓ | ✓ | ✓ |
| **3.8** | Principal component analysis | | | | | ✓ |
| **3.9** | CNV calling | | ✓ | | ✓ | ✓ |
| **3.10** | CNV quality controls | | ✓ | | ✓ | ✓ |

## 8.4. Specific methods

### 8.4.1. Exclusion of small CNV calls from the CNV catalogue

The cleaned CNV catalogue for each study cohort (refer to **Chapter 5, results section**) was filtered to include only the CNV calls with predicted lengths of at least 100 kb. The filter was conducted manually to produce a CNV catalogue that contained only large CNV calls for each study cohort. Singletons were included for the purpose of detecting rare genomic imbalances that could be present in one individual of a study cohort.

### 8.4.2. CNV list and file conversions

All file conversions were conducted using a script by PennCNV (May 2014 version) tool (Wang *et al.*, 2007). For each study cohort, a set of files were prepared manually, as below:

(a) PennCNV-formatted `.cnv` file containing the filtered CNV calls

(b) PLINK-formatted `.fam` file containing sex and family information

(c) PLINK-formatted `.pheno` file containing BMI, WHR (continuous values) and obesity status (binary-coded)

(d) PLINK-formatted `.covar` file containing sex (binary-coded), genotyping batch (continuous), and the first 11 principal components (PC1 to PC11)

The PennCNV-formatted `.cnv` file was converted into its PLINK-readable CNV file by executing the command below in PennCNV directory:

```
./penncnv_to_plink.pl -i LargeNVcatalogue.cnv -o LargeCNVcatalogue.plink.cnv
```

The resulting file was used to create a PLINK-formatted .cnv.map file that contained the genomic positions of all CNV probes in the CNV list. The mapping of the probes was carried out by executing the command below in PLINK, which generated the PLINK-formatted `.cnv.map` file:

```
./plink --cnv-list LargeCNVcatalogue.plink.cnv --cnv-make-map --out
LargeCNVcatalogue.plink
```

The PLINK-formatted `.cnv` file, `.cnv.map` file, `.fam` file, `.pheno` file, and `.covar` file were used in the permutation-based CNV association analyses for each study cohort.

### 8.4.3. Permutation-based CNV association tests

Obesity was classified as individuals with BMI≥30 kg/m$^2$ in an adult population (i.e. NFBC1966), and teens with BMI above the 95$^{th}$ percentile (according to the clinical growth chart) in an adolescent population (i.e. NFBC1986). Individuals were considered as morbidly obese if their BMI reached 40 kg/m$^2$ (for adults) or reached the 99$^{th}$ percentile (for adolescents).

For each study cohort (with a set of files as described in the previous section), CNV association analysis was carried out by performing 100,000 permutation tests to identify specific loci that were significantly associated with BMI. The permutation test was performed according to maxT procedures to control family-wise error rates, adjust for multiple testing and minimise false positive discoveries (Dudoit *et al.*, 2003). Principal component analysis was conducted for each study cohort to identify principal components (PC) that could be used as covariates. In addition to the PCs, sex and genotyping batch were also used as covariates. A minimum of two consecutive probes with empirical p values < 0.005 were considered as

significant to suggest the presence of a large genomic imbalance associated with BMI. Only rare, large BMI-associated genomic imbalances in which the carriers were obese/morbidly obese (according to the classification mentioned earlier), were selected for validation. All maxT permutation tests were carried out in PLINK Version 1.07, according to the developer's protocol (Purcell *et al.*, 2007). Carriers of such genomic imbalances were cross-checked in the CNV catalogue to obtain the predicted chromosomal start and end coordinates.

### 8.4.4. Visual inspections of signal intensity plots

The LRR and BAF plots for the carriers of the large genomic imbalances identified from the permutation tests, were visually inspected to validate the presence of such imbalances. A heterozygous duplication is present if the LRRs for the SNPs within the putative genomic imbalance region was higher than the normalised values (scattered around zero) along the x-axis, while the BAFs for the same SNPs scattered in between 0 and 0.5, and 0.5 and 1. Since 0 and 1 represent homozygosity (i.e. AA and BB genotypes) for that particular SNP and location, and 0.5 represents heterozygosity (i.e. AB genotype), a heterozygous duplication of the SNPs spanning the genomic region can be seen as having an additional copy in the BAF plot (i.e. either AAB or ABB genotypes). As mentioned in the previous chapter (refer to **Chapter 7.4.5: Visual inspections of signal intensity plots of potential 16p11.2 deletion carriers**), a heterozygous deletion is present when the LRRs were lower than the normalised values along the x-axis, and the BAFs for the same SNPs scatter either around 0 or 1.

### 8.4.5. Statistical analyses

The distributions of the residuals of WHR and BMI were compared to a normal distribution as frequency histograms and normal P-P plots. The means of WHR and BMI were compared according to sex and batch using Analysis of Variance (ANOVA) tests. The effects of sex on obesity status (BMI≥30 kg/$m^2$), was determined using Pearson's chi-square tests. Bivariate correlation analyses were conducted between WHR, BMI, sex and genotyping batch. A p-value < 0.05 was considered statistically significant. All statistical analyses were performed using IBM SPSS Statistics for Windows Version 24.0 (Released 2016; IBM Corp., Armonk, New York. USA) and Rstudio version 1.1.419 (RStudio, Inc., Boston, MA, USA) (RStudio, 2015).

### 8.4.6. Public online database search

Upon confirmation of the presence of each top hit in its corresponding carrier, I carried out a comprehensive search of the large genomic imbalances on public online databases. The main source used was the data generated by the DECIPHER community. A full list of the contributed centres for this data generation is available at http://decipher.sanger.ac.uk. Funding for the DECIPHER project was provided by the Wellcome Trust (Firth *et al.*, 2009). In addition, a search was also conducted in the Online Mendelian Inheritance in Men (OMIM) database (National Center for Biotechnology Information (NCBI), 1995). The findings from the online databases were compared with the top hits and their carriers in terms of the genomic coordinates (hg19) and other comparable phenotypic or clinical data, when available.

# 8.5. Results

## 8.5.1. Study cohort overview

In this chapter, the results from **Chapter 5 – Inferring Copy Number Variants** were used. To recap, the cleaned CNV catalogue for the NFBC1966 cohort consists of 25,878 CNVs in 4,905 adults, while the cleaned CNV catalogue for the NFBC1986 cohort contains a total of 37,670 CNVs in 3,227 adolescents. These individuals survived all genotyping QCs and population structure adjustments from **Chapter 4 – Genetic Data Processing**.

## 8.5.2. Distributions of obesity-related traits

Histograms and P-P plots were constructed to visualise the overall distributions of the obesity-related phenotypes (ie. BMI and WHR). The distributions of both phenotypes for the NFBC1966 cohort were normal, as presented in the previous chapter (refer to **Chapter 7.5.1 Distributions of obesity-related traits**).

The distribution of the standardised residuals for both obesity traits in the NFBC1986 cohort are shown in **Figure 8- 1**. The mean residual for BMI was zero (range= -7.77 to 50.05), while the mean residual for WHR was also zero (range= -0.77 to 0.41). Both obesity traits were identified to have an approximately normal distribution and were treated as such.

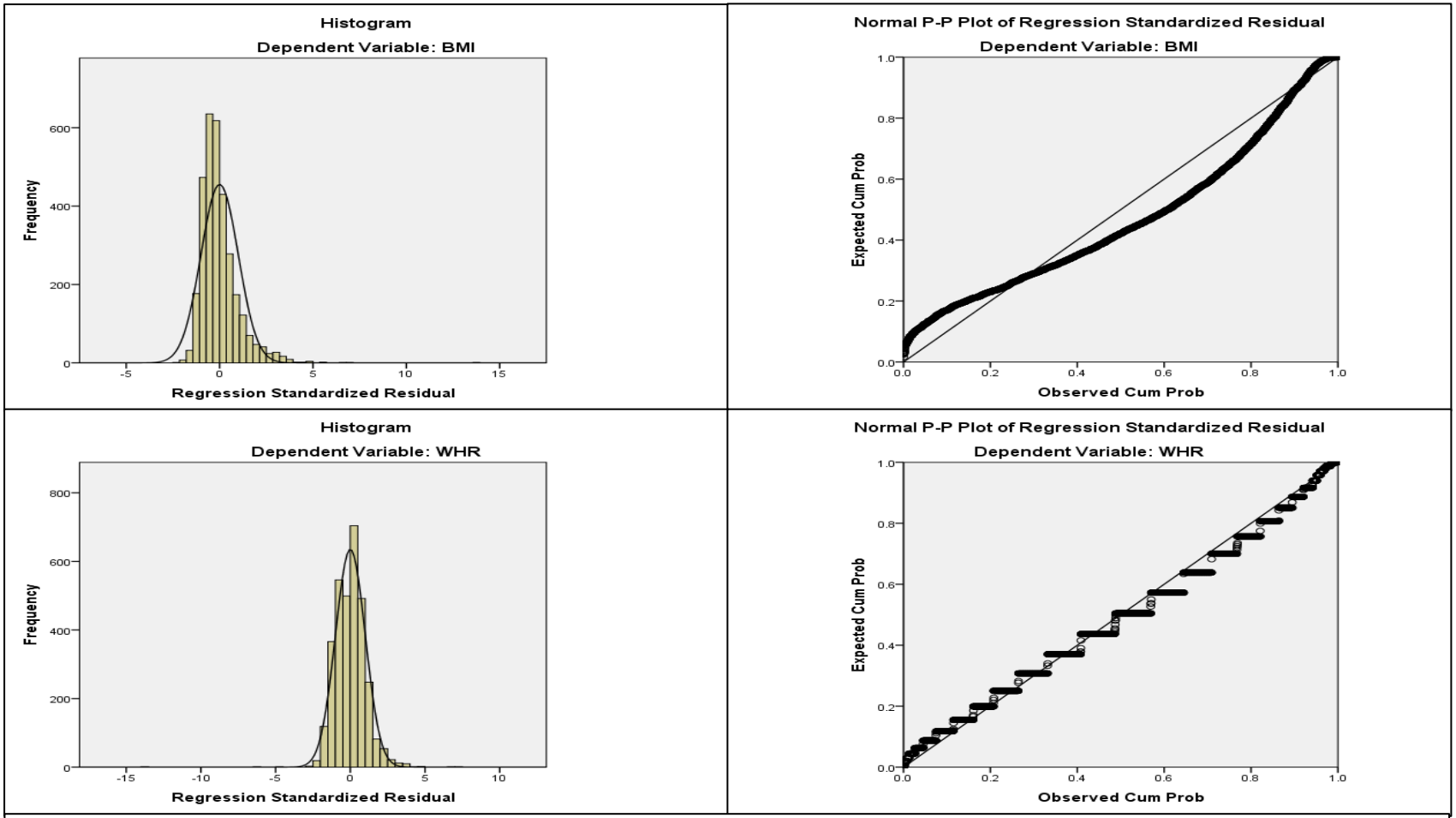


**Figure 8- 1:** The distributions of standardised residuals of obesity traits, BMI (n=3,195) and WHR (n=3,184), in the NFBC1986 cohort.

BMI = body mass index, WHR = waist-to-hip ratio.

### 8.5.3. Sex and batch effects

The mean for each obesity-related phenotype was compared by sex and genotyping batches, using independent t-tests (for sex) or ANOVA (for batch). Chi-square tests were also performed for comparisons between obese and non-obese individuals. The mean comparisons for the NFBC1966 cohort have been presented in the previous chapter (refer to **Chapter 7.5.2 Sex, age and batch effects**). Briefly, different genotyping batches were observed to have significant effects on BMI and WHR, while males had significantly higher mean BMI and mean WHR compared to females in the NFBC1966 cohort.

For the NFBC1986 cohort, independent t-tests revealed that there was no significant difference in mean BMI between males and females ($p=0.432$). However, as expected, the mean WHR was significantly different between males and females ($p=2.64 \times 10^{-158}$). I also found that the means of both obesity-related phenotypes were not significantly different between the genotyping batches ($p>0.5$). Age was not a factor in this cohort since, by definition, all participants were born in the same 12-month period. Chi-square tests were carried out to identify any significant effects of sex or batch between the obese and non-obese people within this cohort. The prevalence of obesity was not significantly different between males and females in this adolescent cohort (Pearson's Chi-squared $p=0.376$). Similarly, different genotyping batches did not have a significant effect on the obesity status of NFBC1986 individuals (Pearson's Chi-squared $p=0.626$). Although non-significant, sex and genotyping batch were included as covariates in association analyses to avoid any residual confounding effect.

### 8.5.4. Principal components as covariates

Principal Component Analysis (PCA) was conducted for each study cohort on the first 20 PCs to reduce the dimensions needed as covariates. **Figure 8- 2** and **Figure 8- 3** show the eigenvalues plotted against its PC in the NFBC1966 and NFBC1986 respectively. The PCs were included as covariates up until the point where there was no difference in eigenvalues in the consecutive PCs, as shown by the red points in **Figure 8- 2** and **Figure 8- 3.** The first 11 PCs were included as covariates in NFBC1966, while the first 10 PCs were chosen as covariates in NFBC1986.

### 8.5.5. Obesity-related phenotype correlation analysis

Bivariate correlation analysis was performed to identify any significant correlations amongst the obesity traits, sex and batch. Traits correlations for the NFBC1966 cohort have been presented in the previous chapter (refer to **Chapter 7.5.3 Obesity-related phenotype correlation analysis**). In brief, BMI and WHR were significantly correlated,

For the NFBC1986 cohort, correlation analysis revealed that BMI and WHR were significantly correlated to each other (n=3,184; $p=3.98 \times 10^{-174}$; r=0.470) as expected.

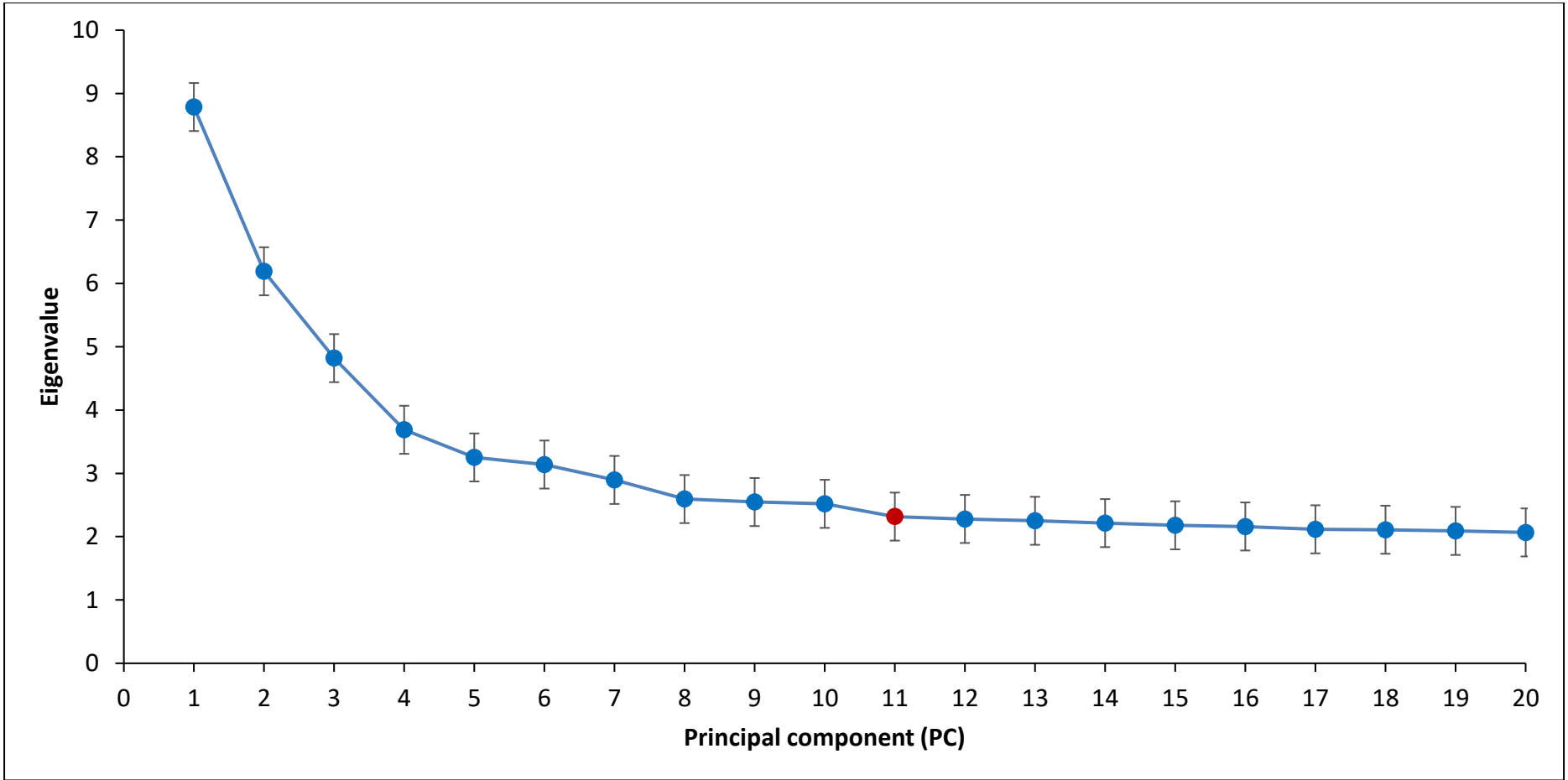


**Figure 8- 2:** A scatter plot of eigenvalues for the first 20 principal components (PC) determined from Principal Component Analysis (PCA) in the NFBC1966 cohort (n=4,905). Error bars represent standard errors per PC. The red point represents the maximum threshold for inclusion of PCs as covariates in downstream association analyses.

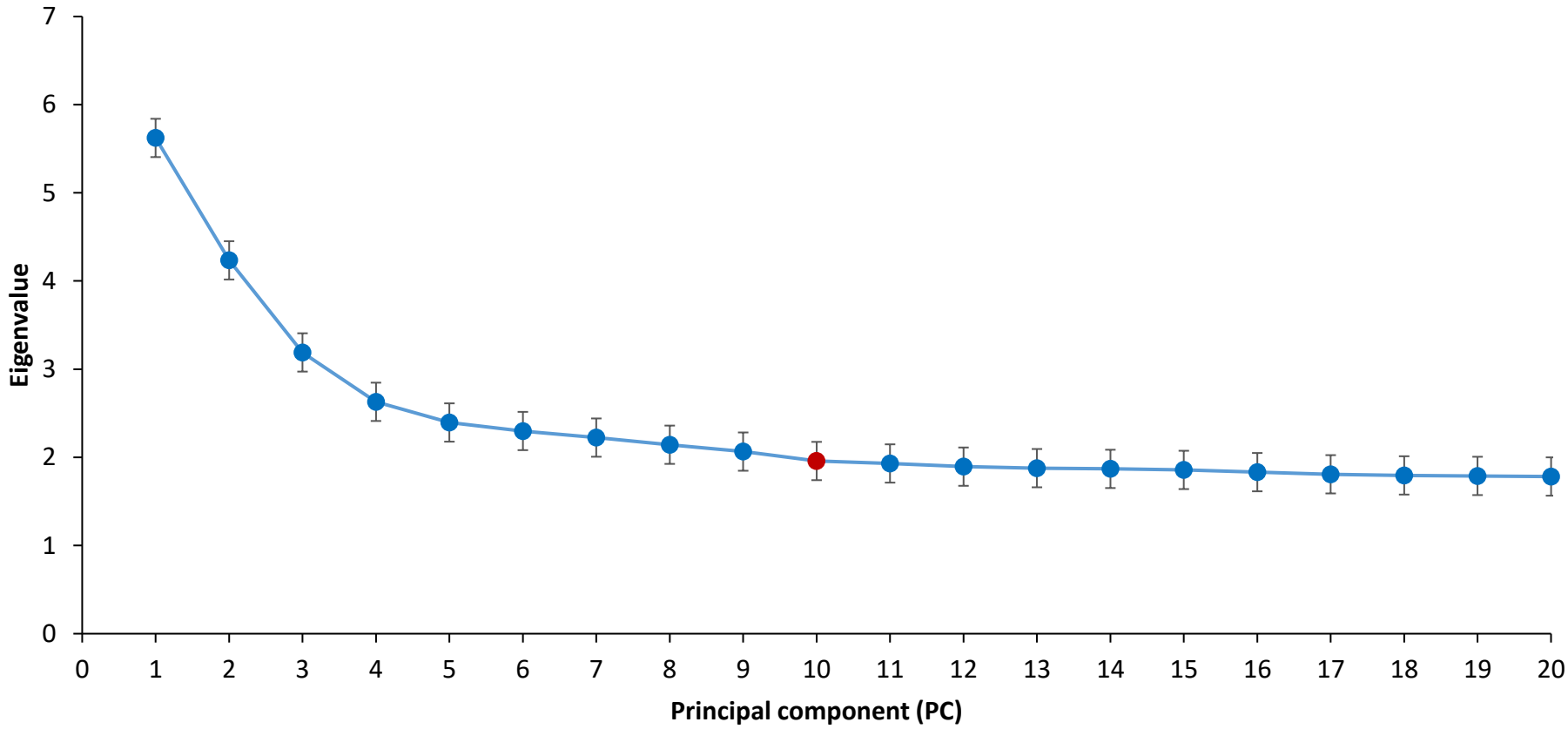


**Figure 8- 3:** A scatter plot of eigenvalues for the first 20 principal components (PC) determined from Principal Component Analysis (PCA) in the NFBC1986 cohort (n=3,227). Error bars represent standard errors per PC. The red point represents the maximum threshold for inclusion of PCs as covariates in downstream association analyses.

### 8.5.6. Discovery and validation of suggestive top hits

A cleaned CNV catalogue containing only large CNV calls (length>100,000 bp) was prepared for the NFBC1966 and NFBC1986 cohorts. These catalogues were converted into dataset in a PLINK-readable format accordingly. Sex, genotyping batch and PCs were included in the covariate file for use in the permutation tests in PLINK for the discovery of potential rare, large genomic imbalances associated with BMI.

**Table 8- 1** shows the suggestive top hits that survived the CNV association tests after 10,000, 50,000, and 100,000 maxT permutations. A minimum of 2 CNV probes within a large duplication in the 17p13.3-p11.2 region (hg19; chr17: 883,050-18,858,049; ~18 Mb) were detected to have a significant association with the BMI in a morbidly obese female of the NFBC1966 cohort ($p = 1.95 \times 10^{-3}$). Her BMI at 31 years of age, was 23.18 kg/m$^2$ higher than the mean BMI of the non-carriers of the 17p13.3-p11.2 duplication variant (**Table 8- 1**). The presence of the 17p13.3-p11.2 duplication was validated by visually inspecting the signal intensity plot (the LRR and BAF plots) of the potential carrier (**Figure 8- 4**). The duplication was seen to include the telomeric end of the 17p13.3-p11.2 region, hence had an increased length compared to the predicted length of the CNV call (confirmed duplication; hg19; chr17: 1-19,000,000; ~19 Mb).

Within the same adult cohort, I have identified another morbidly obese female carrying another suggestive top hit (deletion variant) in the 5p14.3 region (hg19; chr5: 18,852,476-20,455,966; ~1.6 Mb). The association of the 5p14.3 deletion variant and BMI was significant at $p = 1.48 \times 10^{-3}$ (**Table 8- 1**). The 5p14.3 deletion carrier had a BMI of 50.06 kg/m$^2$ at 31 years old (25.41 kg/m$^2$ higher than the mean BMI of non-carriers). The putative deletion variant was confirmed to be present in the female carrier, based on the signal intensity plot shown in **Figure 8- 5**.

| Study cohort | CNV | Chr | Predicted by the PennCNV (CNV catalogue) | | | | CNV carrier information | | | | Mean BMI of the non-carriers, kg/m$^2$ | Mean BMI difference, kg/m$^2$ | Empirical p-value |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Start position, Mb | End position, Mb | Length, Mb | Minimum conf score | Sex | Age | Obesity status | BMI, kg/m$^2$ | | | |
| NFBC1966 | 17p13.3-p11.2 Duplication (1-copy) | 17 | 0.88 | 18.86 | 17.97 | 29.21 | female | 31 | Morbidly obese | 47.83 | 24.65 | 23.18 | **1.95 x 10$^{-3}$** |
| | 5p14.3 Deletion (Heterozygous) | 5 | 18.85 | 20.46 | 1.60 | 255.97 | female | 31 | Morbidly obese | 50.06 | 24.65 | 25.41 | **1.48 x 10$^{-3}$** |
| NFBC1986 | 22q11.2 Duplication (1-copy) | 22 | 18.89 | 21.46 | 2.58 | 103.38 | male | 16 | Morbidly obese | 36.34 | 21.35 | 14.99 | **4.21 x 10$^{-3}$** |

**Table 8- 1:** Potential large syndromic CNVs that were considered as suggestive top hits by the CNV association tests based on 10,000, 50,000 and 100,000 permutations. Each SNP (as a CNV probe) that spanned any given CNV call from the CNV catalogue, was permuted (for a maximum of 100,000 times) on the BMI of each CNV carrier. Sex, genotyping batch and principal components were used as covariates. A minimum of two consecutive CNV probes with empirical p values < 0.005 were considered significant to suggest top hits. CNV = copy number variant, NFBC1966 = Northern Finland Birth Cohort 1966, NFBC1986 = Northern Finland Birth Cohort 1986, CNV = copy number variant, Chr = chromosome, Mb = mega-base pairs, conf = internal confidence from PennCNV, BMI = body mass index, Empirical p-value = P-value of the CNV association test based on 100,000 maxT permutations on BMI.

