Large-scale genetic analysis of quantitative traits

Joshua C. Randall

Jesus College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2011

Abstract

Recent advances in genotyping technology coupled with an improved understanding of the architecture of linkage disequilibrium across the human genome have resulted in genome-wide association studies (GWAS) becoming a useful and widely applied tool for discovering common genetic variants associated with both quantitative traits and disease risk. After each GWAS was completed, it left behind a set of genotypes and phenotypes, often including anthropometric measures used as covariates. Genetic associations with anthropometric measures are not well characterized, perhaps due to lack of power to detect them in the sample sizes of individual studies. To improve power to detect variants associated with complex phenotypes such as anthropometric traits, data from multiple GWAS can be combined. This thesis describes the methods and results of several such analyses performed as part of the Genome-wide Investigation of ANThropemtric measures (GIANT) consortium, and compares various different methods that can be used to perform combined analyses of GWAS. In particular, the comparisons focus on comparing differences between meta-analysis methods, in which only summary statistics that result from within-study association testing are shared between studies, and mega-analysis methods in which individual-level genotype and phenotype data is analysed together. Finally, a brief discussion of technological means that have the potential to help overcome some of the challenges associated with performing mega-analyses is offered in order to suggest future work that could be undertaken in this area.

Acknowledgements

I am grateful for the guidance, encouragement, and support I have received from my supervisors, Cecilia Lindgren and Chris Holmes. Thanks in particular to Cecilia for always making sure I had plenty of interesting research to do.

Thanks to the Nuffield Department of Medicine (NDM) for funding my D.Phil studentship for the first two years, and to the Wellcome Trust, for funding the following years.

Thanks also to all the organisers of and participants in the GIANT consortium and its constituent studies, who provided the data for all of the analyses presented in Chapter 2, and especially to my collaborators in the several analysis working groups who contributed to the analyses.

A special thanks to Jeff Barrett, for first suggesting the possibility that I leave my position in telecommunications research & development to go into genetic research, for connecting me with the people who could make that happen, and for following through to support me along the way.

Finally, thanks to Eric Lander, who, as my Introductory Biology instructor at MIT, suggested that one needn't study biology in order to contribute to biological and genetic research, and who, through his inspiring lectures, led me to see genetics as a most interesting field of research.

Contents

Acronyms **ix**

Gene List **xvii**

1 Introduction **1**

1.1 Genes and genotypes 2

1.1.1 Linkage disequilibrium 3

1.1.2 Genotyping methods 5

1.1.3 Genotype calling algorithms 5

1.1.4 Called genotype data 7

1.1.5 Genotype uncertainty data 10

1.2 Genetic association studies 12

1.2.1 Quality control 14

1.2.1.1 Experimental replicates and basic genotype QC . . . 14

1.2.1.2 Departure from Hardy-Weinberg Equilibrium 14

1.2.2 Phenotype 16

1.2.2.1 Quantitative traits 16

1.2.2.2 Phenotype transformation 17

1.2.2.3 Treating quantitative traits as case-control phenotypes 17

1.2.3 Genotype imputation 19

1.2.3.1 Imputation accuracy 20

1.2.4 Genetic association analysis 22

1.2.4.1 Additive, dominant, and recessive models 23
1.2.5 Confounding factors and methods to control confounding . . . 23
1.2.5.1 Population stratification and cryptic relatedness . . . 24
1.2.5.2 Genomic control 25
1.2.5.3 Principal components analysis 26
1.2.5.4 Statistical methods for genetic association testing . . 28
1.2.5.5 Automation of association analysis 29
1.2.6 Interpretation and visualisation of data 30
1.2.6.1 Intensity plots 30
1.2.6.2 Plotting PCs with reference populations 31
1.2.6.3 Genome-wide association plots 34
1.2.6.4 Genome-wide quantile-quantile plots 36
1.3 Meta-analysis of genome-wide association studies 37
1.3.1 Statistical methods for meta-analysis 40
1.3.1.1 Fisher's combined probability test 40
1.3.1.2 Weighted Z-score 41
1.3.1.3 Inverse-variance 41
1.3.1.4 Testing the difference in effect between two analysis strata 42
1.3.2 Practical issues 43
1.3.2.1 Pruning into independent loci 44
1.3.3 Significance levels in GWAS meta-analysis 45
1.3.3.1 Bonferroni correction 45
1.3.3.2 Genome-wide significance 45
1.3.3.3 False-discovery rate control 46