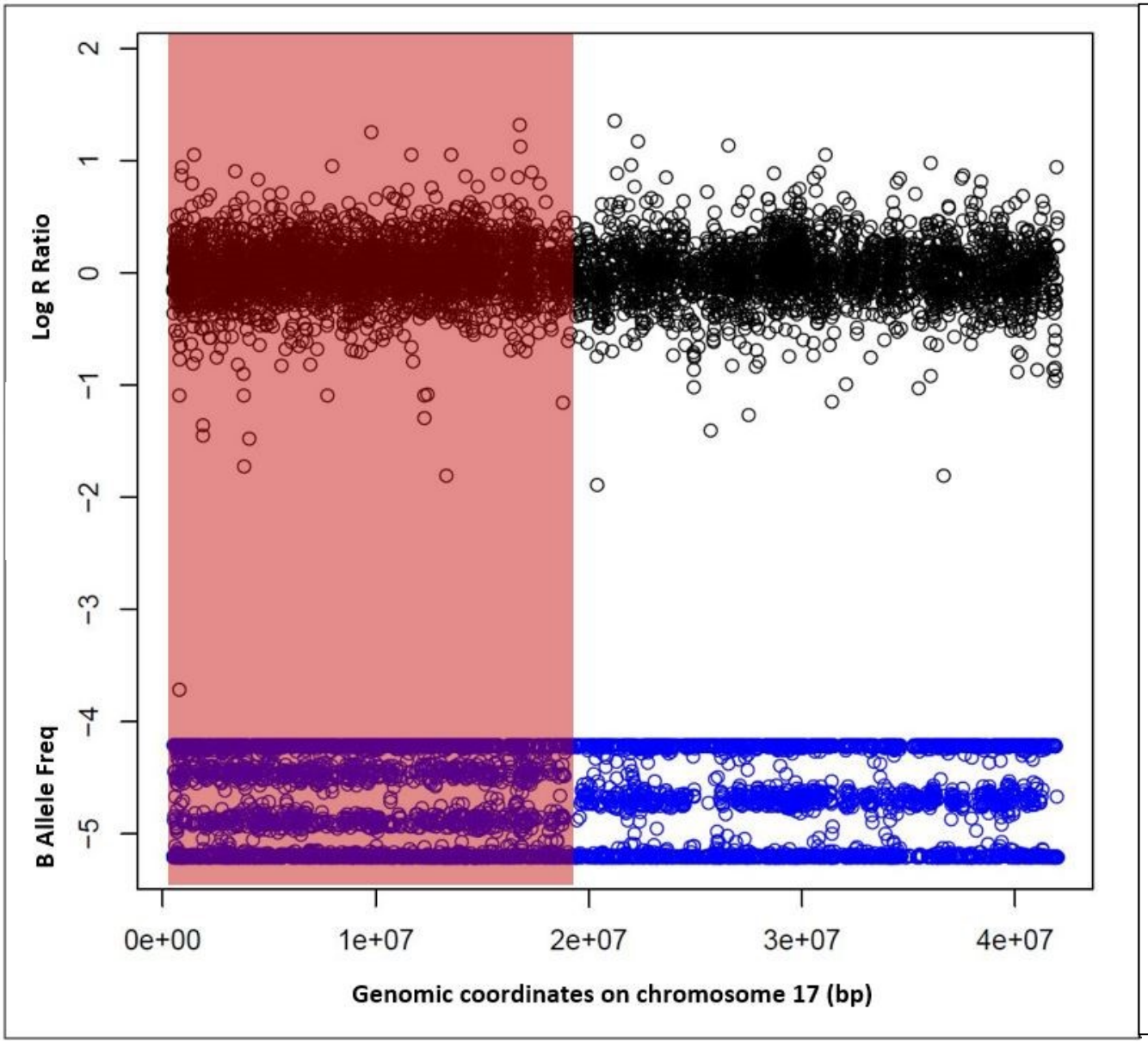


**Figure 8- 4:** Signal intensity plot of the 17p13.3-p11.2 duplication carrier in the NFBC1966 cohort (n=1). The plot consists of Log R Ratio (LRR) values (top) and B Allele Freq (BAF) values (bottom), plotted against the genomic coordinates of SNPs on chromosome 17 (hg19). The red shade displays the duplication (~19 Mb) of the 17p13.3-p11.2 region. The female carrier was morbidly obese at 31 years old (BMI = 47.83 kg/m$^2$). NFBC1966 = Northern Finland Birth Cohort 1966, BMI = body mass index.

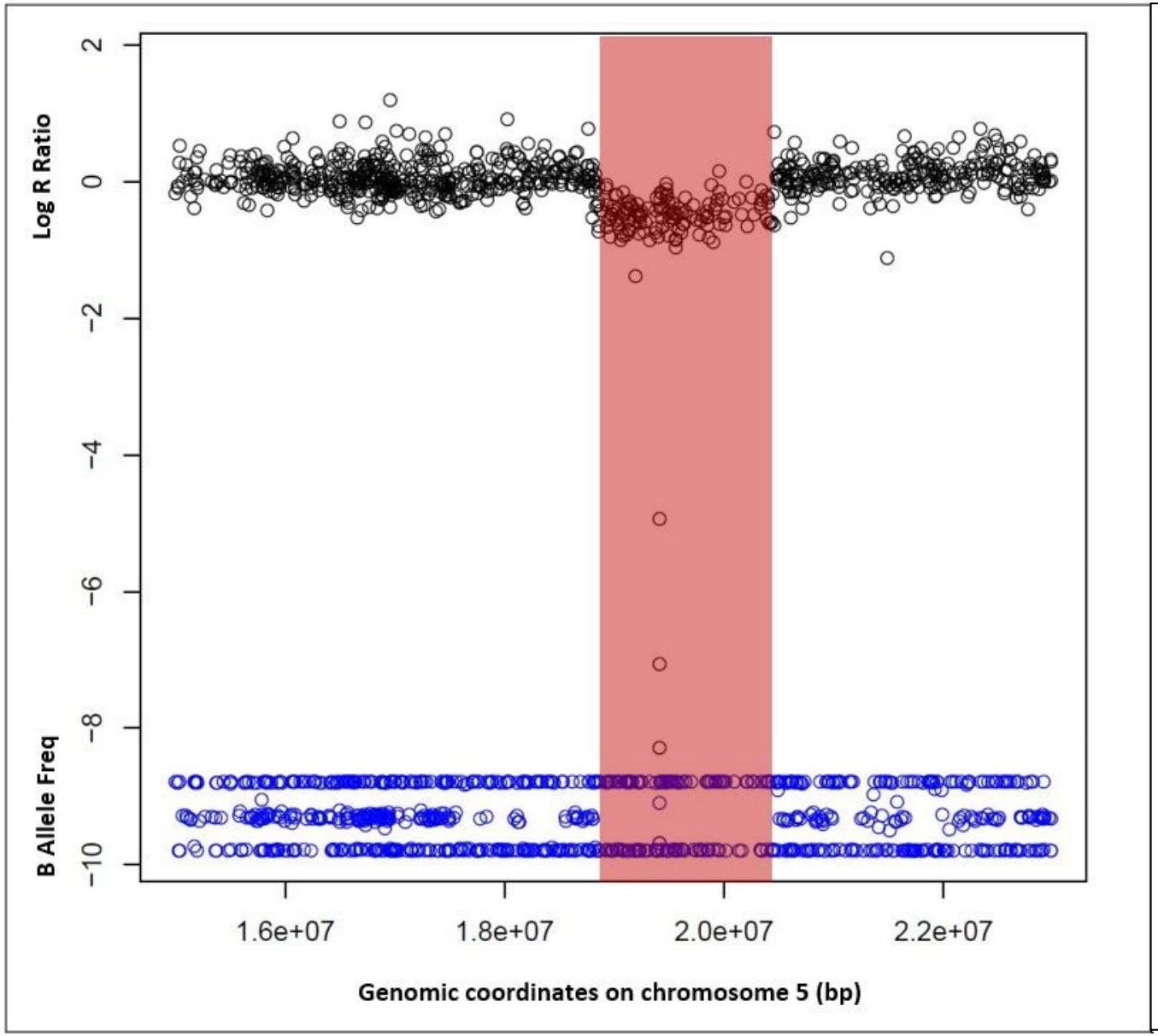


**Figure 8- 5:** Signal intensity plot of the 5p14.3 deletion carrier in the NFBC1966 cohort (n=1). The plot consists of Log R Ratio (LRR) values (top) and B Allele Freq (BAF) values (bottom), plotted against the genomic coordinates of SNPs on chromosome 5 (hg19). The red shade displays the deletion (~1.6 Mb) of the 5p14.3 region. The female carrier was morbidly obese at 31 years old (BMI= 50.06 kg/m$^2$). NFBC1966 = Northern Finland Birth Cohort 1966, BMI = body mass index.

In the CNV catalogue of the NFBC1986 cohort, a large duplication (~2.6 Mb) in the 22q11.2 region (hg19; chr22: 18,886,915-21,463,730) was identified to be significantly associated with BMI ($p = 4.21 \times 10^{-3}$) based on the permutation-based CNV association test (**Table 8- 1**). The 22q11.2 duplication carrier was a morbidly obese male, whose BMI was 14.99 $kg/m^2$ higher than the mean BMI of the other adolescents without the predicted duplication variant (**Table 8- 1**). The presence of the large 22q11.2 duplication was subsequently confirmed from the signal intensity plot of the carrier (**Figure 8- 6**).

I would like to emphasise that the permutation test results were reported here for the purpose of describing the discovery of such singletons from both cohorts. Therefore, the significance of the findings did not give any meaningful impact towards the suspected associative relationship between each of these CNV and BMI.

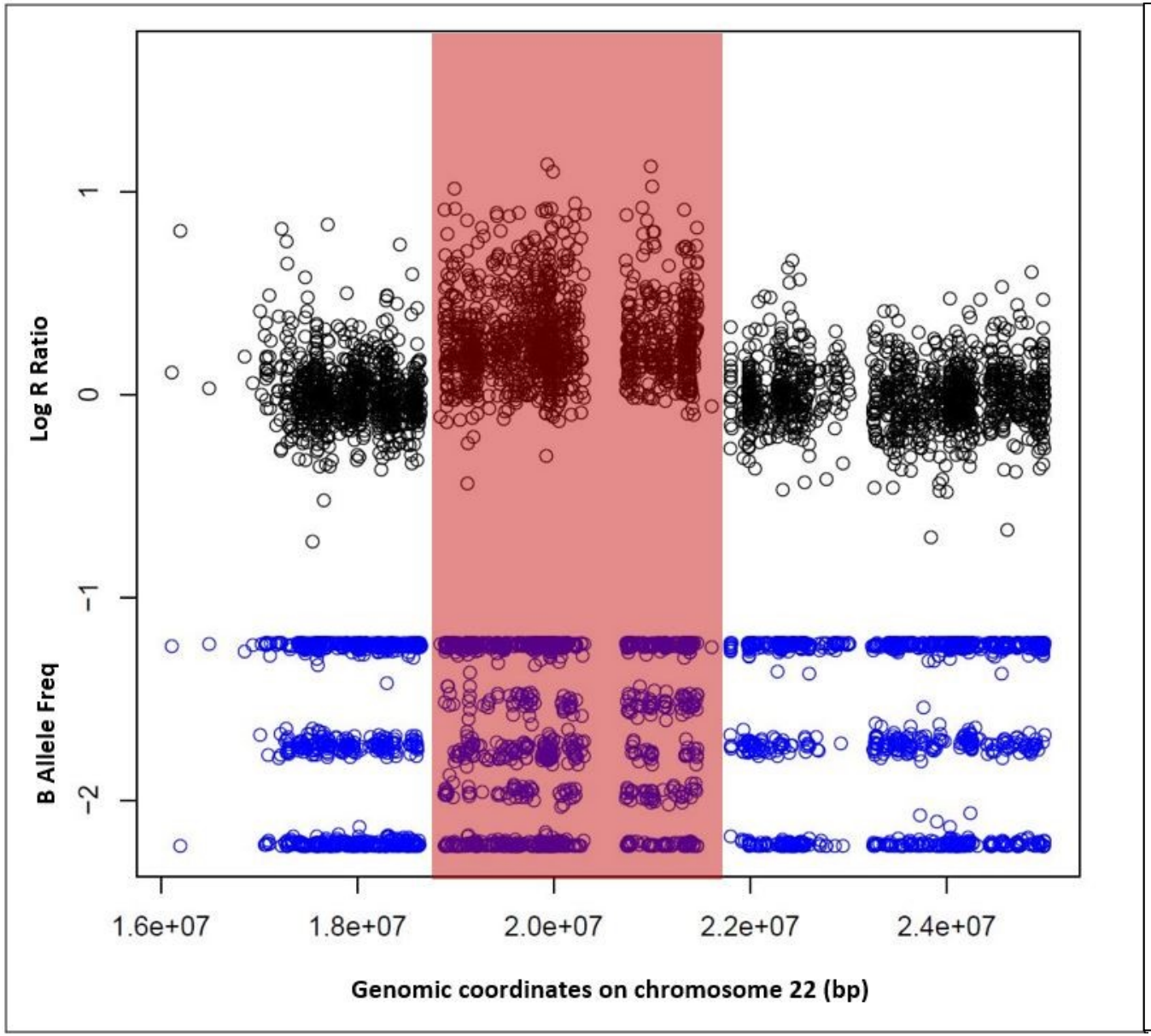


**Figure 8- 6:** Signal intensity plot of the 22q11.2 duplication carrier in the NFBC1986 cohort (n=1). The plot consists of Log R Ratio (LRR) values (top) and B Allele Freq (BAF) values (bottom), plotted against the genomic coordinates of SNPs on chromosome 22 (hg19). The red shade displays the duplication (~2.6 Mb) of the 22q11.2 region. The male carrier was morbidly obese at 16 years old (BMI= 36.34 kg/m$^2$). NFBC1986 = Northern Finland Birth Cohort 1986, BMI = body mass index.

### 8.5.7. Public online database search

The 17p13.3-p11.2 duplication region identified in this chapter, includes part of the critical regions for two genomic duplication syndromes on chromosome 17: Charcot-Marie-Tooth syndrome type 1A (CMT1A; OMIM #118220) and Potocki-Lupski syndrome (PLS; OMIM #610883) (**Figure 8- 7**). In addition, the 17p13.3-p11.2 duplication also covered the critical deletion region that causes Smith-Magenis syndrome (SMS; OMIM #182290) that contains the retinoic acid induced 1 gene (*RAI1*; OMIM *607642).

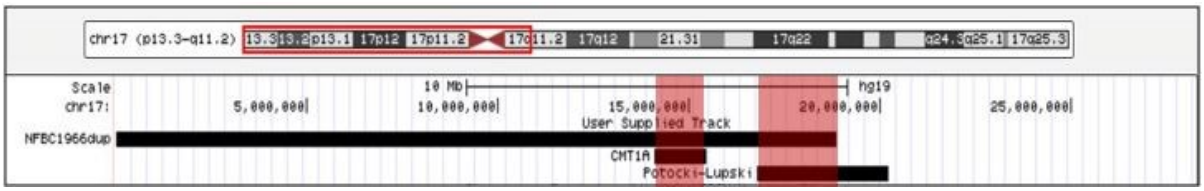


**Figure 8- 7:** A modified screenshot of the chromosomal region spanning the 17p13.3-p11.2 duplication from the UCSC Genome Browser (hg19; chr17:1-30,000,000). The 17p13.3-p11.2 duplication (NFBC1966dup) identified in this chapter overlapped the critical duplication regions that cause Charcot-Marie-Tooth syndrome type 1A (CMT1A) and Potocki-Lupski syndrome (shaded in red) (Firth *et al.*, 2009).

The CMT1A is characterised by various sensory and neural motor impairments, while Potocki-Lupski syndrome (PLS; also known as 17p11.2 duplication syndrome) is characterised by autism, low attention span, and hyperactivity. The corresponding deletion of the PLS critical region, along with any disruption of the *RAI1* gene (i.e. haploinsufficiency) can cause SMS, in which is characterised by obesity and hypercholesterolemia (Burns *et al.*, 2010, Lacaria *et al.*, 2012). In OMIM, this 17p13.3-p11.2 duplication which include the 'Chromosome 17p13.3, Telomeric, Duplication Syndrome' (17p13.3t.DupS; OMIM #612576), was associated

with the split-hand/foot malformation and the long bone deficiency-3 (SHFLD3). Patients with 17p13.3t.DS were characterised by various skeletal malformations (OMIM #612576), in addition to intellectual disabilities (ID) (Bruno *et al.*, 2010). However, the data on anatomical structure and psychiatric conditions were not available for the NFBC1966 individual carrying the larger 17p13.3-p11.2 duplication encompassing 17p13.3t.DupS region to make a comparable assessment.

As reported earlier in this thesis, a female adult with morbid obesity was identified as carrying a potentially pathogenic BMI-associated 5p14.3 deletion (~1.6 Mb). In DECIPHER, six patients were reported to carry deletions overlapping the 5p14.3 deletion region in this study, as shown in **Figure 8- 8**. The first patient (ID: 250008) was found to have a larger de novo 5p heterozygous deletion that spanned from 5p14.1 to 5p15.2 chromosomal regions, including the whole stretch of 5p14.3 and 5p15.1 regions (~14.7 Mb). The same patient exhibited head/neck abnormality and obesity. The second patient (DECIPHER ID: 262182) displayed head/neck abnormality and speech delays, while the third patient (DECIPHER ID: 257567) had abnormalities in the respiratory and digestive systems. The fourth patient (DECIPHER ID: 351341) was diagnosed with head/neck abnormality and ID, but not obesity. However, the phenotype for the fifth patient (DECIPHER ID: 269681) was unknown. In addition, the patient with the smallest deletion overlapping the 5p14.3 deletion region (DECIPHER ID: 270183), was reported to have a mild global developmental delay (DD). Interestingly, all six deletion variants on the 5p region covered a common gene; *CDH18* gene (**Figure 8- 8**).

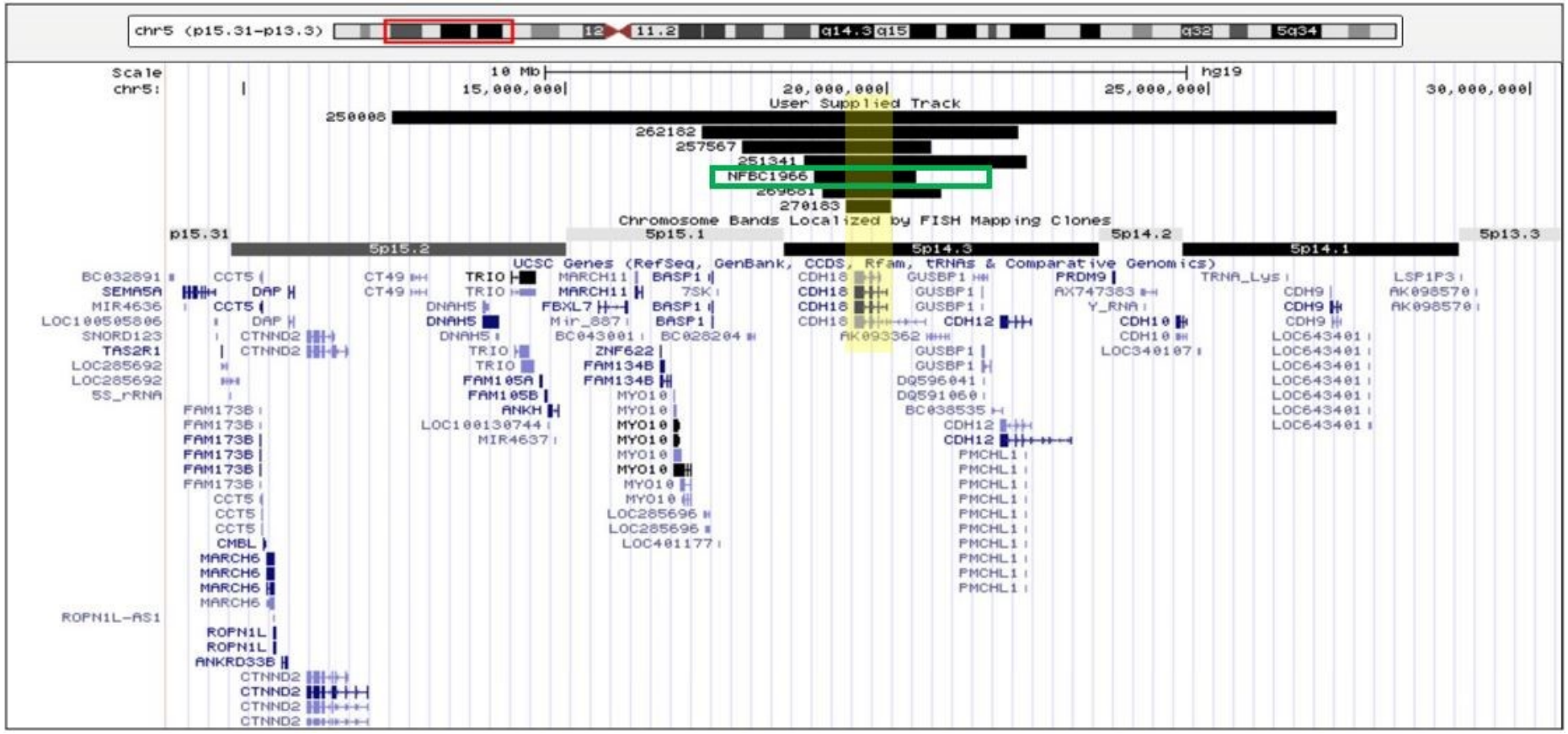


**Figure 8- 8:** A screenshot from the UCSC Genome Browser, of the 5p13.3 to 5p15.3 chromosomal region displaying the 5p14.3 deletion carrier (NFBC1966; green box), and patients carrying deletion variants overlapping the 5p14.3 deletion from the DECIPHER database. All overlapped deletion variants contained a common gene; *CDH18* gene (Yellow highlight). NFBC1966 = Northern Finland Birth Cohort 1966, *CDH18* = Cadherin 18.

Among the Northern Finnish adolescents, a morbidly obese male has been identified to carry a large 22q11.2 duplication (~2.6 Mb) that was associated with BMI in this study. In DECIPHER, I found a minimum of 162 patients with similar duplications, overlapping at least 70% of the 22q11.2 duplication variant described above. These patients have been categorised as patients of '22q11 duplication syndrome' in DECIPHER, characterised by ID, DD, speech delay, autism, cognitive defects, and in some cases, obesity. In OMIM, the 22q11 duplication syndrome was characterised by growth retardation, head/neck abnormalities, speech delay, learning disabilities, and mental retardation (OMIM #608363). These duplication variants encompassed approximately 78 known genes, in which a deletion of the corresponding region causes DiGeorge syndrome (OMIM #188400). DiGeorge syndrome can present a wide spectrum of phenotypes, including head/neck abnormalities, obesity, and several neurobehavioural complications such as schizophrenia, attention deficit disorder (ADD), bipolar disorder and speech delay. **Figure 8- 9** shows the genomic locations of the 22q11.2 duplication variant identified in this chapter, and the genomic regions implicated in DiGeorge syndrome and 22q11 duplication syndrome (22q11.DS). At the time of writing, the clinical diagnosis on genomic syndromes and psychiatric disorders of the morbidly obese adolescent (22q11.2 duplication carrier) was not available to make a comparable conclusion.