2 Results of genome-wide association meta-analyses 47

2.1 Meta-analysis of waist traits, 2008-2009 50
2.1.1 Introduction 50

2.1.2 Results & Discussion 51
2.1.2.1 Stage 1 meta-analysis 51
2.1.2.2 In silico and de novo follow-up 55
2.1.2.3 Sex-specific association analyses 56
2.1.2.4 Confirmation in CHARGE consortium GWAS data . . 57
2.1.2.5 Disentangling effects on overall and central adiposity 59
2.1.2.6 Description of the three loci 60
2.1.2.7 Associations with other phenotypes 64
2.1.2.8 Conclusion 65
2.2 Meta-analysis of WHRadjBMI 66
2.2.1 Introduction 66
2.2.2 Methods 67
2.2.2.1 Contributing studies 67
2.2.2.2 Phenotype definition 67
2.2.2.3 Genotypes and association testing 68
2.2.2.4 Discovery stage GWAMA 69
2.2.2.5 Follow-up meta-analysis 70
2.2.2.6 Testing for sex-difference 70
2.2.2.7 Percentage of variance explained 71
2.2.2.8 Pathway analysis 71
2.2.2.9 Copy number variant analyses 72
2.2.2.10 eQTL analyses 73
2.2.3 Results 75
2.2.3.1 Genome-wide significance association of WHRadjBMI with 14 SNPs 75
2.2.3.2 Sexual dimorphism at several of the WHRadjBMI loci 78
2.2.3.3 Between-study heterogeneity 79
2.2.3.4 Association with other anthropometric measures . . 79
2.2.3.5 Enrichment of association with metabolic traits . . . 81

2.2.3.6 Pathway analysis and potential biological roles . . . 82
2.2.3.7 Evaluation of CNVs and non-synonymous changes . . 83
2.2.3.8 Effect of WHRadjBMI associations on expression in relevant tissues 84
2.2.3.9 RNA expression in gluteal and abdominal fat tissue . 85
2.2.3.10 Description of the fourteen loci 85
2.2.4 Discussion 101
2.3 Meta-analyses of body mass index 104
2.3.1 Introduction 104
2.3.2 Results 105
2.3.2.1 Stage 1 GWA studies identify novel loci associated with BMI 105
2.3.2.2 Stage 2 follow-up leads to additional novel loci for BMI 109
2.3.2.3 Impact of 32 confirmed loci on BMI, obesity, body size, and other metabolic traits 110
2.3.2.4 Potential functional roles and pathways analyses . . 111
2.3.2.5 Identifying possible functional variants 112
2.3.3 Discussion 113
2.4 Meta-analyses of weight 115
2.4.1 Background 115
2.5 Sex-specific meta-analysis of nine anthropometric traits 119
2.5.1 Results of sex-specific analyses 120
2.5.1.1 Discovery meta-analysis of sex-specific genome-wide association study for anthropometric traits 120
2.5.2 Descriptions of the 18 sex-specific loci 140
2.5.2.1 WHRadjBMI loci 142
2.5.2.2 WCadjBMI loci 152
2.5.2.3 Height loci 153
2.5.2.4 Weight loci 162

2.5.3 Sex-specific analysis methods . . . 165
2.5.3.1 Anthropometric phenotypes . . . 165
2.5.3.2 Discovery GWAS . . . 166
2.5.3.3 Follow-up studies . . . 166
2.5.3.4 Analyses and quality control on the study level . . . 167
2.5.3.5 Genome-wide sex-difference analysis . . . 167
2.5.3.6 Sex-specific multi-stage analysis . . . 168
2.5.3.7 Follow-up and joint meta-analyses . . . 169
2.5.3.8 Age-stratified sex-specific meta-analysis . . . 171
2.5.3.9 Association with other phenotypes . . . 171
2.5.4 Discussion . . . 172
2.5.4.1 Summary of findings . . . 172
2.5.5 Conclusion . . . 172

3 Evaluation of summary-statistics meta-analyses compared to individual-level analyses 174

3.1 Meta/mega-analysis in homogeneous populations . . . 174
3.2 Meta/mega-analyses in the presence of within-study population stratification . . . 180
3.3 Meta/mega-analyses in the presence of between-study population stratification . . . 190

4 The future of large-scale genetic analyses 194

4.1 Mega-analysis of individual-level data . . . 194
4.1.1 Covariates . . . 195
4.1.2 Privacy & informed consent . . . 197
4.1.2.1 Data protection . . . 197
4.1.2.2 Informed consent . . . 198
4.1.2.3 Publishing of analysis results . . . 198
4.1.3 Trusted analysis platform . . . 199

4.1.3.1 Platform architecture 200
4.1.3.2 Data 201
4.1.3.3 Analysis pipelines and modules 201
4.1.3.4 Informed consent within a trusted platform 204
4.1.3.5 Cost of a trusted analysis platform 204

Bibliography **207**

Acronyms

1000G 1000 Genomes Project. 4, 12, 74, 84, 111, 145

ADAMTS a disintegrin and metalloproteinase with thrombospondin motif. 93, 146

ANOVA analysis of variance. 28, 74

ASW HapMap population comprising samples of African ancestry in Southwest USA. 32, 34

BED Binary PED for PLINK. 7, 8

β effect estimate. 38, 39, 41–43, 53, 58, 68, 69, 71, 77, 80, 107, 176, 177, 181–185, 199

BLM bleomycin. 156

BMI body mass index. vi, xii, xv, xvi, 17, 48, 50, 54, 55, 59, 60, 66–68, 70, 78–80, 86, 101–105, 109–117, 119, 120, 128, 129, 131, 132, 138, 140, 148, 157, 164–166, 172

CAD Coronary Artery Disease. 66, 115, 150

CD Celiac Disease. 152

CED concordant effect direction. 119, 120, 125, 172

CEPH Centre d'Etude du Polymorphisme Humain. ix, 166

CEU HapMap population comprising samples from the CEPH collection (Utah residents with ancestry from northern and western Europe). 12, 19, 32–34, 51, 68, 71, 72, 74, 115, 166, 175, 181, 190

CHARGE Cohorts for Heart and Aging Research in Genomic Epidemiology. v, 56–58

CHB HapMap population comprising samples of Han Chinese in Beijing, China. 12, 32, 34

CHD HapMap population comprising samples of Chinese in Metropolitan Denver, Colorado. 32, 34

CNV copy number variant. v, vi, x, 2, 72, 73, 83, 84, 95, 111, 148

CTS CNV tag SNP. 72, 73

D-BP D-bifunctional protein. 147

deCODE deCODE genetics, Reykjavik, Iceland. 74, 118

DF degree of freedom. 17, 22, 23, 40, 42

DIAGRAM DIAbetes Genetics Replication And Meta-analysis. 64

DNA deoxyribonucleic acid. x, 1, 5, 98, 157, 195, 196

DNMT3B DNA (cytosine-5-)-methyltransferase 3 beta. 98

$|D'|$ D-prime. 3

DXA dual energy X-ray absorptiometry. 59, 64, 196

EA effect allele. 58, 77, 107

EAF effect allele frequency. 58, 68, 71, 77, 107

eQTL expression QTL. v, 73, 74, 84, 102, 112, 142

ER endoplasmic reticulum. 63, 164

EST expressed sequence tag. 84

FDR false-discovery rate. 45, 46, 85, 127, 129, 168, 169

FFA free fatty acids. 63

FI fasting insulin. 81, 82, 171

GC genomic control. 24, 25, 43, 53, 69, 70, 74, 75, 115, 168, 169, 180–185, 187, 191

GCC genomic control correction. 181–185, 187, 188, 190, 191

GENCODE The GENCODE Project: Encyclopædia of genes and gene variants. 90

GIANT Genome-wide Investigation of ANThropemtric measures. 48–50, 66, 79, 104, 115, 118–120, 138, 171

GIH HapMap population comprising samples of Gujarti Indians in Houston, Texas. 32–34, 181, 190

GLGC Global Lipids Genetic Consortium. 171

GO gene ontology. 111

GPI glycosylphosphatidylinositol. 87

GSV Genomic Structural Variation. 72

GWA genome-wide association. vi, xi, 19, 34, 44, 66, 74, 102, 105, 109, 114, 116, 199