**Figure 8- 9:** A screenshot from the UCSC Genome Browser, displaying the location of the 22q11.2 duplication identified in a morbidly obese adolescent (NFBC1986; green box), and the genomic regions that have been implicated in DiGeorge syndrome (DiGeorge) and 22q11 Duplication syndrome (22q11.DS). NFBC1986 = Northern Finland Birth Cohort 1986.

## 8.6. Discussions

### 8.6.1. Results overview

This chapter presents the discovery of large, rare CNVs in morbidly obese individuals from two birth cohorts; the NFBC1966 cohort consisting adults at 31 years of age, and the NFBC1986 cohort comprising adolescents at 16 years of age.

The investigation has identified 3 rare CNVs in 3 different individuals – a ~19-Mb 17p13.3 duplication in a morbidly obese adult female (BMI = 47.83 kg/$m^2$), a ~1.6-Mb 5p14.3 deletion in another morbidly obese adult female (BMI = 50.06 kg/$m^2$), and a ~2.6-Mb 22q11.2 duplication in a morbidly obese adolescent male (BMI = 36.34 kg/$m^2$). Since these CNVs were singletons, the CNV-BMI relationship shown from the permutation-based CNV association tests was not a convincing argument to suggest a significant association of the CNVs with BMI or obesity. Furthermore, statistical comparison of means could not be performed due to the lack of other carriers of each of these CNVs. However, these discoveries are important to report because two of the CNVs involved the genomic regions that have been implicated in genomic syndromes previously reported in DECIPHER and OMIM databases. In other work by the Blakemore group members, the 16p11.2 deletion discussed in **Chapter 7** of this thesis was also first seen as a singleton in a small collection of extremely obese children. The causative association with obesity was, in that case, only confirmed by comparison with results of other collections throughout Europe (and later, across the globe). Thus, we intend to share these results with members of the 16p11.2 consortium to seek further cases.

The 17p13.3-p11.2 duplication spanned a region that overlapped the syndromic regions causing the 17p13.3 Deletion telomeric syndrome (17p13.3t.DelS), 17p13.3 Duplication telomeric syndrome (17p13.3t.DupS) Charcot-Marie-Tooth syndrome type 1A (CMT1A), Smith-Magenis syndrome (SMS) and Potocki-Lupski syndrome (PLS) on chromosome 17. Patients with CMT1A are characterised by various neural and motor impairments (OMIM #118220), while patients with PLS display neurobehavioural complications and cognitive defects (OMIM #610883). The SMS is commonly caused by deletions at the 17p11.2 region, but in other cases, it can also be caused by haploinsufficiency of the *RAI1* gene. The second CNV discovered in this chapter, the 22q11.2 duplication, spanned the critical region that causes either DiGeorge syndrome (deletion), or the 22q11DS (duplication) on chromosome 22. As for the 5p14.3 deletion, no known syndrome was found to be implicated by the deletion. However, the deletion region covered a potentially important gene for obesity and psychiatric disorders; *CDH18* gene. Further investigations in DECIPHER and OMIM databases revealed several patients carrying deletions similar to the 5p14.3 deletion, in which all of the deletion variants with different breakpoints shared coverage of a common gene (*CDH18*). The patients identified in DECIPHER exhibited a variety of phenotypes, with common reports of head/neck abnormalities, speech delay and cognitive impairments. The three CNVs reported in this chapter are discussed in more detail in the following sections.

### 8.6.2. The 17p13.3 CNVs, cognitive health and obesity

The 17p13 region has been known as a copy number variable region, in which its CNVs have been classified as two separate genomic syndromes, namely 17p13.3 Deletion syndrome (17p13.3t.DelS) and 17p13.3 Duplication syndrome (17p13.3t.DupS), with variable coverage. The region started from the telomeric end of the p arm of chromosome 17, and stretched for

approximately 3.3 Mb in length (hg19; chr17:1-3,300,000). 17p13.3 CNVs are believed to have formed by the non-allelic homologous recombination (NAHR) genomic process mediated by *Alu* repeat elements (Kuroda *et al.*, 2014).

The 17p13.3t.DelS encompassing tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein epsilon (*YWHAE*) and platelet activating factor acetylhydrolase 1b regulatory subunit 1 (*PAFAH1B1*) genes was previously reported in patients with abnormal brain structure, while the 17p13.3t.DupS covering the same genes was described in patients with ID (Bruno *et al.*, 2010). Another variation of 17p13.3t.DelS, encompassing tumour protein p53 (*TP53*) gene was described in a patient with ID and skeletal dysmorphisms (Shlien *et al.*, 2010), while the corresponding duplication has been implicated in two unrelated cases, both displaying ID and obesity (Belligni *et al.*, 2012, Kuroda *et al.*, 2014).

The 17p13.3 duplication variant reported in the morbidly obese female (~19 Mb) spanned several genes, including *YWHAE*, *PAFAH1B1* and *TP53* genes. Other than that, the duplication region encompassed a number of metabolism-related genes, particularly the solute carrier family 2 member 4 (*SLC2A4*) gene (OMIM *138190). The *SLC2A4* gene encodes for a membrane protein called glucose transporter 4 (GLUT4). Upon stimulation by insulin, GLUT4 plays an integral role in the glucose uptake by muscle and adipose tissues (Kuroda *et al.*, 2014). Disruption in the expression and regulation of *SLC2A4* gene have been previously associated with obesity and Type II Diabetes, while overexpression of the *SLC2A4* gene resulted in a boosted (quicker) response rate for insulin and greater utilisation of glucose (Charron *et al.*, 1999, Kuroda *et al.*, 2014). Therefore, it can be suggested that the duplication of *SLC2A4* gene might cause overexpression of the *SLC2A4* gene, thereby increasing the glucose uptake by the muscles and adipocytes, possibly contribute to the obesity phenotype

in the 17p13.3 duplication carrier. These possibilities could be explored by transcriptomic analysis if suitable samples or pre-existing data are available and we intend to discuss this with the NFBC1966 study group.

Overall, the 17p13.3 duplication variants from the previous studies had common phenotypic characteristics, namely ID and obesity. The present case had morbid obesity but the information regarding her cognitive health was not available at the time of writing the thesis. Therefore, this study provides tentative supporting evidence that the ~19-Mb 17p13.3t.DupS encompassing *YWHAE*, *PAFAH1B1* and *TP53* and *SLC2A4* genes, is characterised by obesity. Nonetheless, further investigation is needed to elucidate the effect of *SLC2A4* gene duplication on the glucose uptake in adipocytes and how such a relationship might affect body weight. Other than that, the *SLC2A4* gene could act together with other nearby metabolism-related genes in a tightly regulated system to maintain the energy balance in the body. Hence, any aberrations could result in complex metabolic disruptions that may lead to obesity. It is also possible that obesity is an endophenotype of ID.

### 8.6.3. The 22q11.2 CNVs, cognitive health and obesity

The CNVs of the 22q11.2 genomic region are known to have significant implications in various syndromic disorders. The most recognised disorder related to the 22q11.2 locus is DiGeorge syndrome, in which a deletion causes the 22q11.2 deletion syndrome (22q11.2DelS). With a frequency estimated at 1 in 4,000 births, the 22q11.2DelS is the most common known deletion syndrome in human (Hoeffding *et al.*, 2017, Oskarsdottir *et al.*, 2004). Most patients with 22q11.2DelS harbour a ~3-Mb deletion, while some others carry a smaller, nested deletion within the 3-Mb deletion region. The deletion variants (full and nested) of the 22q11.2DelS were reported to be inherited in less than 10% of the cases, while

the remaining were believed to arise *de novo* via NAHR events between the low copy repeats (LCR) of the 22.q11.2 region (Hoeffding *et al.*, 2017, Stankiewicz and Lupski, 2002). Previous studies reported broad phenotypic expressions of the 22q11.2DelS patients - from congenital heart and immune defects (McDonald-McGinn *et al.*, 1993), to major neurodevelopmental impairments such as schizophrenia (Williams, 2011), autism, ADHD and learning difficulties (Hoeffding *et al.*, 2017, Niklasson *et al.*, 2009, Schneider *et al.*, 2014).

Recent studies have described the reciprocal 22q11.2 duplication syndrome (22q11.2DupS) in patients with ID, DD, autism, learning difficulties and congenital defects, albeit in milder effects (Courtens *et al.*, 2008, Wentzel *et al.*, 2008, Ou *et al.*, 2008, Wenger *et al.*, 2016). However, other studies conducted on individuals without any cognitive ascertainment, have identified carriers of the 22q11.2 duplications who did not display any of the phenotypes described above (Courtens *et al.*, 2008, Van Campenhout *et al.*, 2012). The 22q11.2 duplication has been observed to be frequently inherited (Ou *et al.*, 2008) as well as to arise *de novo* (Coppinger *et al.*, 2009, Ou *et al.*, 2008). The most commonly reported 22q11.2 duplication was ~3Mb in length, spanning the same locus of the ~3-Mb deletion variant of the 22q11.2DelS (Hoeffding *et al.*, 2017, Ou *et al.*, 2008). Based on recent reports, a smaller duplication (~1.5 Mb) nested within the 3-Mb region and a larger duplication (~4 Mb to 6 Mb) spanning the critical 22q11.2 region have also been described in patients with the 22q11.2DupS (Ou *et al.*, 2008, Portnoi, 2009). Although studies on 22q11.2DupS are limited, the duplication frequency in the general population was estimated to be half that of its corresponding deletion (Ou *et al.*, 2008).

All 3 duplication variants (nested, full, and large) of the 22q11.2 region encompassed several common genes, including the solute carrier family 25 member 1 (*SLC25A1*) gene (OMIM *190315). The *SLC25A1* gene encodes for a mitochondrial solute carrier protein that

transports citrate out of mitochondria, while bringing in malate. Within the mitochondria, the two molecules (citrate and malate) are potent in the energy production reactions to support various cellular activities. Crucially, citrate also involves in lipid productions and regulate glycolysis after leaving mitochondria (https://www.genecards.org/cgi-bin/carddisp.pl?gene=SLC25A1).

A metabolomics study in children with 22q11.2DelS revealed the significant role of *SLC25A1* haploinsufficiency in altering energy metabolism (Napoli *et al.*, 2015). The author observed that a shift from oxidative phosphorylation to glycolysis, and increased concentrations of fatty acids and cholesterol, was initiated by the reduced dosage of *SLC25A1* gene copies (Napoli *et al.*, 2015). Other than that, a study by Maynard, *et. al.* (2008) which was performed in mice suggested that altered gene dosage of one or several mitochondrial genes located in the 22q11.2 region (including *SLC25A1*) could significantly affect neuronal signalling and metabolism in the brain (Maynard *et al.*, 2008). These findings could suggest a potentially critical effect of *SLC25A1* gene dosage in maintaining the energy balance and cognitive health in human. Despite that, the effects of having duplicated copies of *SLC25A1* gene on energy metabolism have not yet been systematically investigated. It is interesting to know whether the difference in *SLC25A1* gene dosage is responsible for the obesity phenotype observed in patients with 22q11.2DupS, since an impaired energy balance can lead to altered body weight and obesity. Additionally, the potentially dual effect of *SLC25A1* gene dosage on weight and cognitive development could also add to the growing evidence that obesity is primarily driven by disruptions in the neuro-behavioural pathways in the brain, rather than solely by energy imbalance (O'Rahilly and Farooqi, 2008). Therefore, this warrants further investigations to assess the relationship between the CNVs involving the *SLC25A1* gene and obesity. One relatively unbiased approach to determining phenotypic effects of

gene dosage in this region might be to seek carriers of the duplication and deletion among participants in very large-scale biobanks such as the UKBB and the China Kadoorie Biobank and examine their anthropomorphic, imaging and biochemical data.

### 8.6.4. The 5p14.3 CNVs, cognitive health and obesity

In this chapter, one morbidly obese adult has been confirmed as carrying a ~1.6-Mb heterozygous deletion in the 5p14.3 region (5p14.3DEL). To date, the 5p14.3DEL has not been previously implicated in any known syndromes or disorders, including obesity. Thus, the 5p14.3DEL could be a potentially novel microdeletion for obesity.

As shown earlier in this chapter (**Figure 8- 8)**, 6 other patients were described as harbouring similar deletions, with variable breakpoints, from the DECIPHER database. It is interesting to note that all the deletions shared a common gene; the *CDH18* gene (OMIM *603019). The *CDH18* gene encodes a cell adhesion protein called Type II classical cadherin which is primarily expressed in the brain. The main function of the encoded cadherin is mediating cell-to-cell adhesion. The cadherin superfamily is known to be involved in various intracellular signalling pathways (Hirano and Takeichi, 2012) - crucially in pathways that have been associated with neuropsychiatric disorders (Redies *et al.*, 2012). In general, cadherins have distinct involvements in the development of the brain and the functioning capabilities of a mature brain (Redies *et al.*, 2012). For example in the central nervous system (CNS), cadherins play various important roles in neuronal migration, differentiation of gray matter, and the formation of synapses (Hirano and Takeichi, 2012), with many disease-causing mutations were proposed to increase the risk of neuropsychiatric disorders.

A comprehensive review by Hulpiau and van Roy (2009) revealed that the cadherin encoded by the *CDH18* gene was associated with schizophrenia, while other classical

cadherins were found to be associated with ID, autism, bipolar disease and bipolar alcoholism (Hulpiau and van Roy, 2009). Redies, C., Hertel, N., and Hubner, C. A. (2012) argued that the involvement of cadherins in all steps of developmental and synaptic specifications, from early embryonic patterning to the formation of a mature brain, could indicate the importance of classical cadherins in the overall cognitive development in human brain (Redies *et al.*, 2012). It has been speculated that any disruption in the functional capacity of the cadherin-specific adhesive system, could break the overall functional connectivity and logical processing in the brain, a consequence that has been frequently observed in patients with neuropsychiatric diseases (Redies *et al.*, 2012). Additionally, a GWAS conducted by Terracciano, *et. al.* (2010) revealed that several SNPs within the *CDH18* gene were linked to personality traits, further suggesting that there might be a link between neuropsychiatric disorders and personality (Terracciano *et al.*, 2010). These studies have clearly shown the influential roles of the *CDH18* gene in neuronal growth, connectivity and repair, as well as in overall cognitive functioning of the brain.

Recent GWAS studies on obesity-related phenotypes (i.e. BMI and waist circumference) have identified SNPs near and within the *CDH18* gene to be of significant importance in modifying the phenotypes as well as increasing the risk of obesity (Fox *et al.*, 2007, Zhang *et al.*, 2013b). Interestingly, an investigation to identify the quantitative trait loci (QTL) for adiponectin revealed the 5p14 locus as one of the highly significant adiponectin QTLs (Comuzzie *et al.*, 2001). A QTL is defined as a genomic region that has a significant correlation with the variation in a particular phenotype. Adiponectin which is a protein hormone derived from adipocytes, increases its expression when the total fat level is low (Zhang *et al.*, 2013b) - an indication of adiponectin's possible involvement in body composition. It has been postulated that adiponectin may have an important role as an intermediate phenotype for

obesity by modifying body weight (Comuzzie *et al.*, 2001). A follow-up study on the adiponectin QTLs in metabolic syndromes showed that at least 6 SNPs located in the introns of the *CDH18* gene were significantly associated with waist circumference, weight, and BMI (Zhang *et al.*, 2013b). In metabolic syndromes, it has been suggested that increasing concentration of adiponectin reduces the risk of diabetes and obesity, as well as the tendency of accumulating fat in the arteries (Zhang *et al.*, 2013b). Furthermore, Fox, *et. al.* (2007) have demonstrated that the SNP rs1374489 located within the *CDH18* gene was associated with BMI in individuals from a heart study cohort (Fox *et al.*, 2007). These findings have increased the growing evidence of the importance of *CDH18* gene in the developments of body composition and obesity.

In conclusion, there is a range of evidence to support a possible role of the *CDH18* gene in influencing the body composition, but the extents of its effects on body weight and obesity are still poorly understood. More importantly, the effects of haploinsufficiency of the *CDH18* gene on obesity-related phenotypes, and their relationships with adiponectin levels in modifying the risk of obesity warrant further investigation. Since the 5p14.3 deletion stretches up to ~1.6 Mb in length, including another cadherin family *CDH12* gene in the vicinity, it is possible that *CDH18* gene plays a more mechanistic role with other affected genes in displaying the obesity phenotype.

# CHAPTER 9. Conclusions and Future Work

## 9.1. Summary of findings

The work presented in this thesis includes i) the imputation of CNVs from SNP genotyping data; (ii) analysis of obesity-related phenotypes in relation to common CNVs in the *AMY1* genomic region, and to a pair of rare CNVs at chr16p11.2; and (iii) detection of putative pathogenic CNVs in obese individuals from a general population. The work was initially based on a hypothesis-free approach, in which the availability of SNP genotyping data from various study cohorts made it possible to generate a CNV catalogue for each cohort using signal intensity information as my primary source of data.

As a priority, the scope of CNV interrogation was narrowed down to focus on CNVs that had been previously reported to be associated with obesity. Since the anthropometric data on BMI and WHR are available across all the study cohorts, I decided to look into the associative relationships of known CNVs with these phenotypes, as proxy measures of obesity, in addition to the effects of each CNV on the risk of obesity itself. The summary of the findings from each result chapter **(Chapter 4, Chapter 5, Chapter 6, Chapter 7, Chapter 8**) are below.

### 9.1.1. Quality SNP genotyping data is prerequisite for CNV analysis

The imputation of CNVs from signal intensity data from SNP genotyping is technically challenging and the source data needs to be of the highest quality possible. Prior to any CNV analyses, every dataset had to be scrutinised to remove samples that had poor genotyping quality due to any of several reasons – including poor DNA quality or concentration, sample mix-up, DNA contamination. Furthermore, investigating genetics in a population-level dataset

can be confounded by the presence of sub-groups of different ancestral background within a population (Lemes *et al.*, 2018, Edwards and Gao, 2012). This matter was addressed by assessing the population structure within each study cohort (Edwards and Gao, 2012). Thus, the first result chapter presented in this thesis (**Chapter 4**) contains all the work that was relevant to these quality issues in the NFBC1966, NFBC1986, NutriTech and the PMMO cohorts. The work included post-SNP genotyping data processing, quality control (QC) checks and population structure adjustments. The key findings in **Chapter 4** are as follows:

1) The failure rates from the SNP genotyping quality controls were less than 5% for the NFBC1966, NutriTech and the PMMO cohorts. The NFBC1986 cohort, which was genotyped in two batches and using two different arrays, had a relatively higher failure rate of ~6.5%. The possible reasons, which have been discussed in **Chapter 4**'s Discussion section, being either DNA contamination, various human-related errors, or sub-optimal genotyping condition in the second batch (which contributed ~87% of the samples who failed the QC).
2) The failure rates from the population structure adjustments were less than 5% for the NFBC1966, NFBC1986, NutriTech and the PMMO cohorts.

The overall genotyping quality across the four study cohorts under study was good, while the number of individuals removed due to inbreeding, genetic relatedness, or as ethnic outliers within each study cohort were low as well: this was consistent across the four study cohorts. At the end of the work in **Chapter 4**, I had generated a clean, homogeneous dataset of genotyped samples for each study cohort under study.

### 9.1.2. Cataloguing high-quality CNV calls

In the next results chapter (**Chapter 5**), I followed up by detecting CNVs from analysis of signal intensity data of the cleaned, genotyped datasets using a CNV-calling algorithm, the PennCNV tool. The main aim was to generate a high-quality CNV catalogue specific for each study cohort (NFBC1966, NFBC1986, NutriTech and PMMO). Initially, the proportion of CNV calls was different across the four study cohorts. For example, I observed, on average, very much higher number of CNV calls per patient in the PMMO cohort (more than 5-fold higher), than in the other study cohorts. Furthermore, the majority of the excess CNV calls were 1-copy duplications. I also observed a stark difference in the number of CNV calls between the first and the second batch of the NFBC1986 cohort (ratio of 1 to 3). Again, most of the CNV calls in the second batch were 1-copy duplications. This led to the decision to refine each CNV catalogue by applying various quality parameter checks. To check that each CNV catalogue did not over- or under-estimate CNV calls, the catalogues were compared to one another, as well as with the findings reported in other CNV studies. The key findings from the work in **Chapter 5** are as follows:

1) The low quality CNV calls were successfully filtered out from each CNV catalogue, especially the 1-copy duplication calls accumulated in the second batch of the NFBC1986 cohort, which were reduced by 12-folds.
2) Compared to the other cohorts, the clean CNV catalogue generated in the obesity cohort (PMMO), contained ~18- to ~24-times lower proportion of deletions, and ~2- to ~3.4-times higher proportion of duplications. Since this cohort consisted of very obese patients, the high number of duplications might possibly have a clinical significance in obesity. It will be possible to examine this in future work, as the larger PMMO cohort (N>2,000) is about to undergo genotyping using a custom-designed

array. More attention should be given to interrogation of these CNVs in the PMMO cohort in this follow-up study.

3) Significant differences were observed in the proportions of deletions, duplications, rare and singleton CNVs for the PMMO cohort, compared to the CNV calls for the NFBC1966, NFBC1986 and NutriTech cohorts, as well as to other CNV studies from different populations (Shaikh *et al.*, 2009, Xu *et al.*, 2011, Kanduri *et al.*, 2013).
4) Significant differences were observed in the CNV lengths predicted for the NFBC1966 cohort, compared to the CNV calls for the NFBC1986, NutriTech and PMMO cohorts, as well as to other CNV studies from different populations (Shaikh *et al.*, 2009, Xu *et al.*, 2011, Kanduri *et al.*, 2013).

At the end of **Chapter 5**, I have generated a cleaned CNV catalogue containing high-quality CNV calls specific for each study cohort under study. The proportion of 1-copy duplications, although had been cleaned, was still significantly high in the PMMO cohort (deletion-to-duplication ratio =0.03) compared to other cohorts. As I have discussed in **Chapter 5**, the PMMO cohort was genotyped in two separate arrays, one of which had no exonic markers. The merging of both genotyping batches followed by the merging of overlapping CNV calls could have masked small deletions from being detected by the PennCNV tool, although this suspicion should be investigated in more detail, possibly by visually inspecting a select of those duplications for validation. In addition, CNV calls predicted for the NFBC1966 cohort had, on average, significantly different CNV lengths compared to the other cohorts. This could be due to the limited number of probes on the genotyping arrays used (only >300,000 probes). Other relatively newer arrays contain nearly a million probes. The limited coverage of the array used in genotyping the NFBC1966 samples made it possible for the PennCNV tool to infer two duplications separated by other CNVs in between them, as one large duplication.

Despite these technical issues, to a certain extent, the CNV catalogues can be compared with reported findings from different populations. These CNV catalogues were then used for the initial detection of known CNVs, and to seek new potentially pathogenic CNVs in obese subjects. In addition to their use in the present work, the cleaned CNV catalogues for the two Northern Finnish Birth cohorts will be supplied to the study co-ordinators so that can be used to examine possible CNV associations with other phenotypes.

### 9.1.3. No association of *AMY1* copy number with obesity-related phenotypes

The first reportedly obesity-associated CNV investigated was included the human salivary amylase gene (*AMY1*). Previous reports had suggested its significant relationship with BMI and obesity (Falchi *et al.*, 2014, Marcovecchio *et al.*, 2016, Mejia-Benitez *et al.*, 2015) with p-values of up to 6.93 x $10^{-10}$ (Falchi *et al.*, 2014) which is higher than any known SNP association. Specifically, the report by Falchi, *et. al.* (2014), indicated that having one less copy of *AMY1* gene increases BMI by 0.15 kg/$m^2$, and low *AMY1* copy number increases the risk of obesity by 1.19 times per copy (Falchi *et al.*, 2014). In this thesis, the relationships of *AMY1* copy number with obesity-related phenotypes (i.e. BMI, WHR) and various adiposity measures (total abdominal fat, pancreatic fat, visceral fat, and subcutaneous abdominal fat) were explored comprehensively in **Chapter 6**.

It is important to note that reliable determination of copy number of a highly multi-allelic CNV region such as the amylase gene region, by any SNP-based CNV calling algorithm is highly unlikely. This is because in most cases, the CNV calling algorithms implement two copy numbers as a reference across the genome, and any estimated departure from the reference will usually be reported as a CNV call. For the *AMY1* gene, the normal copy number varies

between 2 to 17 copies (Groot *et al.*, 1991, Perry *et al.*, 2007, Usher *et al.*, 2015). Due to this high variability, the CNV catalogues generated by the CNV calling algorithm were not used in **Chapter 6**.

To address this issue, we collaborated with Armour J., *et al.* (University of Nottingham, UK) to genotype every participant from the NutriTech cohort, using multiplex paralogue ratio tests (PRTs) that they developed specific for the amylase region (Armour *et al.*, 2007, Carpenter *et al.*, 2015, Walker *et al.*, 2009). The key findings from the work in **Chapter 6** are as follows:

1) The *AMY1* copy numbers estimated from the PRT-based methods were significantly correlated with signal intensity data (LRR and RR) of the SNP located in the *AMY1* gene.
2) No significant association between *AMY1* copy number and BMI or WHR in the NutriTech cohort was detected.
3) No significant association between *AMY1* copy number and any of the adiposity measures (i.e. total abdominal fat, pancreatic, subcutaneous and visceral fat content) in the NutriTech cohort was detected.
4) No significant association between *AMY1* copy number and dietary starch intake in the NutriTech cohort was detected.

These findings have been discussed extensively in the Discussion section of **Chapter 6.** Prior to the association analysis, the AMY1 copy numbers genotyped from the PRT-based methods were assessed whether they correlate well with signal intensity data. As expected, the correlation was significantly high, further elucidating the accuracy of the PRT method used in genotyping multi-allelic CNVs. As for the non-association findings, it is possible that the association signals reported previously could have come from the copy numbers of the neighbouring pancreatic amylase genes; *AMY2A* and *AMY2B* genes. These genes have more than 93% of sequence identity with the salivary *AMY1* gene (Carpenter *et al.*, 2015, Kent WJ

SC, 2002), increasing the possibility of having wrongly estimated *AMY1* copy numbers. In addition, the differences in the methodological approaches have also been discussed in **Chapter 6** as potential challenges in addressing the highly multi-allelic nature of the *AMY1* copy number region.

### 9.1.4. Proximal and distal 16p11.2 deletions are associated with obesity-related phenotypes

The second known CNV region investigated included the two obesity-related deletion variants in the copy number variable 16p11.2 region. There are two non-overlapping deletion regions, located proximally and distally from the centromere of chromosome 16. The proximal deletion (~600 kb in length) was initially associated with autism (Kumar *et al.*, 2008, Marshall *et al.*, 2008, Weiss *et al.*, 2008) and mental retardation (Bijlsma *et al.*, 2009), before Walters*, et. al.* (2010) reported its associative relationship with BMI and obesity in addition to neuropsychiatric conditions (Walters *et al.*, 2010). Since then, the proximal deletion variant has been consistently reported to be associated with obesity in other studies (Maillard *et al.*, 2016, Yu *et al.*, 2011). On the other end, the distal deletion region containing *SH2B1* gene (~200 kb in length) has been previously associated with ID and DD (Bachmann-Gagescu *et al.*, 2010), as well as with obesity (Bachmann-Gagescu *et al.*, 2010, Bochukova *et al.*, 2010). In this thesis, my investigation of the relationships of each 16p11.2 deletion variant with obesity-related phenotypes (i.e. BMI and WHR) is presented in **Chapter 7**. Apart from that, I also assessed the effect of each variant on the risk of obesity. For this chapter, I used the CNV catalogues of the NFBC1966 (birth cohort) and the PMMO (obesity patient cohort) cohort (from **Chapter 5**) to identify the carriers of both deletion variants. However, due to the non-availability of CNV catalogue for the UKBB cohort at the time of analysis, I used its genotyped

data to scan for runs of homozygosity overlapping either deletion region. This could potentially detect both true homozygosity or hemizygosity due to a deletion. The presence of deletions was validated by visual inspection of signal intensity plots across the region. The key findings described in **Chapter 7** are as follows:

1) The proximal deletion was significantly associated with BMI and WHR in the NFBC1966 and UKBB cohorts. Regression analysis could not be conducted for the PMMO cohort due to there being only one carrier in that small cohort.
2) The distal deletion was significantly associated with BMI and WHR in the NFBC1966 cohort. Regression analysis could not be performed for the UKBB and PMMO cohorts due to insufficient deletion carriers being detected.
3) The proximal deletion significantly increased the risk of obesity in the NFBC1966 and UKBB cohorts. This type of analysis could not be carried out in the PMMO cohort because all of the patients were obese/morbidly obese.
4) The distal deletion did not significantly affect the risk of obesity in the NFBC1966 and UKBB cohorts. Fisher's exact test could not be carried out for the PMMO cohort because all of the patients were obese/morbidly obese.
5) In a combined analysis of PMMO and UKBB, both deletion variants were significantly enriched in the obese participants, compared to the non-obese.

In summary, I managed to replicate the associations of the proximal 16p11.2 deletion with BMI (Jacquemont *et al.*, 2011, Mace *et al.*, 2017, Zufferey *et al.*, 2012) and WHR (Mace *et al.*, 2017) in the UKBB cohort at or near genome-wide significance threshold, in which having the deletion increases BMI by 12.08 kg/m$^2$ ($p=9.75 \times 10^{-6}$) and reduces WHR by 0.04 ($p=2.3 \times 10^{-7}$). The association findings were similarly detected in the smaller sized NFBC1966 cohort, but with lower significance. The frequency of the proximal 16p11.2 deletions detected in my work

(0.0009%) is relatively lower than other studies that involved the UKBB data (~0.03%) (Kendall *et al.*, 2017, Mace *et al.*, 2017). This difference could be due to the lack of a validation step in the two studies, in which they relied on CNV predictions (Kendall *et al.*, 2017, Mace *et al.*, 2017). In the report by Kendall, *et. al.* (2017) the CNVs called by the PennCNV tool were subjected to similar signal intensity parameter filtering as in my work, and used those to annotate pathogenic CNV, including the proximal 16p11.2 deletion variant (Kendall *et al.*, 2017). Similarly in the report by Mace, *et. al.* (2017), they used the PennCNV-predicted CNVs together with signal intensity data (LRR and BAF) to compute a "quality score (QS)" for high-quality CNV screening (Mace *et al.*, 2016, Mace *et al.*, 2017). However, there was no mention of any laboratory-based or visual-based validation involved. Whereas in my work, I initially identified 242 from 329,419 UKBB participants (~0.07%) to have a homozygosity stretch in that proximal 16p11.2 region. However, upon visual inspections to validate those deletions, I confirmed that only 3 of them had the proximal ~600-kb deletion. Thus, I believe that many of the deletion carriers in both studies were likely to have a homozygosity stretch in the same proximal 16p11.2 region, rather than carrying an actual deletion. I consider that the visual inspection step for validation as one of the strengths of this work. The other strength of this work is the use of a very obese patient cohort (PMMO) and a very large general population cohort (UKBB) in designing the pooled case-control association analysis, in which both 16p11.2 deletion variants were significantly enriched in obese patients, as well as in obese people in general.

### 9.1.5. Discovery of rare CNVs in the 5p14.3, 17p13 and 22q11.2 regions

I extended my investigation to identify large CNVs associated with obesity-related phenotypes across the genome, by interrogating the clean CNV catalogues generated from

two birth cohorts; NFBC1966 and NFBC1986. This work was carried out as an association analysis based on permutation tests, as presented in my final result chapter of this thesis (**Chapter 8**). The association tests were performed on the cleaned CNV calls, as proxy CNVs for the discovery stage. Since the CNV calls were derived from SNP genotyping data, principal components (PC) representing variation values that existed strongly within the genotyped data, were included as covariates, along with sex and genotyping batch. The key findings from the work in **Chapter 8** are as follows:

1) Three rare CNVs (17p13, 22q11.2, and 5p14.3 regions) were identified in three morbidly obese individuals, based on the CNV-BMI permutation tests.
2) In the NFBC1966 cohort, the presence of a large 17p13 1-copy duplication (~20 Mb in length) and a large 5p14.3 heterozygous deletion (~1.6 Mb in length) were confirmed in one adult each.
3) In the NFBC1986 cohort, the presence of a rare 1-copy duplication (~2.6 Mb) in the 22q11.2 region was validated in one adolescent.

Based on these findings, **Chapter 8** provides the description of these rare CNVs discovered in the two birth cohorts. The 17p13 duplication (~19 Mb) which was found in a female adult with morbid obesity, spanned a smaller duplication that have been previously reported in two separate cases exhibiting ID and obesity (Belligni *et al.*, 2012, Kuroda *et al.*, 2014). The reported duplication (~0.8 Mb to ~3.2 Mb) encompassed several genes including *SLC2A4* gene, which is believed to be involved in the glucose uptake stimulated by insulin(Kuroda *et al.*, 2014). Furthermore, disruption of the *SLC2A4* gene has been associated with diabetes and obesity (Kuroda *et al.*, 2014).

Other than that, the 22q11.2 duplication found in a morbidly obese male adolescent encompassed ~78 genes (deletion of the same region causes diGeorge syndrome). Most

importantly, this 22q11.2 duplication has been previously found in a patient with ADHD, autism, sensory and language disorders (Guy *et al.*, 2015). Besides that, a population study on the nationwide Danish registry revealed its association with increased risk of many psychiatric disorders, including ID (Hoeffding *et al.*, 2017). However, none of the previously reported findings characterised the 22q11.2 duplication carriers with obesity. To my knowledge, this present study is the first to report the proposed relationship of the 22q11.2 duplication with increased body weight and morbid obesity, possibly due to impaired *SLC25A1*-mediated neuronal signalling that controls the energy balance in the brain.

As for the 5p14.3 deletion (~1.6 Mb), the carrier was a morbidly obese adult (female). In the DECIPHER database, 1 carrier of a larger ~14.7 Mb deletion, encompassing the 5p14.3 deletion region identified in this study, was reported to exhibit head/neck abnormality and obesity. The other 5 patients carrying deletions of variable overlapping regions with the 5p14.3 deletion, displayed various phenotypes; head/neck abnormalities, cognitive defects and respiratory/digestive problems. Interestingly, all overlapped deletion variants from DECIPHER and the deletion found in this study, spanned a genomic region containing the whole stretch of *CDH18* gene. The *CDH18* gene encodes for Type II classical cadherin proteins (a type of cell adhesion molecule) in the brain, which mediate cell-to-cell adhesion (OMIM *603019). *CDH18*'s cadherin has been previously associated with neuropsychiatric disorders such as schizophrenia (Redies *et al.*, 2012), while genetic polymorphisms within the *CDH18* gene have been associated with increased BMI (Fox *et al.*, 2007), waist circumference (Fox *et al.*, 2007, Zhang *et al.*, 2013b) and increased risk of obesity (Wilson *et al.*, 2015). The present study reports the discovery of a novel, rare heterozygous deletion in the 5p14.3 region, and proposes a contribution of the deletion towards increased body weight and obesity, possibly via the action of adiponectin.

Because I only found one carrier for each of the three large CNVs, I could not make an associative conclusion of their relationships on BMI and obesity. Thus, further investigation is needed to address the potential effects of these large genomic imbalances on obesity-related phenotypes and obesity in general. For example, these regions could be evaluated in the large PMMO analysis currently ongoing, in GIANT consortium data, and in the UK Biobank.

Overall, the strategies used in generating cleaned, cohort-specific CNV catalogues, and validating them using signal intensity plots, has been effective and especially useful for large CNVs. Thus, the catalogues generated are also likely to be useful of preliminary analysis of other phenotypes.

### 9.1.6. Neurobehavioural connection to obesity

This thesis has provided further evidence to support the associations of both deletion variants in the 16p11.2 region, with BMI and obesity and revealed 3 rare CNVs that could be important in obesity aetiology, paving the way for more investigations on these rare, under-reported syndromic CNVs. Like the 16p11.2 deletions, both large duplications (17p13 and 22q11.2), and genetic variations within the *CDH18* gene (located within the 5p14.3 deletion region), have been implicated in neurodevelopmental disorders and/or psychiatric disease. As mentioned previously, all the carriers of such large CNVs were morbidly obese. Adding to the fact that there is a prominent relationship between cognitive impairments and obesity in the 16p11.2 deletion carriers (Walters *et al.*, 2010), these findings further underline the neurobehavioural basis of Mendelian obesity, which seems to almost exclusively result from defects in appetite control, rather than energy balance disturbance (O'Rahilly and Farooqi, 2008). Furthermore, mental retardation is known to have a strong correlation with the prevalence of obesity (Boeka and Lokken, 2008), and the increased risk of obesity in

individuals with learning difficulties or autism may reflect shared aetiological mechanisms (Chen *et al.*, 2010).

Monogenic obesity genes, such as leptin receptor (*LEPR*), leptin (*LEP*), melanocortin receptor 4 (*MC4R*), and proopiomelanocortin (*POMC*) are central for appetite regulation and control(Walley *et al.*, 2009), while other monogenic obesity genes such as single-minded (Drosophila) homolog 1 (*SIM1*), neutrotrophic tyrosine kinase receptor type2 (*NTRK2*) and brain-derived neutrotrophic factor (*BDNF*) are important genes in pathways related to neuro-development and hypothalamus functioning (Friedel *et al.*, 2005, Holder *et al.*, 2000, Yeo *et al.*, 2004). Of note, the fat-mass and obesity-associated (*FTO*) gene, which was the first GWAS-identified gene to be associated with common obesity, is also expressed in the brain (Frayling *et al.*, 2007). Furthermore, genetic polymorphisms of the human *FTO* gene have associative relationships with decreased satiety and increased dietary intake, in addition to the increased risk of obesity (Speakman *et al.*, 2008, Wardle *et al.*, 2008a). These examples emphasise the connection between obesity phenotypes with neurodevelopmental functioning in humans and could indicate a shared pathway between body weight regulation and cognition.

The discoveries of rare duplications in the 17p13 and 22q11.2 regions, as well as a possibly novel 5p14.3 deletion, have increased the list of rare structural variants that could be important in the genetics of obesity. It is imperative to investigate such rare CNVs, especially in individuals with extreme phenotypes (i.e. morbid obesity), in order to determine their contributions on complex diseases, such as obesity.

## 9.2. Limitations & Future Work

### 9.2.1. Investigation of salivary and pancreatic amylase CNVs

The findings from **Chapter 6** of this thesis have suggested the non-associative relationship between *AMY1* copy numbers and obesity-related phenotypes, as well as body fat, at least for the overweight-to-obese individuals in the NutriTech cohort. However, further investigation is required to assess the association in a general population. In addition, the role of pancreatic amylase genes (*AMY2A* and *AMY2B*) copy numbers in obesity is still relatively unknown and was not investigated in great detail in this thesis. It is also important to know that the sample size of the NutriTech cohort is very small, thus the power to detect meaningful significant results is limited. However, it is slightly mitigated by the excellent, well-characterised phenotyping in the cohort. In addition, I have shown that the *AMY1* copy number estimated by the PRT-based methods correlates well with signal intensity data of the only probe located on the *AMY1* gene. Although this could suggest the possible use of signal intensity data to predict the copy numbers, more probe coverage is needed to ensure that the signal intensity of the probe is not affected by other nearby probes corresponding to other amylase genes in the region. Furthermore, the LRR is a normalised value that already takes into account the intensity signals from other probes, thereby affecting the actual signal intensity of that probe.

As discussed in **Chapter 6**, although dPCR has been proposed as a sensitive method to genotype highly polymorphic CNVs (Falchi *et al.*, 2014, Mejia-Benitez *et al.*, 2015, Whale *et al.*, 2012), the opposing findings by Viljakainen *et. al.* (2015) (Viljakainen *et al.*, 2015) and Usher *et al.* (2015) (Usher *et al.*, 2015) suggest the need for a stable, validated control probe

that targets a locus located outside of the polymorphic region to obtain a more accurate reference copy number. In addition, the difference in sample size is likely to have a major role in estimating the copy number. In the PRT-based methods developed by Armour, *et. al*., the author takes advantage of the presence of ERV sequence within the *AMY1* gene since the ERV sequence is also present on chromosome 12 (Carpenter *et al.*, 2015, Armour *et al.*, 2007). By amplifying both ERV sequences, the subsequent calibration mitigates the issue of misclassifying the copy numbers of other amylase genes (which have high sequence identity with the *AMY1* gene) as *AMY1* copy numbers, providing a sensitive estimation of the salivary *AMY1* gene copy number. Therefore, I believe that, to date, the use of specific PRTs provides the most sensitive and accurate measurement in addressing the genotyping of multi-allelic CNV regions, such as in the amylase gene cluster region on chromosome 1. However, the scalability for massive cohorts, such as the UKBB cohort, may prove difficult for PRTs to be conducted and laborious. This could provide an opportunity for improved accuracy in exploiting the signal intensity data from DNA microarrays, and more attention should be given to extend the functionality of high-throughput platforms, or to develop statistical algorithms specific for multi-allelic CNV genotyping.

As an extension of the work from this thesis, a number of works can be carried out in the future to replicate the findings in cohorts with bigger sample sizes. In addition, the effects of copy numbers for the pancreatic amylase genes on obesity-related phenotypes are still relatively unknown (and were not included in the scope of **Chapter 6**). Therefore, a similar investigation can be performed for the pancreatic *AMY2A* and *AMY2B* copy numbers as well. These suggestions are listed in more detail below:

1) Send the DNA samples of the individuals from both birth cohorts, and possibly from the obesity patient cohort, to Armour J., *et al.* (the University of Nottingham, UK), for

*AMY1* CNV genotyping. From here, the relationship of *AMY1* copy number with BMI, WHR and obesity can be assessed in adults and adolescents, as well as in morbidly obese patients.

2) Conduct a comprehensive association analysis for the pancreatic *AMY2A* and *AMY2B* copy numbers to explore their potential effects on BMI, WHR and obesity. This investigation can be performed in the overweight-to-obese cohort before exploring it in adults and adolescents from the birth cohorts.
3) Explore the possibility that dietary starch intake is an important modifier of the relationship between *AMY1* copy number and BMI in the two birth cohorts. In order to do this in adults and adolescents, I need to find out if the information of their dietary intakes is available. This investigation should also be conducted on the pancreatic amylase copy numbers.

### 9.2.2. Investigation of the proximal and distal 16p11.2 duplications

In **Chapter 7**, the comprehensive investigation on the associations of proximal and distal 16p11.2 deletions on BMI, WHR and obesity have elucidated the significance of the relationship between each deletion variant and the obesity-related phenotypes mentioned above. The reciprocal duplication of the same proximal 16p11.2 region (~600-kb) has been previously associated with decreased BMI and underweight in children and adults, mirroring the phenotypic expression exhibited by the corresponding deletion carriers (Jacquemont *et al.*, 2011). As for the distal 16p11.2 region, to my knowledge, its corresponding duplication has not been thoroughly investigated in terms of its associative relationship with obesity-related phenotypes (i.e. BMI, WHR) and obesity in general.

One of the limitations that I encountered in the work described in **Chapter 7** was the unavailability of the CNV catalogue for the UKBB cohort. For that reason, ROH scan was carried out to detect potential CNVs in the 16p11.2 region – this step is effective and reliable in identifying simple deletions with known breakpoints. However, partial deletions, duplications (partial and full), and much smaller CNV (i.e. exonic or genic deletion) of the same region could not be reliably detected by the ROH scan. Therefore, there are many potential works that can be carried out to extend the work presented in **Chapter 7**. First of all, a more sensitive method for detecting duplications should be prioritised. This can be realised in a number of ways, as described below:

1) Using a deep machine-learning method to train a high-level neural network application programming interface (API) to recognise a pattern corresponding to duplication events on the signal intensity plots. From this training, the machine-learning tool can extract specific information based on the percentage similarity between the actual duplication pattern, and its trained pattern. One of the software available for this purpose is TensorFlow (https://github.com/tensorflow/tensorflow).
2) As I have mentioned previously in this thesis, heterozygous SNPs will have their BAF values around 0.5 (representing either AB or BA genotypes), while homozygous SNPs will have their BAF values either around 0 representing AA genotype) or 1 (representing BB genotype). Consecutive homozygosity could indicate a potential deletion, just as I used the ROH. On the other hand, consecutive SNPs with BAF values around 0.33 (representing AAB genotype) or 0.67 (representing ABB genotype) could indicate a duplication event. In theory, I could use a combination of linux-based commands to extract the individual IDs as potential duplication carriers, by scanning a specific range of values from the BAF data table.

The next step is to conduct a similar investigation on the association of 16p11.2 duplications for both proximal and distal region, with obesity-related phenotypes (i.e. BMI and WHR), and their potential effects on the risk of obesity. However, due to the rarity of the 16p11.2 CNVs for most small- and moderately-sized general population cohorts, large biobanks, such as the UKBB, is essential to define the frequency of such duplications in general populations. In addition, we can also use specialised collection cohort such as PMMO to define the frequency of the duplications in clinical populations. However, the phenotypic effects of the putative duplications may be over-estimated due to ascertainment bias.

Genetic counselling is essential for individuals carrying the proximal and distal 16p11.2 CNVs, since these CNVs confer risks for obesity and neurocognitive defects. But in order to offer families the best advice, we are in urgent need for the most sensitive and accurate assessment of the phenotype. The best strategy is to detect partial deletions or duplications in both 16p11.2 regions – this could help us define critical regions responsible for the phenotype, just like in the case of PWS (de Smith *et al.*, 2009).

### 9.2.3. Interrogation of the rare, novel CNVs in other populations

In **Chapter 8**, I have identified three rare CNVs present in different morbidly-obese individuals from two birth cohorts. However, the low frequency of carriers of such CNVs have limited the statistical analysis that can be performed. The difficulty in this type of work includes replication since these CNVs are extremely rare in general populations. Thus, a global effort is needed to further characterise the phenotype of these CNVs. Although DECIPHER has been very helpful in defining the characteristics, it is currently oriented to clinical populations.

The work described in **Chapter 8** involves the use of the Northern Finland population. Although it is comprehensive, this population is an isolated population with possibly different frequencies of CNVs compared to other general populations. The representations of some CNVs may be over-estimated, while other CNVs may be under-estimated. Therefore, there is a need to look for the three CNVs in different birth cohorts, such as Avon Longitudinal Study of Parents and Children (ALSPAC; http://www.bristol.ac.uk/alspac/) and Born in Bradford (https://borninbradford.nhs.uk/) to compare the frequencies of such rare CNVs. In addition, identification of such CNVs in specialised patient cohorts such as the obesity patient cohort (PMMO) is essential because such rare CNVs could be enriched in obese patients as compared to the general population.

# Bibliography

AERTS, E., BECKERS, S., ZEGERS, D., VAN CAMP, J. K., VAN HOORENBEECK, K., MASSA, G., VERRIJKEN, A., MERTENS, I. L., VERHULST, S. L., ROOMAN, R. R., VAN GAAL, L. F. & VAN HUL, W. 2015. Genetic and structural variation in the SH2B1 gene in the Belgian population. *Mol Genet Metab,* 115**,** 193-8.

ALMAL, S. H. & PADH, H. 2012. Implications of gene copy-number variation in health and diseases. *J Hum Genet,* 57**,** 6-13.

ALSTERS, S. 2016. *Genetic analysis of extreme obesity.* Doctoral PhD, Imperial College London.

ALSTERS, S. I., GOLDSTONE, A. P., BUXTON, J. L., ZEKAVATI, A., SOSINSKY, A., YIORKAS, A. M., HOLDER, S., KLABER, R. E., BRIDGES, N., VAN HAELST, M. M., LE ROUX, C. W., WALLEY, A. J., WALTERS, R. G., MUELLER, M. & BLAKEMORE, A. I. 2015. Truncating Homozygous Mutation of Carboxypeptidase E (CPE) in a Morbidly Obese Female with Type 2 Diabetes Mellitus, Intellectual Disability and Hypogonadotrophic Hypogonadism. *PLoS One,* 10**,** e0131417.

ANDERSON, C. A., PETTERSSON, F. H., CLARKE, G. M., CARDON, L. R., MORRIS, A. P. & ZONDERVAN, K. T. 2010. Data quality control in genetic case-control association studies. *Nat Protoc,* 5**,** 1564-73.

ARMOUR, J. A., PALLA, R., ZEEUWEN, P. L., DEN HEIJER, M., SCHALKWIJK, J. & HOLLOX, E. J. 2007. Accurate, high-throughput typing of copy number variation using paralogue ratios from dispersed repeats. *Nucleic Acids Res,* 35**,** e19.

BACHMANN-GAGESCU, R., MEFFORD, H. C., COWAN, C., GLEW, G. M., HING, A. V., WALLACE, S., BADER, P. I., HAMATI, A., REITNAUER, P. J., SMITH, R., STOCKTON, D. W., MUHLE, H., HELBIG, I., EICHLER, E. E., BALLIF, B. C., ROSENFELD, J. & TSUCHIYA, K. D. 2010. Recurrent 200-kb deletions of 16p11.2 that include the SH2B1 gene are associated with developmental delay and obesity. *Genet Med,* 12**,** 641-7.

BAILEY, J. A., BAERTSCH, R., KENT, W. J., HAUSSLER, D. & EICHLER, E. E. 2004. Hotspots of mammalian chromosomal evolution. *Genome Biol,* 5**,** R23.

BAILEY, J. A., GU, Z., CLARK, R. A., REINERT, K., SAMONTE, R. V., SCHWARTZ, S., ADAMS, M. D., MYERS, E. W., LI, P. W. & EICHLER, E. E. 2002. Recent segmental duplications in the human genome. *Science,* 297**,** 1003-7.

BAKER, M. 2012. Structural variation: the genome's hidden architecture. *Nat Methods,* 9**,** 133-7.

BARRETT, J. C. & CARDON, L. R. 2006. Evaluating coverage of genome-wide association studies. *Nat Genet,* 38**,** 659-62.

BELLIGNI, E. F., DI GREGORIO, E., BIAMINO, E., CALCIA, A., MOLINATTO, C., TALARICO, F., FERRERO, G. B., BRUSCO, A. & SILENGO, M. C. 2012. 790 Kb microduplication in chromosome band 17p13.1 associated with intellectual disability, afebrile seizures, dysmorphic features, diabetes, and hypothyroidism. *Eur J Med Genet,* 55**,** 222-4.

BENNETT, A., SOVIO, U., RUOKONEN, A., MARTIKAINEN, H., POUTA, A., TAPONEN, S., HARTIKAINEN, A.-L., FRANKS, S., PELTONEN, L., ELLIOTT, P., JÄRVELIN, M.-R. &

MCCARTHY, M. I. 2005. No association between insulin gene variation and adult metabolic phenotypes in a large Finnish birth cohort. *Diabetologia,* 48**,** 886-891.

BENNETT, A. J., SOVIO, U., RUOKONEN, A., MARTIKAINEN, H., POUTA, A., TAPONEN, S., HARTIKAINEN, A. L., KING, V. J., ELLIOTT, P., JARVELIN, M. R. & MCCARTHY, M. I. 2004. Variation at the insulin gene VNTR (variable number tandem repeat) polymorphism and early growth: studies in a large Finnish birth cohort. *Diabetes,* 53**,** 2126-31.

BIJLSMA, E. K., GIJSBERS, A. C., SCHUURS-HOEIJMAKERS, J. H., VAN HAERINGEN, A., FRANSEN VAN DE PUTTE, D. E., ANDERLID, B. M., LUNDIN, J., LAPUNZINA, P., PEREZ JURADO, L. A., DELLE CHIAIE, B., LOEYS, B., MENTEN, B., OOSTRA, A., VERHELST, H., AMOR, D. J., BRUNO, D. L., VAN ESSEN, A. J., HORDIJK, R., SIKKEMA-RADDATZ, B., VERBRUGGEN, K. T., JONGMANS, M. C., PFUNDT, R., REESER, H. M., BREUNING, M. H. & RUIVENKAMP, C. A. 2009. Extending the phenotype of recurrent rearrangements of 16p11.2: deletions in mentally retarded patients without autism and in normal individuals. *Eur J Med Genet,* 52**,** 77-87.

BLACKLESS, M., CHARUVASTRA, A., DERRYCK, A., FAUSTO-STERLING, A., LAUZANNE, K. & LEE, E. 2000. How sexually dimorphic are we? Review and synthesis. *Am J Hum Biol,* 12**,** 151-166.

BOCHUKOVA, E. G., HUANG, N., KEOGH, J., HENNING, E., PURMANN, C., BLASZCZYK, K., SAEED, S., HAMILTON-SHIELD, J., CLAYTON-SMITH, J., O'RAHILLY, S., HURLES, M. E. & FAROOQI, I. S. 2010. Large, rare chromosomal deletions associated with severe early-onset obesity. *Nature,* 463**,** 666-70.

BOEKA, A. G. & LOKKEN, K. L. 2008. Neuropsychological performance of a clinical sample of extremely obese individuals. *Arch Clin Neuropsychol,* 23**,** 467-74.

BOJESEN, A., JUUL, S. & GRAVHOLT, C. H. 2003. Prenatal and postnatal prevalence of Klinefelter syndrome: a national registry study. *J Clin Endocrinol Metab,* 88**,** 622-6.

BOUTIN, P., DINA, C., VASSEUR, F., DUBOIS, S., CORSET, L., SERON, K., BEKRIS, L., CABELLON, J., NEVE, B., VASSEUR-DELANNOY, V., CHIKRI, M., CHARLES, M. A., CLEMENT, K., LERNMARK, A. & FROGUEL, P. 2003. GAD2 on chromosome 10p12 is a candidate gene for human obesity. *PLoS Biol,* 1**,** E68.

BRUNO, D. L., ANDERLID, B. M., LINDSTRAND, A., VAN RAVENSWAAIJ-ARTS, C., GANESAMOORTHY, D., LUNDIN, J., MARTIN, C. L., DOUGLAS, J., NOWAK, C., ADAM, M. P., KOOY, R. F., VAN DER AA, N., REYNIERS, E., VANDEWEYER, G., STOLTE-DIJKSTRA, I., DIJKHUIZEN, T., YEUNG, A., DELATYCKI, M., BORGSTROM, B., THELIN, L., CARDOSO, C., VAN BON, B., PFUNDT, R., DE VRIES, B. B., WALLIN, A., AMOR, D. J., JAMES, P. A., SLATER, H. R. & SCHOUMANS, J. 2010. Further molecular and clinical delineation of co-locating 17p13.3 microdeletions and microduplications that show distinctive phenotypes. *J Med Genet,* 47**,** 299-311.

BURNS, B., SCHMIDT, K., WILLIAMS, S. R., KIM, S., GIRIRAJAN, S. & ELSEA, S. H. 2010. Rai1 haploinsufficiency causes reduced Bdnf expression resulting in hyperphagia, obesity and altered fat distribution in mice and humans with no evidence of metabolic syndrome. *Hum Mol Genet,* 19**,** 4026-42.

CABREJO, L., GUYANT-MARECHAL, L., LAQUERRIERE, A., VERCELLETTO, M., DE LA FOURNIERE, F., THOMAS-ANTERION, C., VERNY, C., LETOURNEL, F., PASQUIER, F., VITAL, A., CHECLER, F., FREBOURG, T., CAMPION, D. & HANNEQUIN, D. 2006. Phenotype associated with APP duplication in five families. *Brain,* 129**,** 2966-76.

CALLE, E. E., THUN, M. J., PETRELLI, J. M., RODRIGUEZ, C. & HEATH, C. W., JR. 1999. Body-mass index and mortality in a prospective cohort of U.S. adults. *N Engl J Med,* 341**,** 1097-105.

CAMPBELL, M. C. & TISHKOFF, S. A. 2008. African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. *Annu Rev Genomics Hum Genet,* 9**,** 403-33.

CARPENTER, D., DHAR, S., MITCHELL, L. M., FU, B., TYSON, J., SHWAN, N. A., YANG, F., THOMAS, M. G. & ARMOUR, J. A. 2015. Obesity, starch digestion and amylase: association between copy number variants at human salivary (AMY1) and pancreatic (AMY2) amylase genes. *Hum Mol Genet,* 24**,** 3472-80.

CASTRO, V. M., MINNIER, J., MURPHY, S. N., KOHANE, I., CHURCHILL, S. E., GAINER, V., CAI, T., HOFFNAGLE, A. G., DAI, Y., BLOCK, S., WEILL, S. R., NADAL-VICENS, M., POLLASTRI, A. R., ROSENQUIST, J. N., GORYACHEV, S., ONGUR, D., SKLAR, P., PERLIS, R. H., SMOLLER, J. W. & INTERNATIONAL COHORT COLLECTION FOR BIPOLAR DISORDER, C. 2015. Validation of electronic health record phenotyping of bipolar disorder cases and controls. *Am J Psychiatry,* 172**,** 363-72.

CHARRON, M. J., KATZ, E. B. & OLSON, A. L. 1999. GLUT4 gene regulation and manipulation. *J Biol Chem,* 274**,** 3253-6.

CHEN, A. Y., KIM, S. E., HOUTROW, A. J. & NEWACHECK, P. W. 2010. Prevalence of obesity among children with chronic conditions. *Obesity (Silver Spring),* 18**,** 210-3.

CHEN, Y., DALES, R., TANG, M. & KREWSKI, D. 2002. Obesity may increase the incidence of asthma in women but not in men: longitudinal observations from the Canadian National Population Health Surveys. *Am J Epidemiol,* 155**,** 191-7.

CHINN, S. 2006. Definitions of childhood obesity: current practice. *Eur J Clin Nutr,* 60**,** 1189-94.

CLARE BYCROFT, C. F., DESISLAVA PETKOVA, GAVIN BAND,, LLOYD T. ELLIOTT, K. S., ALLAN MOTYER, DAMJAN VUKCEVIC,, OLIVIER DELANEAU, J. O. C., ADRIAN CORTES, SAMANTHA WELSH, & GIL MCVEAN, S. L., PETER DONNELLY, JONATHAN MARCHINI. 2017. Genome-wide genetic data on ~500,000 UK Biobank participants.

CLEMENT, K., VAISSE, C., LAHLOU, N., CABROL, S., PELLOUX, V., CASSUTO, D., GOURMELEN, M., DINA, C., CHAMBAZ, J., LACORTE, J. M., BASDEVANT, A., BOUGNERES, P., LEBOUC, Y., FROGUEL, P. & GUY-GRAND, B. 1998. A mutation in the human leptin receptor gene causes obesity and pituitary dysfunction. *Nature,* 392**,** 398-401.

CLINICALTRIALS.GOV. 2016a. *Application of New Technologies and Tools to Nutrition Research (NutriTech). Identifier: NCT01684917.* [Online]. Available: https://clinicaltrials.gov/ct2/show/study/NCT01684917 [Accessed 19 May 2018].

CLINICALTRIALS.GOV 2016b. Personalised Medicine for Morbid Obesity. Identifier: NCT01365416.

COLELLA, S., YAU, C., TAYLOR, J. M., MIRZA, G., BUTLER, H., CLOUSTON, P., BASSETT, A. S., SELLER, A., HOLMES, C. C. & RAGOUSSIS, J. 2007. QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Res,* 35**,** 2013-25.

COLEMAN, D. L. & HUMMEL, K. P. 1967. Studies with the mutation, diabetes, in the mouse. *Diabetologia,* 3**,** 238-48.

COLEMAN, D. L. & HUMMEL, K. P. 1973. The influence of genetic background on the expression of the obese (Ob) gene in the mouse. *Diabetologia,* 9**,** 287-93.

COMUZZIE, A. G., FUNAHASHI, T., SONNENBERG, G., MARTIN, L. J., JACOB, H. J., BLACK, A. E., MAAS, D., TAKAHASHI, M., KIHARA, S., TANAKA, S., MATSUZAWA, Y., BLANGERO, J., COHEN, D. & KISSEBAH, A. 2001. The genetic basis of plasma variation in adiponectin, a global endophenotype for obesity and the metabolic syndrome. *J Clin Endocrinol Metab,* 86**,** 4321-5.

COOPER, G. M., NICKERSON, D. A. & EICHLER, E. E. 2007. Mutational and selective effects on copy-number variants in the human genome. *Nat Genet,* 39**,** S22-9.

COPPINGER, J., MCDONALD-MCGINN, D., ZACKAI, E., SHANE, K., ATKIN, J. F., ASAMOAH, A., LELAND, R., WEAVER, D. D., LANSKY-SHAFER, S., SCHMIDT, K., FELDMAN, H., COHEN, W., PHALIN, J., POWELL, B., BALLIF, B. C., THEISEN, A., GEIGER, E., HALDEMAN-ENGLERT, C., SHAIKH, T. H., SAITTA, S., BEJJANI, B. A. & SHAFFER, L. G. 2009. Identification of familial and de novo microduplications of 22q11.21-q11.23 distal to the 22q11.21 microdeletion syndrome region. *Hum Mol Genet,* 18**,** 1377-83.

COURTENS, W., SCHRAMME, I. & LARIDON, A. 2008. Microduplication 22q11.2: a benign polymorphism or a syndrome with a very large clinical variability and reduced penetrance?--Report of two families. *Am J Med Genet A,* 146A**,** 758-63.

DE SMITH, A. J., PURMANN, C., WALTERS, R. G., ELLIS, R. J., HOLDER, S. E., VAN HAELST, M. M., BRADY, A. F., FAIRBROTHER, U. L., DATTANI, M., KEOGH, J. M., HENNING, E., YEO, G. S., O'RAHILLY, S., FROGUEL, P., FAROOQI, I. S. & BLAKEMORE, A. I. 2009. A deletion of the HBII-85 class of small nucleolar RNAs (snoRNAs) is associated with hyperphagia, obesity and hypogonadism. *Hum Mol Genet,* 18**,** 3257-65.

DE SMITH, A. J., TREWICK, A. L. & BLAKEMORE, A. I. 2010. Implications of copy number variation in people with chromosomal abnormalities: potential for greater variation in copy number state may contribute to variability of phenotype. *Hugo J,* 4**,** 1-9.

DE SMITH, A. J., TSALENKO, A., SAMPAS, N., SCHEFFER, A., YAMADA, N. A., TSANG, P., BEN-DOR, A., YAKHINI, Z., ELLIS, R. J., BRUHN, L., LADERMAN, S., FROGUEL, P. & BLAKEMORE, A. I. 2007. Array CGH analysis of copy number variation identifies 1284 new genes variant in healthy white males: implications for association studies of complex diseases. *Hum Mol Genet,* 16**,** 2783-94.

DISKIN, S. J., LI, M., HOU, C., YANG, S., GLESSNER, J., HAKONARSON, H., BUCAN, M., MARIS, J. M. & WANG, K. 2008. Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. *Nucleic Acids Res,* 36**,** e126.

DUDOIT, S., SHAFFER, J. P. & BOLDRICK, J. C. 2003. Multiple hypothesis testing in microarray experiments. *Statistical Science,* 18**,** 71-103.

DURAND, E., BOUTIN, P., MEYRE, D., CHARLES, M. A., CLEMENT, K., DINA, C. & FROGUEL, P. 2004. Polymorphisms in the amino acid transporter solute carrier family 6 (neurotransmitter transporter) member 14 gene contribute to polygenic obesity in French Caucasians. *Diabetes,* 53**,** 2483-6.

EDWARDS, T. L. & GAO, X. 2012. Methods for detecting and correcting for population stratification. *Curr Protoc Hum Genet,* Chapter 1**,** Unit 1 22 1-14.

EICHLER, E. E. 2001. Recent duplication, domain accretion and the dynamic mutation of the human genome. *Trends Genet,* 17**,** 661-9.

ELLIOTT, P., PEAKMAN, T. C. & BIOBANK, U. K. 2008. The UK Biobank sample handling and storage protocol for the collection, processing and archiving of human blood and urine. *Int J Epidemiol,* 37**,** 234-44.

FALCHI, M., EL-SAYED MOUSTAFA, J. S., TAKOUSIS, P., PESCE, F., BONNEFOND, A., ANDERSSON-ASSARSSON, J. C., SUDMANT, P. H., DORAJOO, R., AL-SHAFAI, M. N.,

BOTTOLO, L., OZDEMIR, E., SO, H. C., DAVIES, R. W., PATRICE, A., DENT, R., MANGINO, M., HYSI, P. G., DECHAUME, A., HUYVAERT, M., SKINNER, J., PIGEYRE, M., CAIAZZO, R., RAVERDY, V., VAILLANT, E., FIELD, S., BALKAU, B., MARRE, M., VISVIKIS-SIEST, S., WEILL, J., POULAIN-GODEFROY, O., JACOBSON, P., SJOSTROM, L., HAMMOND, C. J., DELOUKAS, P., SHAM, P. C., MCPHERSON, R., LEE, J., TAI, E. S., SLADEK, R., CARLSSON, L. M., WALLEY, A., EICHLER, E. E., PATTOU, F., SPECTOR, T. D. & FROGUEL, P. 2014. Low copy number of the salivary amylase gene predisposes to obesity. *Nat Genet,* 46**,** 492-7.

FAN, C. M., KUWANA, E., BULFONE, A., FLETCHER, C. F., COPELAND, N. G., JENKINS, N. A., CREWS, S., MARTINEZ, S., PUELLES, L., RUBENSTEIN, J. L. & TESSIER-LAVIGNE, M. 1996. Expression patterns of two murine homologs of Drosophila single-minded suggest possible roles in embryonic patterning and in the pathogenesis of Down syndrome. *Mol Cell Neurosci,* 7**,** 1-16.

FAROOQI, I. S. 2006. The severely obese patient--a genetic work-up. *Nat Clin Pract Endocrinol Metab,* 2**,** 172-7; quiz following 177.

FAROOQI, I. S., MATARESE, G., LORD, G. M., KEOGH, J. M., LAWRENCE, E., AGWU, C., SANNA, V., JEBB, S. A., PERNA, F., FONTANA, S., LECHLER, R. I., DEPAOLI, A. M. & O'RAHILLY, S. 2002. Beneficial effects of leptin on obesity, T cell hyporesponsiveness, and neuroendocrine/metabolic dysfunction of human congenital leptin deficiency. *J Clin Invest,* 110**,** 1093-103.

FAROOQI, I. S. & O'RAHILLY, S. 2009. Leptin: a pivotal regulator of human energy homeostasis. *Am J Clin Nutr,* 89**,** 980S-984S.

FAROOQI, I. S., WANGENSTEEN, T., COLLINS, S., KIMBER, W., MATARESE, G., KEOGH, J. M., LANK, E., BOTTOMLEY, B., LOPEZ-FERNANDEZ, J., FERRAZ-AMARO, I., DATTANI, M. T., ERCAN, O., MYHRE, A. G., RETTERSTOL, L., STANHOPE, R., EDGE, J. A., MCKENZIE, S., LESSAN, N., GHODSI, M., DE ROSA, V., PERNA, F., FONTANA, S., BARROSO, I., UNDLIEN, D. E. & O'RAHILLY, S. 2007. Clinical and molecular genetic spectrum of congenital deficiency of the leptin receptor. *N Engl J Med,* 356**,** 237-47.

FERNANDEZ, B. A., ROBERTS, W., CHUNG, B., WEKSBERG, R., MEYN, S., SZATMARI, P., JOSEPH-GEORGE, A. M., MACKAY, S., WHITTEN, K., NOBLE, B., VARDY, C., CROSBIE, V., LUSCOMBE, S., TUCKER, E., TURNER, L., MARSHALL, C. R. & SCHERER, S. W. 2010. Phenotypic spectrum associated with de novo and inherited deletions and duplications at 16p11.2 in individuals ascertained for diagnosis of autism spectrum disorder. *J Med Genet,* 47**,** 195-203.

FEUK, L., CARSON, A. R. & SCHERER, S. W. 2006. Structural variation in the human genome. *Nat Rev Genet,* 7**,** 85-97.

FIRTH, H. V., RICHARDS, S. M., BEVAN, A. P., CLAYTON, S., CORPAS, M., RAJAN, D., VAN VOOREN, S., MOREAU, Y., PETTETT, R. M. & CARTER, N. P. 2009. DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *Am J Hum Genet,* 84**,** 524-33.

FOX, C. S., HEARD-COSTA, N., CUPPLES, L. A., DUPUIS, J., VASAN, R. S. & ATWOOD, L. D. 2007. Genome-wide association to body mass index and waist circumference: the Framingham Heart Study 100K project. *BMC Med Genet,* 8 Suppl 1**,** S18.

FRAYLING, T. M., TIMPSON, N. J., WEEDON, M. N., ZEGGINI, E., FREATHY, R. M., LINDGREN, C. M., PERRY, J. R., ELLIOTT, K. S., LANGO, H., RAYNER, N. W., SHIELDS, B., HARRIES, L. W., BARRETT, J. C., ELLARD, S., GROVES, C. J., KNIGHT, B., PATCH, A. M., NESS, A. R., EBRAHIM, S., LAWLOR, D. A., RING, S. M., BEN-SHLOMO, Y., JARVELIN, M. R., SOVIO,

U., BENNETT, A. J., MELZER, D., FERRUCCI, L., LOOS, R. J., BARROSO, I., WAREHAM, N. J., KARPE, F., OWEN, K. R., CARDON, L. R., WALKER, M., HITMAN, G. A., PALMER, C. N., DONEY, A. S., MORRIS, A. D., SMITH, G. D., HATTERSLEY, A. T. & MCCARTHY, M. I. 2007. A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science,* 316**,** 889-94.

FRIEDEL, S., HORRO, F. F., WERMTER, A. K., GELLER, F., DEMPFLE, A., REICHWALD, K., SMIDT, J., BRONNER, G., KONRAD, K., HERPERTZ-DAHLMANN, B., WARNKE, A., HEMMINGER, U., LINDER, M., KIEFL, H., GOLDSCHMIDT, H. P., SIEGFRIED, W., REMSCHMIDT, H., HINNEY, A. & HEBEBRAND, J. 2005. Mutation screen of the brain derived neurotrophic factor gene (BDNF): identification of several genetic variants and association studies in patients with obesity, eating disorders, and attention-deficit/hyperactivity disorder. *Am J Med Genet B Neuropsychiatr Genet,* 132B**,** 96-9.

FRIEDMAN, K. E., REICHMANN, S. K., COSTANZO, P. R. & MUSANTE, G. J. 2002. Body image partially mediates the relationship between obesity and psychological distress. *Obes Res,* 10**,** 33-41.

FRIEDRICH, M. J. 2017. Global Obesity Epidemic Worsening. *JAMA,* 318**,** 603.

GALLAGHER, D., HEYMSFIELD, S. B., HEO, M., JEBB, S. A., MURGATROYD, P. R. & SAKAMOTO, Y. 2000. Healthy percentage body fat ranges: an approach for developing guidelines based on body mass index. *Am J Clin Nutr,* 72**,** 694-701.

GE, T., CHEN, C. Y., NEALE, B. M., SABUNCU, M. R. & SMOLLER, J. W. 2018. Correction: Phenome-wide heritability analysis of the UK Biobank. *PLoS Genet,* 14**,** e1007228.

GENOMES PROJECT, C., AUTON, A., BROOKS, L. D., DURBIN, R. M., GARRISON, E. P., KANG, H. M., KORBEL, J. O., MARCHINI, J. L., MCCARTHY, S., MCVEAN, G. A. & ABECASIS, G. R. 2015. A global reference for human genetic variation. *Nature,* 526**,** 68-74.

GIBBONS, H., MICHIELSEN, C. J. R., RUNDLE, M., FROST, G., MCNULTY, B. A., NUGENT, A. P., WALTON, J., FLYNN, A., GIBNEY, M. J. & BRENNAN, L. 2017. Demonstration of the utility of biomarkers for dietary intake assessment; proline betaine as an example. *Mol Nutr Food Res,* 61.

GOLDSTEIN, J. I., CRENSHAW, A., CAREY, J., GRANT, G. B., MAGUIRE, J., FROMER, M., O'DUSHLAINE, C., MORAN, J. L., CHAMBERT, K., STEVENS, C., SWEDISH SCHIZOPHRENIA, C., CONSORTIUM, A. A. S., SKLAR, P., HULTMAN, C. M., PURCELL, S., MCCARROLL, S. A., SULLIVAN, P. F., DALY, M. J. & NEALE, B. M. 2012. zCall: a rare variant caller for array-based genotyping: genetics and population analysis. *Bioinformatics,* 28**,** 2543-5.

GOODARZI, M. O. 2018. Genetics of obesity: what genetic association studies have taught us about the biology of obesity and its complications. *Lancet Diabetes Endocrinol,* 6**,** 223-236.

GROOT, P. C., BLEEKER, M. J., PRONK, J. C., ARWERT, F., MAGER, W. H., PLANTA, R. J., ERIKSSON, A. W. & FRANTS, R. R. 1989. The human alpha-amylase multigene family consists of haplotypes with variable numbers of genes. *Genomics,* 5**,** 29-42.

GROOT, P. C., MAGER, W. H. & FRANTS, R. R. 1991. Interpretation of polymorphic DNA patterns in the human alpha-amylase multigene family. *Genomics,* 10**,** 779-85.

GUBBAY, J., COLLIGNON, J., KOOPMAN, P., CAPEL, B., ECONOMOU, A., MUNSTERBERG, A., VIVIAN, N., GOODFELLOW, P. & LOVELL-BADGE, R. 1990. A gene mapping to the sex-determining region of the mouse Y chromosome is a member of a novel family of embryonically expressed genes. *Nature,* 346**,** 245-50.

GUO, Y., HE, J., ZHAO, S., WU, H., ZHONG, X., SHENG, Q., SAMUELS, D. C., SHYR, Y. & LONG, J. 2014. Illumina human exome genotyping array clustering and quality control. *Nat Protoc,* 9**,** 2643-62.

GUY, C., WANG, X., LU, X., LU, J. & LI, S. 2015. Two patients with small chromosome 22q11.21 alterations and central nervous system abnormalities. *Mol Cytogenet,* 8**,** 102.

HAMILTON, D., DEE, A. & PERRY, I. J. 2018. The lifetime costs of overweight and obesity in childhood and adolescence: a systematic review. *Obes Rev,* 19**,** 452-463.

HARTIKAINEN-SORRI, A. L., RANTAKALLIO, P. & SIPILA, P. 1990. Changes in prognosis of twin births over 20 years. *Ann Med,* 22**,** 131-5.

HASTINGS, P. J., LUPSKI, J. R., ROSENBERG, S. M. & IRA, G. 2009. Mechanisms of change in gene copy number. *Nat Rev Genet,* 10**,** 551-64.

HENRICHSEN, C. N., VINCKENBOSCH, N., ZOLLNER, S., CHAIGNAT, E., PRADERVAND, S., SCHUTZ, F., RUEDI, M., KAESSMANN, H. & REYMOND, A. 2009. Segmental copy number variation shapes tissue transcriptomes. *Nat Genet,* 41**,** 424-9.

HERRERA, B. M. & LINDGREN, C. M. 2010. The genetics of obesity. *Curr Diab Rep,* 10**,** 498-505.

HINDORFF, L. A., SETHUPATHY, P., JUNKINS, H. A., RAMOS, E. M., MEHTA, J. P., COLLINS, F. S. & MANOLIO, T. A. 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A,* 106**,** 9362-7.

HINDSON, C. M., CHEVILLET, J. R., BRIGGS, H. A., GALLICHOTTE, E. N., RUF, I. K., HINDSON, B. J., VESSELLA, R. L. & TEWARI, M. 2013. Absolute quantification by droplet digital PCR versus analog real-time PCR. *Nat Methods,* 10**,** 1003-5.

HINNEY, A., NGUYEN, T. T., SCHERAG, A., FRIEDEL, S., BRONNER, G., MULLER, T. D., GRALLERT, H., ILLIG, T., WICHMANN, H. E., RIEF, W., SCHAFER, H. & HEBEBRAND, J. 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**,** e1361.

HINRICHS, A. S., KAROLCHIK, D., BAERTSCH, R., BARBER, G. P., BEJERANO, G., CLAWSON, H., DIEKHANS, M., FUREY, T. S., HARTE, R. A., HSU, F., HILLMAN-JACKSON, J., KUHN, R. M., PEDERSEN, J. S., POHL, A., RANEY, B. J., ROSENBLOOM, K. R., SIEPEL, A., SMITH, K. E., SUGNET, C. W., SULTAN-QURRAIE, A., THOMAS, D. J., TRUMBOWER, H., WEBER, R. J., WEIRAUCH, M., ZWEIG, A. S., HAUSSLER, D. & KENT, W. J. 2006. The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res,* 34**,** D590-8.

HIRANO, S. & TAKEICHI, M. 2012. Cadherins in brain morphogenesis and wiring. *Physiol Rev,* 92**,** 597-634.

HOEFFDING, L. K., TRABJERG, B. B., OLSEN, L., MAZIN, W., SPARSO, T., VANGKILDE, A., MORTENSEN, P. B., PEDERSEN, C. B. & WERGE, T. 2017. Risk of Psychiatric Disorders Among Individuals With the 22q11.2 Deletion or Duplication: A Danish Nationwide, Register-Based Study. *JAMA Psychiatry,* 74**,** 282-290.

HOLDER, J. L., JR., BUTTE, N. F. & ZINN, A. R. 2000. Profound obesity associated with a balanced translocation that disrupts the SIM1 gene. *Hum Mol Genet,* 9**,** 101-8.

HOLLOX, E. J., HUFFMEIER, U., ZEEUWEN, P. L., PALLA, R., LASCORZ, J., RODIJK-OLTHUIS, D., VAN DE KERKHOF, P. C., TRAUPE, H., DE JONGH, G., DEN HEIJER, M., REIS, A., ARMOUR, J. A. & SCHALKWIJK, J. 2008. Psoriasis is associated with increased beta-defensin genomic copy number. *Nat Genet,* 40**,** 23-5.

HU, F. B., MANSON, J. E., STAMPFER, M. J., COLDITZ, G., LIU, S., SOLOMON, C. G. & WILLETT, W. C. 2001. Diet, lifestyle, and the risk of type 2 diabetes mellitus in women. *N Engl J Med,* 345**,** 790-7.

HUBERT, H. B., FEINLEIB, M., MCNAMARA, P. M. & CASTELLI, W. P. 1983. Obesity as an independent risk factor for cardiovascular disease: a 26-year follow-up of participants in the Framingham Heart Study. *Circulation,* 67**,** 968-77.

HULPIAU, P. & VAN ROY, F. 2009. Molecular evolution of the cadherin superfamily. *Int J Biochem Cell Biol,* 41**,** 349-69.

HYSI, P. G., YOUNG, T. L., MACKEY, D. A., ANDREW, T., FERNANDEZ-MEDARDE, A., SOLOUKI, A. M., HEWITT, A. W., MACGREGOR, S., VINGERLING, J. R., LI, Y. J., IKRAM, M. K., FAI, L. Y., SHAM, P. C., MANYES, L., PORTEROS, A., LOPES, M. C., CARBONARO, F., FAHY, S. J., MARTIN, N. G., VAN DUIJN, C. M., SPECTOR, T. D., RAHI, J. S., SANTOS, E., KLAVER, C. C. & HAMMOND, C. J. 2010. A genome-wide association study for myopia and refractive error identifies a susceptibility locus at 15q25. *Nat Genet,* 42**,** 902-5.

IAFRATE, A. J., FEUK, L., RIVERA, M. N., LISTEWNIK, M. L., DONAHOE, P. K., QI, Y., SCHERER, S. W. & LEE, C. 2004. Detection of large-scale variation in the human genome. *Nat Genet,* 36**,** 949-51.

ILLUMINA, I. 2014. *Infinium Genotyping Data Analysis: A guide for analyzing Infinium genotyping data using the GenomeStudio Genotyping Module* [Online]. Illumina, Inc. Available: https://www.illumina.com/Documents/products/technotes/technote_infinium_genotyping_data_analysis.pdf [Accessed 25 April 2017].

INTERNATIONAL HAPMAP, C. 2003. The International HapMap Project. *Nature,* 426**,** 789-96.

INTERNATIONAL HAPMAP, C. 2005. A haplotype map of the human genome. *Nature,* 437**,** 1299-320.

INTERNATIONAL HUMAN GENOME SEQUENCING, C. 2004. Finishing the euchromatic sequence of the human genome. *Nature,* 431**,** 931-45.

JACKSON, R. S., CREEMERS, J. W., OHAGI, S., RAFFIN-SANSON, M. L., SANDERS, L., MONTAGUE, C. T., HUTTON, J. C. & O'RAHILLY, S. 1997. Obesity and impaired prohormone processing associated with mutations in the human prohormone convertase 1 gene. *Nat Genet,* 16**,** 303-6.

JACKSON, S. E. & STEPTOE, A. 2017. Association between perceived weight discrimination and physical activity: a population-based study among English middle-aged and older adults. *BMJ Open,* 7**,** e014592.

JACQUEMONT, S., REYMOND, A., ZUFFEREY, F., HAREWOOD, L., WALTERS, R. G., KUTALIK, Z., MARTINET, D., SHEN, Y., VALSESIA, A., BECKMANN, N. D., THORLEIFSSON, G., BELFIORE, M., BOUQUILLON, S., CAMPION, D., DE LEEUW, N., DE VRIES, B. B., ESKO, T., FERNANDEZ, B. A., FERNANDEZ-ARANDA, F., FERNANDEZ-REAL, J. M., GRATACOS, M., GUILMATRE, A., HOYER, J., JARVELIN, M. R., KOOY, R. F., KURG, A., LE CAIGNEC, C., MANNIK, K., PLATT, O. S., SANLAVILLE, D., VAN HAELST, M. M., VILLATORO GOMEZ, S., WALHA, F., WU, B. L., YU, Y., ABOURA, A., ADDOR, M. C., ALEMBIK, Y., ANTONARAKIS, S. E., ARVEILER, B., BARTH, M., BEDNAREK, N., BENA, F., BERGMANN, S., BERI, M., BERNARDINI, L., BLAUMEISER, B., BONNEAU, D., BOTTANI, A., BOUTE, O., BRUNNER, H. G., CAILLEY, D., CALLIER, P., CHIESA, J., CHRAST, J., COIN, L., COUTTON, C., CUISSET, J. M., CUVELLIER, J. C., DAVID, A., DE FREMINVILLE, B., DELOBEL, B., DELRUE, M. A., DEMEER, B., DESCAMPS, D., DIDELOT, G., DIETERICH, K., DISCIGLIO, V., DOCO-FENZY, M., DRUNAT, S., DUBAN-BEDU, B., DUBOURG, C., EL-

SAYED MOUSTAFA, J. S., ELLIOTT, P., FAAS, B. H., FAIVRE, L., FAUDET, A., FELLMANN, F., FERRARINI, A., FISHER, R., FLORI, E., FORER, L., GAILLARD, D., GERARD, M., GIEGER, C., GIMELLI, S., GIMELLI, G., GRABE, H. J., GUICHET, A., GUILLIN, O., HARTIKAINEN, A. L., HERON, D., HIPPOLYTE, L., HOLDER, M., HOMUTH, G., ISIDOR, B., JAILLARD, S., JAROS, Z., JIMENEZ-MURCIA, S., HELAS, G. J., *et al.* 2011. Mirror extreme BMI phenotypes associated with gene dosage at the chromosome 16p11.2 locus. *Nature,* 478**,** 97-102.

JAMES, P. T., LEACH, R., KALAMARA, E. & SHAYEGHI, M. 2001. The worldwide obesity epidemic. *Obes Res,* 9 Suppl 4**,** 228S-233S.

JAMES, W. P. T. & MCPHERSON, K. 2017. The costs of overweight. *Lancet Public Health,* 2**,** e203-e204.

JANSSEN, I., HEYMSFIELD, S. B., ALLISON, D. B., KOTLER, D. P. & ROSS, R. 2002. Body mass index and waist circumference independently contribute to the prediction of nonabdominal, abdominal subcutaneous, and visceral fat. *Am J Clin Nutr,* 75**,** 683-8.

JARVELIN, M. R., ELLIOTT, P., KLEINSCHMIDT, I., MARTUZZI, M., GRUNDY, C., HARTIKAINEN, A. L. & RANTAKALLIO, P. 1997. Ecological and individual predictors of birthweight in a northern Finland birth cohort 1986. *Paediatr Perinat Epidemiol,* 11**,** 298-312.

KALLEN, B., MASTROIACOVO, P. & ROBERT, E. 1996. Major congenital malformations in Down syndrome. *Am J Med Genet,* 65**,** 160-6.

KAN, Y. W. & DOZY, A. M. 1978. Polymorphism of DNA sequence adjacent to human beta-globin structural gene: relationship to sickle mutation. *Proc Natl Acad Sci U S A,* 75**,** 5631-5.

KANDURI, C., UKKOLA-VUOTI, L., OIKKONEN, J., BUCK, G., BLANCHER, C., RAIJAS, P., KARMA, K., LAHDESMAKI, H. & JARVELA, I. 2013. The genome-wide landscape of copy number variations in the MUSGEN study provides evidence for a founder effect in the isolated Finnish population. *Eur J Hum Genet,* 21**,** 1411-6.

KAROLCHIK, D., HINRICHS, A. S., FUREY, T. S., ROSKIN, K. M., SUGNET, C. W., HAUSSLER, D. & KENT, W. J. 2004. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res,* 32**,** D493-6.

KAUR, Y., DE SOUZA, R. J., GIBSON, W. T. & MEYRE, D. 2017. A systematic review of genetic syndromes with obesity. *Obes Rev,* 18**,** 603-634.

KENDALL, K. M., REES, E., ESCOTT-PRICE, V., EINON, M., THOMAS, R., HEWITT, J., O'DONOVAN, M. C., OWEN, M. J., WALTERS, J. T. R. & KIROV, G. 2017. Cognitive Performance Among Carriers of Pathogenic Copy Number Variants: Analysis of 152,000 UK Biobank Subjects. *Biol Psychiatry,* 82**,** 103-110.

KENT WJ SC, F. T., ROSKIN KM, PRINGLE TH, ZAHLER AM, HAUSSLER D. 2002. The human genome browser at UCSC. *Genome Res.,* 12**,** 996-1006.

KEYS, A., FIDANZA, F., KARVONEN, M. J., KIMURA, N. & TAYLOR, H. L. 2014. Indices of relative weight and obesity. *Int J Epidemiol,* 43**,** 655-65.

KOOPMAN, P., GUBBAY, J., VIVIAN, N., GOODFELLOW, P. & LOVELL-BADGE, R. 1991. Male development of chromosomally female mice transgenic for Sry. *Nature,* 351**,** 117-21.

KORBEL, J. O., URBAN, A. E., AFFOURTIT, J. P., GODWIN, B., GRUBERT, F., SIMONS, J. F., KIM, P. M., PALEJEV, D., CARRIERO, N. J., DU, L., TAILLON, B. E., CHEN, Z., TANZER, A., SAUNDERS, A. C., CHI, J., YANG, F., CARTER, N. P., HURLES, M. E., WEISSMAN, S. M., HARKINS, T. T., GERSTEIN, M. B., EGHOLM, M. & SNYDER, M. 2007. Paired-end mapping reveals extensive structural variation in the human genome. *Science,* 318**,** 420-6.

KORN, J. M., KURUVILLA, F. G., MCCARROLL, S. A., WYSOKER, A., NEMESH, J., CAWLEY, S., HUBBELL, E., VEITCH, J., COLLINS, P. J., DARVISHI, K., LEE, C., NIZZARI, M. M., GABRIEL, S. B., PURCELL, S., DALY, M. J. & ALTSHULER, D. 2008. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat Genet,* 40**,** 1253-60.

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**,** 155-7.

KUMAR, R. A., KARAMOHAMED, S., SUDI, J., CONRAD, D. F., BRUNE, C., BADNER, J. A., GILLIAM, T. C., NOWAK, N. J., COOK, E. H., JR., DOBYNS, W. B. & CHRISTIAN, S. L. 2008. Recurrent 16p11.2 microdeletions in autism. *Hum Mol Genet,* 17**,** 628-38.

KURODA, Y., OHASHI, I., TOMINAGA, M., SAITO, T., NAGAI, J., IDA, K., NARUTO, T., MASUNO, M. & KUROSAWA, K. 2014. De novo duplication of 17p13.1-p13.2 in a patient with intellectual disability and obesity. *Am J Med Genet A,* 164A**,** 1550-4.

LACARIA, M., SAHA, P., POTOCKI, L., BI, W., YAN, J., GIRIRAJAN, S., BURNS, B., ELSEA, S., WALZ, K., CHAN, L., LUPSKI, J. R. & GU, W. 2012. A duplication CNV that conveys traits reciprocal to metabolic syndrome and protects against diet-induced obesity in mice and men. *PLoS Genet,* 8**,** e1002713.

LEE, J. A., CARVALHO, C. M. & LUPSKI, J. R. 2007. A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. *Cell,* 131**,** 1235-47.

LEMES, R. B., NUNES, K., CARNAVALLI, J. E. P., KIMURA, L., MINGRONI-NETTO, R. C., MEYER, D. & OTTO, P. A. 2018. Inbreeding estimates in human populations: Applying new approaches to an admixed Brazilian isolate. *PLoS One,* 13**,** e0196360.

LI, C., BEROUKHIM, R., WEIR, B. A., WINCKLER, W., GARRAWAY, L. A., SELLERS, W. R. & MEYERSON, M. 2008. Major copy proportion analysis of tumor samples using SNP arrays. *BMC Bioinformatics,* 9**,** 204.

LITTLEJOHNS, T. J., SUDLOW, C., ALLEN, N. E. & COLLINS, R. 2017. UK Biobank: opportunities for cardiovascular research. *Eur Heart J*.

LOCKE, A. E., KAHALI, B., BERNDT, S. I., JUSTICE, A. E., PERS, T. H., DAY, F. R., POWELL, C., VEDANTAM, S., BUCHKOVICH, M. L., YANG, J., CROTEAU-CHONKA, D. C., ESKO, T., FALL, T., FERREIRA, T., GUSTAFSSON, S., KUTALIK, Z., LUAN, J., MAGI, R., RANDALL, J. C., WINKLER, T. W., WOOD, A. R., WORKALEMAHU, T., FAUL, J. D., SMITH, J. A., ZHAO, J. H., ZHAO, W., CHEN, J., FEHRMANN, R., HEDMAN, A. K., KARJALAINEN, J., SCHMIDT, E. M., ABSHER, D., AMIN, N., ANDERSON, D., BEEKMAN, M., BOLTON, J. L., BRAGG-GRESHAM, J. L., BUYSKE, S., DEMIRKAN, A., DENG, G., EHRET, G. B., FEENSTRA, B., FEITOSA, M. F., FISCHER, K., GOEL, A., GONG, J., JACKSON, A. U., KANONI, S., KLEBER, M. E., KRISTIANSSON, K., LIM, U., LOTAY, V., MANGINO, M., LEACH, I. M., MEDINA-GOMEZ, C., MEDLAND, S. E., NALLS, M. A., PALMER, C. D., PASKO, D., PECHLIVANIS, S., PETERS, M. J., PROKOPENKO, I., SHUNGIN, D., STANCAKOVA, A., STRAWBRIDGE, R. J., SUNG, Y. J., TANAKA, T., TEUMER, A., TROMPET, S., VAN DER LAAN, S. W., VAN SETTEN, J., VAN VLIET-OSTAPTCHOUK, J. V., WANG, Z., YENGO, L., ZHANG, W., ISAACS, A., ALBRECHT, E., ARNLOV, J., ARSCOTT, G. M., ATTWOOD, A. P., BANDINELLI, S., BARRETT, A., BAS, I. N., BELLIS, C., BENNETT, A. J., BERNE, C., BLAGIEVA, R., BLUHER, M., BOHRINGER, S., BONNYCASTLE, L. L., BOTTCHER, Y., BOYD, H. A., BRUINENBERG, M., CASPERSEN, I. H., CHEN, Y. I., CLARKE, R., DAW, E. W., DE CRAEN, A. J. M., DELGADO, G., DIMITRIOU, M., *et al.* 2015.

Genetic studies of body mass index yield new insights for obesity biology. *Nature,* 518**,** 197-206.

LOOS, R. J. 2018. The genetics of adiposity. *Curr Opin Genet Dev,* 50**,** 86-95.

LOOS, R. J. F. & JANSSENS, A. 2017. Predicting Polygenic Obesity Using Genetic Information. *Cell Metab,* 25**,** 535-543.

LOVELL-BADGE, R. & ROBERTSON, E. 1990. XY female mice resulting from a heritable mutation in the primary testis-determining gene, Tdy. *Development,* 109**,** 635-46.

LOVIGLIO, M. N., LELEU, M., MANNIK, K., PASSEGGERI, M., GIANNUZZI, G., VAN DER WERF, I., WASZAK, S. M., ZAZHYTSKA, M., ROBERTS-CALDEIRA, I., GHELDOF, N., MIGLIAVACCA, E., ALFAIZ, A. A., HIPPOLYTE, L., MAILLARD, A. M., P, C., P, C., VAN DIJCK, A., KOOY, R. F., SANLAVILLE, D., ROSENFELD, J. A., SHAFFER, L. G., ANDRIEUX, J., MARSHALL, C., SCHERER, S. W., SHEN, Y., GUSELLA, J. F., THORSTEINSDOTTIR, U., THORLEIFSSON, G., DERMITZAKIS, E. T., DEPLANCKE, B., BECKMANN, J. S., ROUGEMONT, J., JACQUEMONT, S. & REYMOND, A. 2017. Chromosomal contacts connect loci associated with autism, BMI and head circumference phenotypes. *Mol Psychiatry,* 22**,** 836-849.

MACE, A., TUKE, M. A., BECKMANN, J. S., LIN, L., JACQUEMONT, S., WEEDON, M. N., REYMOND, A. & KUTALIK, Z. 2016. New quality measure for SNP array based CNV detection. *Bioinformatics,* 32**,** 3298-3305.

MACE, A., TUKE, M. A., DEELEN, P., KRISTIANSSON, K., MATTSSON, H., NOUKAS, M., SAPKOTA, Y., SCHICK, U., PORCU, E., RUEGER, S., MCDAID, A. F., PORTEOUS, D., WINKLER, T. W., SALVI, E., SHRINE, N., LIU, X., ANG, W. Q., ZHANG, W., FEITOSA, M. F., VENTURINI, C., VAN DER MOST, P. J., ROSENGREN, A., WOOD, A. R., BEAUMONT, R. N., JONES, S. E., RUTH, K. S., YAGHOOTKAR, H., TYRRELL, J., HAVULINNA, A. S., BOERS, H., MAGI, R., KRIEBEL, J., MULLER-NURASYID, M., PEROLA, M., NIEMINEN, M., LOKKI, M. L., KAHONEN, M., VIIKARI, J. S., GELLER, F., LAHTI, J., PALOTIE, A., KOPONEN, P., LUNDQVIST, A., RISSANEN, H., BOTTINGER, E. P., AFAQ, S., WOJCZYNSKI, M. K., LENZINI, P., NOLTE, I. M., SPARSO, T., SCHUPF, N., CHRISTENSEN, K., PERLS, T. T., NEWMAN, A. B., WERGE, T., SNIEDER, H., SPECTOR, T. D., CHAMBERS, J. C., KOSKINEN, S., MELBYE, M., RAITAKARI, O. T., LEHTIMAKI, T., TOBIN, M. D., WAIN, L. V., SINISALO, J., PETERS, A., MEITINGER, T., MARTIN, N. G., WRAY, N. R., MONTGOMERY, G. W., MEDLAND, S. E., SWERTZ, M. A., VARTIAINEN, E., BORODULIN, K., MANNISTO, S., MURRAY, A., BOCHUD, M., JACQUEMONT, S., RIVADENEIRA, F., HANSEN, T. F., OLDEHINKEL, A. J., MANGINO, M., PROVINCE, M. A., DELOUKAS, P., KOONER, J. S., FREATHY, R. M., PENNELL, C., FEENSTRA, B., STRACHAN, D. P., LETTRE, G., HIRSCHHORN, J., CUSI, D., HEID, I. M., HAYWARD, C., MANNIK, K., BECKMANN, J. S., LOOS, R. J. F., NYHOLT, D. R., METSPALU, A., ERIKSSON, J. G., *et al.* 2017. CNV-association meta-analysis in 191,161 European adults reveals new loci associated with anthropometric traits. *Nat Commun,* 8**,** 744.

MAILLARD, A. M., HIPPOLYTE, L., RODRIGUEZ-HERREROS, B., CHAWNER, S. J., DREMMEL, D., AGUERA, Z., FAGUNDO, A. B., PAIN, A., MARTIN-BREVET, S., HILBERT, A., KURZ, S., ETIENNE, R., DRAGANSKI, B., JIMENEZ-MURCIA, S., MANNIK, K., METSPALU, A., REIGO, A., ISIDOR, B., LE CAIGNEC, C., DAVID, A., MIGNOT, C., KEREN, B., P11.2 EUROPEAN, C., VAN DEN BREE, M. B., MUNSCH, S., FERNANDEZ-ARANDA, F., BECKMANN, J. S., REYMOND, A. & JACQUEMONT, S. 2016. 16p11.2 Locus modulates response to satiety before the onset of obesity. *Int J Obes (Lond),* 40**,** 870-6.

MAILLARD, A. M., RUEF, A., PIZZAGALLI, F., MIGLIAVACCA, E., HIPPOLYTE, L., ADASZEWSKI, S., DUKART, J., FERRARI, C., CONUS, P., MANNIK, K., ZAZHYTSKA, M., SIFFREDI, V., MAEDER, P., KUTALIK, Z., KHERIF, F., HADJIKHANI, N., BECKMANN, J. S., REYMOND, A., DRAGANSKI, B., JACQUEMONT, S. & P11.2 EUROPEAN, C. 2015. The 16p11.2 locus modulates brain structures common to autism, schizophrenia and obesity. *Mol Psychiatry,* 20**,** 140-7.

MANICHAIKUL, A., MYCHALECKYJ, J. C., RICH, S. S., DALY, K., SALE, M. & CHEN, W. M. 2010. Robust relationship inference in genome-wide association studies. *Bioinformatics,* 26**,** 2867-73.

MANOLIO, T. A., COLLINS, F. S., COX, N. J., GOLDSTEIN, D. B., HINDORFF, L. A., HUNTER, D. J., MCCARTHY, M. I., RAMOS, E. M., CARDON, L. R., CHAKRAVARTI, A., CHO, J. H., GUTTMACHER, A. E., KONG, A., KRUGLYAK, L., MARDIS, E., ROTIMI, C. N., SLATKIN, M., VALLE, D., WHITTEMORE, A. S., BOEHNKE, M., CLARK, A. G., EICHLER, E. E., GIBSON, G., HAINES, J. L., MACKAY, T. F., MCCARROLL, S. A. & VISSCHER, P. M. 2009. Finding the missing heritability of complex diseases. *Nature,* 461**,** 747-53.

MANSON, J. E., COLDITZ, G. A., STAMPFER, M. J., WILLETT, W. C., ROSNER, B., MONSON, R. R., SPEIZER, F. E. & HENNEKENS, C. H. 1990. A prospective study of obesity and risk of coronary heart disease in women. *N Engl J Med,* 322**,** 882-9.

MARCOVECCHIO, M. L., FLORIO, R., VERGINELLI, F., DE LELLIS, L., CAPELLI, C., VERZILLI, D., CHIARELLI, F., MOHN, A. & CAMA, A. 2016. Low AMY1 Gene Copy Number Is Associated with Increased Body Mass Index in Prepubertal Boys. *PLoS One,* 11**,** e0154961.

MARSHALL, C. R., NOOR, A., VINCENT, J. B., LIONEL, A. C., FEUK, L., SKAUG, J., SHAGO, M., MOESSNER, R., PINTO, D., REN, Y., THIRUVAHINDRAPDURAM, B., FIEBIG, A., SCHREIBER, S., FRIEDMAN, J., KETELAARS, C. E., VOS, Y. J., FICICIOGLU, C., KIRKPATRICK, S., NICOLSON, R., SLOMAN, L., SUMMERS, A., GIBBONS, C. A., TEEBI, A., CHITAYAT, D., WEKSBERG, R., THOMPSON, A., VARDY, C., CROSBIE, V., LUSCOMBE, S., BAATJES, R., ZWAIGENBAUM, L., ROBERTS, W., FERNANDEZ, B., SZATMARI, P. & SCHERER, S. W. 2008. Structural variation of chromosomes in autism spectrum disorder. *Am J Hum Genet,* 82**,** 477-88.

MATTSSON, E., LARSSON, U. E. & ROSSNER, S. 1997. Is walking for exercise too exhausting for obese women? *Int J Obes Relat Metab Disord,* 21**,** 380-6.

MAYNARD, T. M., MEECHAN, D. W., DUDEVOIR, M. L., GOPALAKRISHNA, D., PETERS, A. Z., HEINDEL, C. C., SUGIMOTO, T. J., WU, Y., LIEBERMAN, J. A. & LAMANTIA, A. S. 2008. Mitochondrial localization and function of a subset of 22q11 deletion syndrome candidate genes. *Mol Cell Neurosci,* 39**,** 439-51.

MCCARROLL, S. A. 2008. Extending genome-wide association studies to copy-number variation. *Hum Mol Genet,* 17**,** R135-42.

MCDONALD-MCGINN, D. M., EMANUEL, B. S. & ZACKAI, E. H. 1993. 22q11.2 Deletion Syndrome. *In:* ADAM, M. P., ARDINGER, H. H., PAGON, R. A., WALLACE, S. E., BEAN, L. J. H., STEPHENS, K. & AMEMIYA, A. (eds.) *GeneReviews((R)).* Seattle (WA).

MEJIA-BENITEZ, M. A., BONNEFOND, A., YENGO, L., HUYVAERT, M., DECHAUME, A., PERALTA-ROMERO, J., KLUNDER-KLUNDER, M., GARCIA MENA, J., EL-SAYED MOUSTAFA, J. S., FALCHI, M., CRUZ, M. & FROGUEL, P. 2015. Beneficial effect of a high number of copies of salivary amylase AMY1 gene on obesity risk in Mexican children. *Diabetologia,* 58**,** 290-4.

MEYRE, D., BOUATIA-NAJI, N., TOUNIAN, A., SAMSON, C., LECOEUR, C., VATIN, V., GHOUSSAINI, M., WACHTER, C., HERCBERG, S., CHARPENTIER, G., PATSCH, W., PATTOU, F., CHARLES, M. A., TOUNIAN, P., CLEMENT, K., JOURET, B., WEILL, J., MADDUX, B. A., GOLDFINE, I. D., WALLEY, A., BOUTIN, P., DINA, C. & FROGUEL, P. 2005. Variants of ENPP1 are associated with childhood and adult obesity and increase the risk of glucose intolerance and type 2 diabetes. *Nat Genet,* 37**,** 863-7.

MIN, J., CHIU, D. T. & WANG, Y. 2013. Variation in the heritability of body mass index based on diverse twin studies: a systematic review. *Obes Rev,* 14**,** 871-82.

MONTAGUE, C. T., FAROOQI, I. S., WHITEHEAD, J. P., SOOS, M. A., RAU, H., WAREHAM, N. J., SEWTER, C. P., DIGBY, J. E., MOHAMMED, S. N., HURST, J. A., CHEETHAM, C. H., EARLEY, A. R., BARNETT, A. H., PRINS, J. B. & O'RAHILLY, S. 1997. Congenital leptin deficiency is associated with severe early-onset obesity in humans. *Nature,* 387**,** 903-8.

MULLER, M. J., BOSY-WESTPHAL, A. & KRAWCZAK, M. 2010. Genetic studies of common types of obesity: a critique of the current use of phenotypes. *Obes Rev,* 11**,** 612-8.

MUST, A., SPADANO, J., COAKLEY, E. H., FIELD, A. E., COLDITZ, G. & DIETZ, W. H. 1999. The disease burden associated with overweight and obesity. *JAMA,* 282**,** 1523-9.

NAPOLI, E., TASSONE, F., WONG, S., ANGKUSTSIRI, K., SIMON, T. J., SONG, G. & GIULIVI, C. 2015. Mitochondrial Citrate Transporter-dependent Metabolic Signature in the 22q11.2 Deletion Syndrome. *J Biol Chem,* 290**,** 23240-53.

NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION (NCBI), C. 1995. Online Mendelian Inheritance in Man, OMIM®. 1995 ed. Baltimore, MD: McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University.

NIKLASSON, L., RASMUSSEN, P., OSKARSDOTTIR, S. & GILLBERG, C. 2009. Autism, ADHD, mental retardation and behavior problems in 100 individuals with 22q11 deletion syndrome. *Res Dev Disabil,* 30**,** 763-73.

O'RAHILLY, S. & FAROOQI, I. S. 2008. Human obesity: a heritable neurobehavioral disorder that is highly sensitive to environmental conditions. *Diabetes,* 57**,** 2905-10.

O'GORMAN, A., GIBBONS, H., RYAN, M. F., GIBNEY, E. R., GIBNEY, M. J., FROST, G. S., ROCHE, H. M. & BRENNAN, L. 2017. Exploring the Links between Diet and Health in an Irish Cohort: A Lipidomic Approach. *Journal of Proteome Research,* 16**,** 1280-1287.

OHTA, T., BUITING, K., KOKKONEN, H., MCCANDLESS, S., HEEGER, S., LEISTI, H., DRISCOLL, D. J., CASSIDY, S. B., HORSTHEMKE, B. & NICHOLLS, R. D. 1999. Molecular mechanism of angelman syndrome in two large families involves an imprinting mutation. *Am J Hum Genet,* 64**,** 385-96.

OSKARSDOTTIR, S., VUJIC, M. & FASTH, A. 2004. Incidence and prevalence of the 22q11 deletion syndrome: a population-based study in Western Sweden. *Arch Dis Child,* 89**,** 148-51.

OU, Z., BERG, J. S., YONATH, H., ENCISO, V. B., MILLER, D. T., PICKER, J., LENZI, T., KEEGAN, C. E., SUTTON, V. R., BELMONT, J., CHINAULT, A. C., LUPSKI, J. R., CHEUNG, S. W., ROEDER, E. & PATEL, A. 2008. Microduplications of 22q11.2 are frequently inherited and are associated with variable phenotypes. *Genet Med,* 10**,** 267-77.

PATERNOSTER, L., STANDL, M., CHEN, C. M., RAMASAMY, A., BONNELYKKE, K., DUIJTS, L., FERREIRA, M. A., ALVES, A. C., THYSSEN, J. P., ALBRECHT, E., BAURECHT, H., FEENSTRA, B., SLEIMAN, P. M., HYSI, P., WARRINGTON, N. M., CURJURIC, I., MYHRE, R., CURTIN, J. A., GROEN-BLOKHUIS, M. M., KERKHOF, M., SAAF, A., FRANKE, A., ELLINGHAUS, D., FOLSTER-HOLST, R., DERMITZAKIS, E., MONTGOMERY, S. B.,

PROKISCH, H., HEIM, K., HARTIKAINEN, A. L., POUTA, A., PEKKANEN, J., BLAKEMORE, A. I., BUXTON, J. L., KAAKINEN, M., DUFFY, D. L., MADDEN, P. A., HEATH, A. C., MONTGOMERY, G. W., THOMPSON, P. J., MATHESON, M. C., LE SOUEF, P., AUSTRALIAN ASTHMA GENETICS, C., ST POURCAIN, B., SMITH, G. D., HENDERSON, J., KEMP, J. P., TIMPSON, N. J., DELOUKAS, P., RING, S. M., WICHMANN, H. E., MULLER-NURASYID, M., NOVAK, N., KLOPP, N., RODRIGUEZ, E., MCARDLE, W., LINNEBERG, A., MENNE, T., NOHR, E. A., HOFMAN, A., UITTERLINDEN, A. G., VAN DUIJN, C. M., RIVADENEIRA, F., DE JONGSTE, J. C., VAN DER VALK, R. J., WJST, M., JOGI, R., GELLER, F., BOYD, H. A., MURRAY, J. C., KIM, C., MENTCH, F., MARCH, M., MANGINO, M., SPECTOR, T. D., BATAILLE, V., PENNELL, C. E., HOLT, P. G., SLY, P., TIESLER, C. M., THIERING, E., ILLIG, T., IMBODEN, M., NYSTAD, W., SIMPSON, A., HOTTENGA, J. J., POSTMA, D., KOPPELMAN, G. H., SMIT, H. A., SODERHALL, C., CHAWES, B., KREINER-MOLLER, E., BISGAARD, H., MELEN, E., BOOMSMA, D. I., CUSTOVIC, A., JACOBSSON, B., PROBST-HENSCH, N. M., PALMER, L. J., GLASS, D., HAKONARSON, H., *et al.* 2011. Meta-analysis of genome-wide association studies identifies three new risk loci for atopic dermatitis. *Nat Genet,* 44**,** 187-92.

PERRY, G. H., DOMINY, N. J., CLAW, K. G., LEE, A. S., FIEGLER, H., REDON, R., WERNER, J., VILLANEA, F. A., MOUNTAIN, J. L., MISRA, R., CARTER, N. P., LEE, C. & STONE, A. C. 2007. Diet and the evolution of human amylase gene copy number variation. *Nat Genet,* 39**,** 1256-60.

PERRY, G. H., TCHINDA, J., MCGRATH, S. D., ZHANG, J., PICKER, S. R., CACERES, A. M., IAFRATE, A. J., TYLER-SMITH, C., SCHERER, S. W., EICHLER, E. E., STONE, A. C. & LEE, C. 2006. Hotspots for copy number variation in chimpanzees and humans. *Proc Natl Acad Sci U S A,* 103**,** 8006-11.

PHE, P. H. E. 2017. *Health matters: obesity and the food environment.* [Online]. Available: https://www.gov.uk/government/publications/health-matters-obesity-and-the-food-environment/health-matters-obesity-and-the-food-environment--2 [Accessed 3 June 2018 2018].

PINOLA, P., LASHEN, H., BLOIGU, A., PUUKKA, K., ULMANEN, M., RUOKONEN, A., MARTIKAINEN, H., POUTA, A., FRANKS, S., HARTIKAINEN, A. L., JARVELIN, M. R. & MORIN-PAPUNEN, L. 2012. Menstrual disorders in adolescence: a marker for hyperandrogenaemia and increased metabolic risks in later life? Finnish general population-based birth cohort study. *Hum Reprod,* 27**,** 3279-86.

PIQUE-REGI, R., MONSO-VARONA, J., ORTEGA, A., SEEGER, R. C., TRICHE, T. J. & ASGHARZADEH, S. 2008. Sparse representation and Bayesian detection of genome copy number alterations from microarray data. *Bioinformatics,* 24**,** 309-18.

POLDERMAN, T. J., BENYAMIN, B., DE LEEUW, C. A., SULLIVAN, P. F., VAN BOCHOVEN, A., VISSCHER, P. M. & POSTHUMA, D. 2015. Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nat Genet,* 47**,** 702-9.

PORTNOI, M. F. 2009. Microduplication 22q11.2: a new chromosomal syndrome. *Eur J Med Genet,* 52**,** 88-93.

POUTA, A., HARTIKAINEN, A. L., SOVIO, U., GISSLER, M., LAITINEN, J., MCCARTHY, M. I., RUOKONEN, A., ELLIOTT, P. & JARVELIN, M. R. 2004. Manifestations of metabolic syndrome after hypertensive pregnancy. *Hypertension,* 43**,** 825-31.

PRZEWORSKI, M., HUDSON, R. R. & DI RIENZO, A. 2000. Adjusting the focus on human variation. *Trends Genet,* 16**,** 296-302.

PURCELL, S., NEALE, B., TODD-BROWN, K., THOMAS, L., FERREIRA, M. A., BENDER, D., MALLER, J., SKLAR, P., DE BAKKER, P. I., DALY, M. J. & SHAM, P. C. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet,* 81**,** 559-75.

R CORE, T. 2013. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

RACETTE, S. B., DEUSINGER, S. S. & DEUSINGER, R. H. 2003. Obesity: overview of prevalence, etiology, and treatment. *Phys Ther,* 83**,** 276-88.

RAHBARI, R., ZUCCHERATO, L. W., TISCHLER, G., CHIHOTA, B., OZTURK, H., SALEEM, S., TARAZONA-SANTOS, E., MACHADO, L. R. & HOLLOX, E. J. 2017. Understanding the Genomic Structure of Copy-Number Variation of the Low-Affinity Fcgamma Receptor Region Allows Confirmation of the Association of FCGR3B Deletion with Rheumatoid Arthritis. *Hum Mutat,* 38**,** 390-399.

RANKINEN, T., ZUBERI, A., CHAGNON, Y. C., WEISNAGEL, S. J., ARGYROPOULOS, G., WALTS, B., PERUSSE, L. & BOUCHARD, C. 2006. The human obesity gene map: the 2005 update. *Obesity (Silver Spring),* 14**,** 529-644.

RANTAKALLIO, P. 1988. The longitudinal study of the northern Finland birth cohort of 1966. *Paediatr Perinat Epidemiol,* 2**,** 59-88.

RANTAKALLIO, P. & OJA, H. 1990. Perinatal risk for infants of unmarried mothers over a period of 20 years. *Early Hum Dev,* 22**,** 157-69.

REDIES, C., HERTEL, N. & HUBNER, C. A. 2012. Cadherins and neuropsychiatric disorders. *Brain Res,* 1470**,** 130-44.

REDON, R., ISHIKAWA, S., FITCH, K. R., FEUK, L., PERRY, G. H., ANDREWS, T. D., FIEGLER, H., SHAPERO, M. H., CARSON, A. R., CHEN, W., CHO, E. K., DALLAIRE, S., FREEMAN, J. L., GONZALEZ, J. R., GRATACOS, M., HUANG, J., KALAITZOPOULOS, D., KOMURA, D., MACDONALD, J. R., MARSHALL, C. R., MEI, R., MONTGOMERY, L., NISHIMURA, K., OKAMURA, K., SHEN, F., SOMERVILLE, M. J., TCHINDA, J., VALSESIA, A., WOODWARK, C., YANG, F., ZHANG, J., ZERJAL, T., ZHANG, J., ARMENGOL, L., CONRAD, D. F., ESTIVILL, X., TYLER-SMITH, C., CARTER, N. P., ABURATANI, H., LEE, C., JONES, K. W., SCHERER, S. W. & HURLES, M. E. 2006. Global variation in copy number in the human genome. *Nature,* 444**,** 444-54.

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**,** 397-406.

ROBERTS, R. E., KAPLAN, G. A., SHEMA, S. J. & STRAWBRIDGE, W. J. 2000. Are the obese at greater risk for depression? *Am J Epidemiol,* 152**,** 163-70.

RONAN, A., FAGAN, K., CHRISTIE, L., CONROY, J., NOWAK, N. J. & TURNER, G. 2007. Familial 4.3 Mb duplication of 21q22 sheds new light on the Down syndrome critical region. *J Med Genet,* 44**,** 448-51.

ROSE, K. M., NEWMAN, B., MAYER-DAVIS, E. J. & SELBY, J. V. 1998. Genetic and behavioral determinants of waist-hip ratio and waist circumference in women twins. *Obes Res,* 6**,** 383-92.

ROSENFELD, J. A., COPPINGER, J., BEJJANI, B. A., GIRIRAJAN, S., EICHLER, E. E., SHAFFER, L. G. & BALLIF, B. C. 2010. Speech delays and behavioral problems are the predominant features in individuals with developmental delays and 16p11.2 microdeletions and microduplications. *J Neurodev Disord,* 2**,** 26-38.

RSTUDIO, T. 2015. RStudio: Integrated Development for R. Boston, MA: RStudio, Inc.

RUKH, G., ERICSON, U., ANDERSSON-ASSARSSON, J., ORHO-MELANDER, M. & SONESTEDT, E. 2017. Dietary starch intake modifies the relation between copy number variation in the salivary amylase gene and BMI. *Am J Clin Nutr,* 106**,** 256-262.

SABATTI, C., SERVICE, S. K., HARTIKAINEN, A. L., POUTA, A., RIPATTI, S., BRODSKY, J., JONES, C. G., ZAITLEN, N. A., VARILO, T., KAAKINEN, M., SOVIO, U., RUOKONEN, A., LAITINEN, J., JAKKULA, E., COIN, L., HOGGART, C., COLLINS, A., TURUNEN, H., GABRIEL, S., ELLIOT, P., MCCARTHY, M. I., DALY, M. J., JARVELIN, M. R., FREIMER, N. B. & PELTONEN, L. 2009. Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nat Genet,* 41**,** 35-46.

SAMUELSON, L. C., PHILLIPS, R. S. & SWANBERG, L. J. 1996. Amylase gene structures in primates: retroposon insertions and promoter evolution. *Mol Biol Evol,* 13**,** 767-79.

SCHNEIDER, M., DEBBANE, M., BASSETT, A. S., CHOW, E. W., FUNG, W. L., VAN DEN BREE, M., OWEN, M., MURPHY, K. C., NIARCHOU, M., KATES, W. R., ANTSHEL, K. M., FREMONT, W., MCDONALD-MCGINN, D. M., GUR, R. E., ZACKAI, E. H., VORSTMAN, J., DUIJFF, S. N., KLAASSEN, P. W., SWILLEN, A., GOTHELF, D., GREEN, T., WEIZMAN, A., VAN AMELSVOORT, T., EVERS, L., BOOT, E., SHASHI, V., HOOPER, S. R., BEARDEN, C. E., JALBRZIKOWSKI, M., ARMANDO, M., VICARI, S., MURPHY, D. G., OUSLEY, O., CAMPBELL, L. E., SIMON, T. J., ELIEZ, S., INTERNATIONAL CONSORTIUM ON, B. & BEHAVIOR IN 22Q11.2 DELETION, S. 2014. Psychiatric disorders from childhood to adulthood in 22q11.2 deletion syndrome: results from the International Consortium on Brain and Behavior in 22q11.2 Deletion Syndrome. *Am J Psychiatry,* 171**,** 627-39.

SEIDELL, J. C. 2010. Waist circumference and waist/hip ratio in relation to all-cause mortality, cancer and sleep apnea. *Eur J Clin Nutr,* 64**,** 35-41.

SHAIKH, T. H., GAI, X., PERIN, J. C., GLESSNER, J. T., XIE, H., MURPHY, K., O'HARA, R., CASALUNOVO, T., CONLIN, L. K., D'ARCY, M., FRACKELTON, E. C., GEIGER, E. A., HALDEMAN-ENGLERT, C., IMIELINSKI, M., KIM, C. E., MEDNE, L., ANNAIAH, K., BRADFIELD, J. P., DABAGHYAN, E., ECKERT, A., ONYIAH, C. C., OSTAPENKO, S., OTIENO, F. G., SANTA, E., SHANER, J. L., SKRABAN, R., SMITH, R. M., ELIA, J., GOLDMUNTZ, E., SPINNER, N. B., ZACKAI, E. H., CHIAVACCI, R. M., GRUNDMEIER, R., RAPPAPORT, E. F., GRANT, S. F., WHITE, P. S. & HAKONARSON, H. 2009. High-resolution mapping and analysis of copy number variations in the human genome: a data resource for clinical and research applications. *Genome Res,* 19**,** 1682-90.

SHINAWI, M., LIU, P., KANG, S. H., SHEN, J., BELMONT, J. W., SCOTT, D. A., PROBST, F. J., CRAIGEN, W. J., GRAHAM, B. H., PURSLEY, A., CLARK, G., LEE, J., PROUD, M., STOCCO, A., RODRIGUEZ, D. L., KOZEL, B. A., SPARAGANA, S., ROEDER, E. R., MCGREW, S. G., KURCZYNSKI, T. W., ALLISON, L. J., AMATO, S., SAVAGE, S., PATEL, A., STANKIEWICZ, P., BEAUDET, A. L., CHEUNG, S. W. & LUPSKI, J. R. 2010. Recurrent reciprocal 16p11.2 rearrangements associated with global developmental delay, behavioural problems, dysmorphism, epilepsy, and abnormal head size. *J Med Genet,* 47**,** 332-41.

SHLIEN, A., BASKIN, B., ACHATZ, M. I., STAVROPOULOS, D. J., NICHOLS, K. E., HUDGINS, L., MOREL, C. F., ADAM, M. P., ZHUKOVA, N., ROTIN, L., NOVOKMET, A., DRUKER, H., SHAGO, M., RAY, P. N., HAINAUT, P. & MALKIN, D. 2010. A common molecular mechanism underlies two phenotypically distinct 17p13.1 microdeletion syndromes. *Am J Hum Genet,* 87**,** 631-42.

SILVERBERG, M. S., CHO, J. H., RIOUX, J. D., MCGOVERN, D. P., WU, J., ANNESE, V., ACHKAR, J. P., GOYETTE, P., SCOTT, R., XU, W., BARMADA, M. M., KLEI, L., DALY, M. J., ABRAHAM, C., BAYLESS, T. M., BOSSA, F., GRIFFITHS, A. M., IPPOLITI, A. F., LAHAIE, R.

G., LATIANO, A., PARE, P., PROCTOR, D. D., REGUEIRO, M. D., STEINHART, A. H., TARGAN, S. R., SCHUMM, L. P., KISTNER, E. O., LEE, A. T., GREGERSEN, P. K., ROTTER, J. I., BRANT, S. R., TAYLOR, K. D., ROEDER, K. & DUERR, R. H. 2009. Ulcerative colitis-risk loci on chromosomes 1p36 and 12q15 found by genome-wide association study. *Nat Genet,* 41**,** 216-20.

SINGER, M. F. 1982. SINEs and LINEs: highly repeated short and long interspersed sequences in mammalian genomes. *Cell,* 28**,** 433-4.

SLATTERY, M. L., SWEENEY, C., EDWARDS, S., HERRICK, J., BAUMGARTNER, K., WOLFF, R., MURTAUGH, M., BAUMGARTNER, R., GIULIANO, A. & BYERS, T. 2007. Body size, weight change, fat distribution and breast cancer risk in Hispanic and non-Hispanic white women. *Breast Cancer Res Treat,* 102**,** 85-101.

SPEAKMAN, J. R., RANCE, K. A. & JOHNSTONE, A. M. 2008. Polymorphisms of the FTO gene are associated with variation in energy intake, but not energy expenditure. *Obesity (Silver Spring),* 16**,** 1961-5.

STANKIEWICZ, P. & LUPSKI, J. R. 2002. Genome architecture, rearrangements and genomic disorders. *Trends Genet,* 18**,** 74-82.

STOLL, B. A. 1999. Perimenopausal weight gain and progression of breast cancer precursors. *Cancer Detect Prev,* 23**,** 31-6.

STROBEL, A., ISSAD, T., CAMOIN, L., OZATA, M. & STROSBERG, A. D. 1998. A leptin missense mutation associated with hypogonadism and morbid obesity. *Nat Genet,* 18**,** 213-5.

STUART, P. E., HUFFMEIER, U., NAIR, R. P., PALLA, R., TEJASVI, T., SCHALKWIJK, J., ELDER, J. T., REIS, A. & ARMOUR, J. A. L. 2012. Association of beta-defensin copy number and psoriasis in three cohorts of European origin. *J Invest Dermatol,* 132**,** 2407-2413.

STUNKARD, A. J., FOCH, T. T. & HRUBEC, Z. 1986a. A twin study of human obesity. *JAMA,* 256**,** 51-4.

STUNKARD, A. J., SORENSEN, T. I., HANIS, C., TEASDALE, T. W., CHAKRABORTY, R., SCHULL, W. J. & SCHULSINGER, F. 1986b. An adoption study of human obesity. *N Engl J Med,* 314**,** 193-8.

SUDMANT, P. H., KITZMAN, J. O., ANTONACCI, F., ALKAN, C., MALIG, M., TSALENKO, A., SAMPAS, N., BRUHN, L., SHENDURE, J., GENOMES, P. & EICHLER, E. E. 2010. Diversity of human copy number variation and multicopy genes. *Science,* 330**,** 641-6.

SUTTON, E., HUGHES, J., WHITE, S., SEKIDO, R., TAN, J., ARBOLEDA, V., ROGERS, N., KNOWER, K., ROWLEY, L., EYRE, H., RIZZOTI, K., MCANINCH, D., GONCALVES, J., SLEE, J., TURBITT, E., BRUNO, D., BENGTSSON, H., HARLEY, V., VILAIN, E., SINCLAIR, A., LOVELL-BADGE, R. & THOMAS, P. 2011. Identification of SOX3 as an XX male sex reversal gene in mice and humans. *J Clin Invest,* 121**,** 328-41.

SUVIOLAHTI, E., OKSANEN, L. J., OHMAN, M., CANTOR, R. M., RIDDERSTRALE, M., TUOMI, T., KAPRIO, J., RISSANEN, A., MUSTAJOKI, P., JOUSILAHTI, P., VARTIAINEN, E., SILANDER, K., KILPIKARI, R., SALOMAA, V., GROOP, L., KONTULA, K., PELTONEN, L. & PAJUKANTA, P. 2003. The SLC6A14 gene shows evidence of association with obesity. *J Clin Invest,* 112**,** 1762-72.

TAPONEN, S., MARTIKAINEN, H., JARVELIN, M. R., SOVIO, U., LAITINEN, J., POUTA, A., HARTIKAINEN, A. L., MCCARTHY, M. I., FRANKS, S., PALDANIUS, M., RUOKONEN, A. & NORTHERN FINLAND BIRTH COHORT, S. 2004. Metabolic cardiovascular disease risk factors in women with self-reported symptoms of oligomenorrhea and/or hirsutism: Northern Finland Birth Cohort 1966 Study. *J Clin Endocrinol Metab,* 89**,** 2114-8.

TEAM, R. C. 2013. R: A language and environment for statistical computing. . Vienna, Austria: R Foundation for Statistical Computing.

TEMEL, S. G., GULTEN, T., YAKUT, T., SAGLAM, H., KILIC, N., BAUSCH, E., JIN, W. J., LEIPOLDT, M. & SCHERER, G. 2007. Extended pedigree with multiple cases of XX sex reversal in the absence of SRY and of a mutation at the SOX9 locus. *Sex Dev,* 1**,** 24-34.

TERRACCIANO, A., SANNA, S., UDA, M., DEIANA, B., USALA, G., BUSONERO, F., MASCHIO, A., SCALLY, M., PATRICIU, N., CHEN, W. M., DISTEL, M. A., SLAGBOOM, E. P., BOOMSMA, D. I., VILLAFUERTE, S., SLIWERSKA, E., BURMEISTER, M., AMIN, N., JANSSENS, A. C., VAN DUIJN, C. M., SCHLESSINGER, D., ABECASIS, G. R. & COSTA, P. T., JR. 2010. Genome-wide association scan for five major dimensions of personality. *Mol Psychiatry,* 15**,** 647-56.

THOMAS, E. L., PARKINSON, J. R., FROST, G. S., GOLDSTONE, A. P., DORE, C. J., MCCARTHY, J. P., COLLINS, A. L., FITZPATRICK, J. A., DURIGHEL, G., TAYLOR-ROBINSON, S. D. & BELL, J. D. 2012. The missing risk: MRI and MRS phenotyping of abdominal adiposity and ectopic fat. *Obesity (Silver Spring),* 20**,** 76-87.

THOMAS, E. L., SAEED, N., HAJNAL, J. V., BRYNES, A., GOLDSTONE, A. P., FROST, G. & BELL, J. D. 1998. Magnetic resonance imaging of total body fat. *J Appl Physiol (1985),* 85**,** 1778-85.

TSURITANI, I., HONDA, R., NOBORISAKA, Y., ISHIDA, M., ISHIZAKI, M. & YAMADA, Y. 2002. Impact of obesity on musculoskeletal pain and difficulty of daily movements in Japanese middle-aged women. *Maturitas,* 42**,** 23-30.

TURULA, M., KAPRIO, J., RISSANEN, A. & KOSKENVUO, M. 1990. Body weight in the Finnish Twin Cohort. *Diabetes Res Clin Pract,* 10 Suppl 1**,** S33-6.

TYLER, C. & EDMAN, J. C. 2004. Down syndrome, Turner syndrome, and Klinefelter syndrome: primary care throughout the life span. *Prim Care,* 31**,** 627-48, x-xi.

UK BIOBANK, C. 2007. UK Biobank: Protocol for a large-scale prospective epidemiological resource. UK Biobank Coordinating Centre.

UK BIOBANK, C. 2014. *UK Biobank Axiom Array Content Summary* [Online]. Available: http://www.ukbiobank.ac.uk/wp-content/uploads/2014/04/UK-Biobank-Axiom-Array-Content-Summary-2014.pdf [Accessed].

UK BIOBANK, C. 2018. *About UK Biobank.* [Online]. UK Biobank. Available: http://www.ukbiobank.ac.uk/about-biobank-uk/ [Accessed 21 May 2018].

USHER, C. L., HANDSAKER, R. E., ESKO, T., TUKE, M. A., WEEDON, M. N., HASTIE, A. R., CAO, H., MOON, J. E., KASHIN, S., FUCHSBERGER, C., METSPALU, A., PATO, C. N., PATO, M. T., MCCARTHY, M. I., BOEHNKE, M., ALTSHULER, D. M., FRAYLING, T. M., HIRSCHHORN, J. N. & MCCARROLL, S. A. 2015. Structural forms of the human amylase locus and their relationships to SNPs, haplotypes and obesity. *Nat Genet,* 47**,** 921-5.

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**,** 113-4.

VAN CAMPENHOUT, S., DEVRIENDT, K., BRECKPOT, J., FRIJNS, J. P., PEETERS, H., VAN BUGGENHOUT, G., VAN ESCH, H., MAES, B. & SWILLEN, A. 2012. Microduplication 22q11.2: a description of the clinical, developmental and behavioral characteristics during childhood. *Genet Couns,* 23**,** 135-48.

VILJAKAINEN, H., ANDERSSON-ASSARSSON, J. C., ARMENIO, M., PEKKINEN, M., PETTERSSON, M., VALTA, H., LIPSANEN-NYMAN, M., MAKITIE, O. & LINDSTRAND, A. 2015. Low

Copy Number of the AMY1 Locus Is Associated with Early-Onset Female Obesity in Finland. *PLoS One,* 10**,** e0131883.

WAIN, L. V., ARMOUR, J. A. & TOBIN, M. D. 2009. Genomic copy number variation, human health, and disease. *Lancet,* 374**,** 340-50.

WAIN, L. V., SHRINE, N., MILLER, S., JACKSON, V. E., NTALLA, I., SOLER ARTIGAS, M., BILLINGTON, C. K., KHEIRALLAH, A. K., ALLEN, R., COOK, J. P., PROBERT, K., OBEIDAT, M., BOSSE, Y., HAO, K., POSTMA, D. S., PARE, P. D., RAMASAMY, A., CONSORTIUM, U. K. B. E., MAGI, R., MIHAILOV, E., REINMAA, E., MELEN, E., O'CONNELL, J., FRANGOU, E., DELANEAU, O., OX, G. S. K. C., FREEMAN, C., PETKOVA, D., MCCARTHY, M., SAYERS, I., DELOUKAS, P., HUBBARD, R., PAVORD, I., HANSELL, A. L., THOMSON, N. C., ZEGGINI, E., MORRIS, A. P., MARCHINI, J., STRACHAN, D. P., TOBIN, M. D. & HALL, I. P. 2015. Novel insights into the genetics of smoking behaviour, lung function, and chronic obstructive pulmonary disease (UK BiLEVE): a genetic association study in UK Biobank. *Lancet Respir Med,* 3**,** 769-81.

WALKER, S., JANYAKHANTIKUL, S. & ARMOUR, J. A. 2009. Multiplex Paralogue Ratio Tests for accurate measurement of multiallelic CNVs. *Genomics,* 93**,** 98-103.

WALLEY, A. J., ASHER, J. E. & FROGUEL, P. 2009. The genetic contribution to non-syndromic human obesity. *Nat Rev Genet,* 10**,** 431-42.

WALTERS, R. G., JACQUEMONT, S., VALSESIA, A., DE SMITH, A. J., MARTINET, D., ANDERSSON, J., FALCHI, M., CHEN, F., ANDRIEUX, J., LOBBENS, S., DELOBEL, B., STUTZMANN, F., EL-SAYED MOUSTAFA, J. S., CHEVRE, J. C., LECOEUR, C., VATIN, V., BOUQUILLON, S., BUXTON, J. L., BOUTE, O., HOLDER-ESPINASSE, M., CUISSET, J. M., LEMAITRE, M. P., AMBRESIN, A. E., BRIOSCHI, A., GAILLARD, M., GIUSTI, V., FELLMANN, F., FERRARINI, A., HADJIKHANI, N., CAMPION, D., GUILMATRE, A., GOLDENBERG, A., CALMELS, N., MANDEL, J. L., LE CAIGNEC, C., DAVID, A., ISIDOR, B., CORDIER, M. P., DUPUIS-GIROD, S., LABALME, A., SANLAVILLE, D., BERI-DEXHEIMER, M., JONVEAUX, P., LEHEUP, B., OUNAP, K., BOCHUKOVA, E. G., HENNING, E., KEOGH, J., ELLIS, R. J., MACDERMOT, K. D., VAN HAELST, M. M., VINCENT-DELORME, C., PLESSIS, G., TOURAINE, R., PHILIPPE, A., MALAN, V., MATHIEU-DRAMARD, M., CHIESA, J., BLAUMEISER, B., KOOY, R. F., CAIAZZO, R., PIGEYRE, M., BALKAU, B., SLADEK, R., BERGMANN, S., MOOSER, V., WATERWORTH, D., REYMOND, A., VOLLENWEIDER, P., WAEBER, G., KURG, A., PALTA, P., ESKO, T., METSPALU, A., NELIS, M., ELLIOTT, P., HARTIKAINEN, A. L., MCCARTHY, M. I., PELTONEN, L., CARLSSON, L., JACOBSON, P., SJOSTROM, L., HUANG, N., HURLES, M. E., O'RAHILLY, S., FAROOQI, I. S., MANNIK, K., JARVELIN, M. R., PATTOU, F., MEYRE, D., WALLEY, A. J., COIN, L. J., BLAKEMORE, A. I., FROGUEL, P. & BECKMANN, J. S. 2010. A new highly penetrant form of obesity due to deletions on chromosome 16p11.2. *Nature,* 463**,** 671-5.

WANG, D., SUN, Y., STANG, P., BERLIN, J. A., WILCOX, M. A. & LI, Q. 2009. Comparison of methods for correcting population stratification in a genome-wide association study of rheumatoid arthritis: principal-component analysis versus multidimensional scaling. *BMC Proc,* 3 Suppl 7**,** S109.

WANG, K., LI, M., HADLEY, D., LIU, R., GLESSNER, J., GRANT, S. F., 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**,** 1665-74.

WANG, M. H., CORDELL, H. J. & VAN STEEN, K. 2018. Statistical methods for genome-wide association studies. *Semin Cancer Biol*.

WARDLE, J., CARNELL, S., HAWORTH, C. M., FAROOQI, I. S., O'RAHILLY, S. & PLOMIN, R. 2008a. Obesity associated genetic variation in FTO is associated with diminished satiety. *J Clin Endocrinol Metab,* 93**,** 3640-3.

WARDLE, J., CARNELL, S., HAWORTH, C. M. & PLOMIN, R. 2008b. Evidence for a strong genetic influence on childhood adiposity despite the force of the obesogenic environment. *Am J Clin Nutr,* 87**,** 398-404.

WATSON, J. D. & CRICK, F. H. 1953. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature,* 171**,** 737-8.

WEAVER, S., DUBE, S., MIR, A., QIN, J., SUN, G., RAMAKRISHNAN, R., JONES, R. C. & LIVAK, K. J. 2010. Taking qPCR to a higher level: Analysis of CNV reveals the power of high throughput qPCR to enhance quantitative resolution. *Methods,* 50**,** 271-6.

WEISS, L. A., SHEN, Y., KORN, J. M., ARKING, D. E., MILLER, D. T., FOSSDAL, R., SAEMUNDSEN, E., STEFANSSON, H., FERREIRA, M. A., GREEN, T., PLATT, O. S., RUDERFER, D. M., WALSH, C. A., ALTSHULER, D., CHAKRAVARTI, A., TANZI, R. E., STEFANSSON, K., SANTANGELO, S. L., GUSELLA, J. F., SKLAR, P., WU, B. L., DALY, M. J. & AUTISM, C. 2008. Association between microdeletion and microduplication at 16p11.2 and autism. *N Engl J Med,* 358**,** 667-75.

WELLCOME TRUST CASE CONTROL, C. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature,* 447**,** 661-78.

WELLING, M., AUVINEN, J., LEHENKARI, P., MANNIKKO, M., KARPPINEN, J. & ESKOLA, P. J. 2017. Association between height and osteoarthritis of the knee and hip: The Northern Finland Birth Cohort 1966 Study. *Int J Rheum Dis,* 20**,** 1095-1104.

WELSH, S. 2017. *Genotyping of 500,000 UK Biobank participants: Description of sample processing workflow and preparation of DNA for genotyping.* [Online]. Available: http://biobank.ndph.ox.ac.uk/showcase/refer.cgi?id=807 [Accessed].

WENGER, T. L., MILLER, J. S., DEPOLO, L. M., DE MARCHENA, A. B., CLEMENTS, C. C., EMANUEL, B. S., ZACKAI, E. H., MCDONALD-MCGINN, D. M. & SCHULTZ, R. T. 2016. 22q11.2 duplication syndrome: elevated rate of autism spectrum disorder and need for medical screening. *Mol Autism,* 7**,** 27.

WENTZEL, C., FERNSTROM, M., OHRNER, Y., ANNEREN, G. & THURESSON, A. C. 2008. Clinical variability of the 22q11.2 duplication syndrome. *Eur J Med Genet,* 51**,** 501-10.

WEST, J., DAHLQVIST LEINHARD, O., ROMU, T., COLLINS, R., GARRATT, S., BELL, J. D., BORGA, M. & THOMAS, L. 2016. Feasibility of MR-Based Body Composition Analysis in Large Scale Population Studies. *PLoS One,* 11**,** e0163332.

WHALE, A. S., HUGGETT, J. F., COWEN, S., SPEIRS, V., SHAW, J., ELLISON, S., FOY, C. A. & SCOTT, D. J. 2012. Comparison of microfluidic digital PCR and conventional quantitative PCR for measuring copy number variation. *Nucleic Acids Res,* 40**,** e82.

WHO, W. H. O. 2000. Obesity: preventing and managing the global epidemic. Report of a WHO consultation. *World Health Organ Tech Rep Ser,* 894**,** i-xii, 1-253.

WILLIAMS, N. M. 2011. Molecular mechanisms in 22q11 deletion syndrome. *Schizophr Bull,* 37**,** 882-9.

WILSON, C. L., LIU, W., YANG, J. J., KANG, G., OJHA, R. P., NEALE, G. A., SRIVASTAVA, D. K., GURNEY, J. G., HUDSON, M. M., ROBISON, L. L. & NESS, K. K. 2015. Genetic and clinical factors associated with obesity among adult survivors of childhood cancer: A report from the St. Jude Lifetime Cohort. *Cancer,* 121**,** 2262-70.

WINCHESTER, L., YAU, C. & RAGOUSSIS, J. 2009. Comparing CNV detection methods for SNP arrays. *Brief Funct Genomic Proteomic,* 8**,** 353-66.

WOLFF, G. L. 1987. Body weight and cancer. *Am J Clin Nutr,* 45**,** 168-80.
XU, H., POH, W. T., SIM, X., ONG, R. T., SUO, C., TAY, W. T., KHOR, C. C., SEIELSTAD, M., LIU, J., AUNG, T., TAI, E. S., WONG, T. Y., CHIA, K. S. & TEO, Y. Y. 2011. SgD-CNV, a database for common and rare copy number variants in three Asian populations. *Hum Mutat,* 32**,** 1341-9.
YANG, Q., RASMUSSEN, S. A. & FRIEDMAN, J. M. 2002. Mortality associated with Down's syndrome in the USA from 1983 to 1997: a population-based study. *Lancet,* 359**,** 1019-25.
YANG, Z. M., LIN, J., CHEN, L. H., ZHANG, M., CHEN, W. W. & YANG, X. R. 2015. The roles of AMY1 copies and protein expression in human salivary alpha-amylase activity. *Physiol Behav,* 138**,** 173-8.
YEO, G. S., CONNIE HUNG, C. C., ROCHFORD, J., KEOGH, J., GRAY, J., SIVARAMAKRISHNAN, S., O'RAHILLY, S. & FAROOQI, I. S. 2004. A de novo mutation affecting human TrkB associated with severe obesity and developmental delay. *Nat Neurosci,* 7**,** 1187-9.
YEO, G. S., FAROOQI, I. S., AMINIAN, S., HALSALL, D. J., STANHOPE, R. G. & O'RAHILLY, S. 1998. A frameshift mutation in MC4R associated with dominantly inherited human obesity. *Nat Genet,* 20**,** 111-2.
YONG, R. Y., MUSTAFFA, S. B., WASAN, P. S., SHENG, L., MARSHALL, C. R., SCHERER, S. W., TEO, Y. Y. & YAP, E. P. 2016. Complex Copy Number Variation of AMY1 does not Associate with Obesity in two East Asian Cohorts. *Hum Mutat,* 37**,** 669-78.
YU, Y., ZHU, H., MILLER, D. T., GUSELLA, J. F., PLATT, O. S., WU, B. L., SHEN, Y. & CHILDREN'S HOSPITAL BOSTON GENOTYPE PHENOTYPE STUDY, G. 2011. Age- and gender-dependent obesity in individuals with 16p11.2 deletion. *J Genet Genomics,* 38**,** 403-9.
ZHANG, F., CARVALHO, C. M. & LUPSKI, J. R. 2009a. Complex human chromosomal and genomic rearrangements. *Trends Genet,* 25**,** 298-307.
ZHANG, F. & DENG, H. W. 2010. Correcting for cryptic relatedness in population-based association studies of continuous traits. *Hum Hered,* 69**,** 28-33.
ZHANG, F., GU, W., HURLES, M. E. & LUPSKI, J. R. 2009b. Copy number variation in human health, disease, and evolution. *Annu Rev Genomics Hum Genet,* 10**,** 451-81.
ZHANG, F., KHAJAVI, M., CONNOLLY, A. M., TOWNE, C. F., BATISH, S. D. & LUPSKI, J. R. 2009c. The DNA replication FoSTeS/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans. *Nat Genet,* 41**,** 849-53.
ZHANG, X., DU, R., LI, S., ZHANG, F., JIN, L. & WANG, H. 2014. Evaluation of copy number variation detection for a SNP array platform. *BMC Bioinformatics,* 15**,** 50.
ZHANG, Y., HARAKSINGH, R., GRUBERT, F., ABYZOV, A., GERSTEIN, M., WEISSMAN, S. & URBAN, A. E. 2013a. Child development and structural variation in the human genome. *Child Dev,* 84**,** 34-48.
ZHANG, Y., KENT, J. W., JR., OLIVIER, M., ALI, O., CERJAK, D., BROECKEL, U., ABDOU, R. M., DYER, T. D., COMUZZIE, A., CURRAN, J. E., CARLESS, M. A., RAINWATER, D. L., GORING, H. H., BLANGERO, J. & KISSEBAH, A. H. 2013b. A comprehensive analysis of adiponectin QTLs using SNP association, SNP cis-effects on peripheral blood gene expression and gene expression correlation identified novel metabolic syndrome (MetS) genes with potential role in carcinogenesis and systemic inflammation. *BMC Med Genomics,* 6**,** 14.
ZHENG, Z., ZHANG, L., LI, S., ZHAO, F., WANG, Y., HUANG, L., HUANG, J., ZOU, R., QU, Y. & MU, D. 2017. Association among obesity, overweight and autism spectrum disorder: a systematic review and meta-analysis. *Sci Rep,* 7**,** 11697.

ZUFFEREY, F., SHERR, E. H., BECKMANN, N. D., HANSON, E., MAILLARD, A. M., HIPPOLYTE, L., MACE, A., FERRARI, C., KUTALIK, Z., ANDRIEUX, J., AYLWARD, E., BARKER, M., BERNIER, R., BOUQUILLON, S., CONUS, P., DELOBEL, B., FAUCETT, W. A., GOIN-KOCHEL, R. P., GRANT, E., HAREWOOD, L., HUNTER, J. V., LEBON, S., LEDBETTER, D. H., MARTIN, C. L., MANNIK, K., MARTINET, D., MUKHERJEE, P., RAMOCKI, M. B., SPENCE, S. J., STEINMAN, K. J., TJERNAGEL, J., SPIRO, J. E., REYMOND, A., BECKMANN, J. S., CHUNG, W. K., JACQUEMONT, S., SIMONS, V. I. P. C. & P11.2 EUROPEAN, C. 2012. A 600 kb deletion syndrome at 16p11.2 leads to energy imbalance and neuropsychiatric disorders. *J Med Genet,* 49**,** 660-8.

# Appendices