GWAMA GWA meta-analysis. v, 19, 38, 45, 66, 69, 83, 102, 104, 157, 166

GWAS genome-wide association study. iv–vii, 1, 12, 19, 21–23, 25, 27, 37, 45, 47, 48, 51, 55, 56, 59, 64, 67, 69, 79, 81, 82, 104, 113, 119, 120, 134, 166, 172, 174

GWS genome-wide significance. v, 45, 75, 78, 101, 105, 109, 113, 185

H_0 the null hypothesis. 15, 24, 36, 40, 53, 54, 75, 105, 109

HapMap International HapMap Project. ix–xiii, xv, xvi, 2–4, 12, 14, 19, 26, 31–34, 38, 39, 44, 51, 60, 62, 68, 69, 71, 72, 74, 84, 85, 111, 115, 140, 166, 168, 175, 181, 185, 190

HC hip circumference. xii, 60, 67–69, 80, 81, 102, 120, 128, 129, 131, 132, 138, 165, 166

HCadjBMI HC-adjusted-for-BMI. 120, 128, 129, 131, 132, 138, 165, 166

HCC hepatocellular carcinoma. 159

HDL-C high-density lipoprotein cholesterol. 90, 171

HMM hidden markov model. 19

HOMA homeostasis model assessment. 81

HWE Hardy-Weinberg equilibrium. 14, 15, 20, 21

IBD inflammatory bowel disease. 1

IC informed consent. 197, 198

INDEL insertion-deletion polymorphism. 113

IR insulin resistance. 82

IV inverse-variance. 43, 53, 177, 178, 180, 183, 188, 190, 191

JPT HapMap population comprising samples of Japanese in Tokyo, Japan. 12, 32, 34

KEGG Kyoto encyclopedia of genes and genomes. 111

LD linkage disequilibrium. 2–4, 12, 19, 25, 36, 44, 45, 60, 68, 72, 73, 78, 83–85, 91, 109, 111, 140, 142, 150, 157, 175, 181

LDL low-density lipoprotein. 81

MAC minor allele count. 40, 69, 167

MAF minor allele frequency. 4, 12, 21, 28, 115, 167, 175, 181, 190, 194, 195

MAGENTA Meta-Analysis Gene-set Enrichment of variaNT Associations. 111

MAGIC Meta-Analyses of Glucose and Insulin-related traits Consortium. 171

MC MetaboChip. 134

MCOPCT2 microphthalmia with cataract type 2. 154

μ mean. 181

MGH the Massachusetts General Hospital in Boston, Massachusetts, USA. 74

MHC major histocompatibility complex. 71

μRNA microRNA. 93, 146, 151, 159

MXL HapMap population comprising samples of Mexican ancestry in Los Angeles, California. 32–34, 181, 190

N sample size. 38, 58, 68, 70, 74, 79, 84, 107, 110

NCBI National Center for Biotechnology Information, NLM. xiii, xiv, 60

NCBI36 NCBI genome build 36. 60, 85, 140

NHGRI National Human Genome Research Institute. 47

NLM US National Library of Medicine. xiii

OED opposite effect direction. 119, 120, 125, 172

PANTHER Protein ANalysis THrough Evolutionary Relationships. 71, 83, 111

PC principal component. iv, 26, 27, 31–34, 48, 184, 185, 188, 190, 191

PCA principal components analysis. 15, 26, 27, 31, 32, 48, 180, 184, 185, 187, 188, 190, 191, 197

PDGF platelet-derived growth factor. 111

PED Pedigree file for PLINK. ix, 7, 8

PIN personal identification number. 199

QC quality control. iii, 14–16, 19, 30, 38, 48, 69, 70, 115, 167, 168

QQ quantile-quantile. 36, 43, 44, 54, 75, 105, 127, 129, 132

QTL quantiative trait locus. x, 2, 73

QTN quantitative trait nucleotide. 2

r^2 r-squared. 3

$\hat{r^2}$ r-squared hat. 20, 21

RA Rheumatoid Arthritis. 152

RefSeq NCBI reference sequence database. 60, 85, 140

RFLP restriction fragment length polymorphism. 2

RNA ribonucleic acid. vi, xiii, 74, 85, 93, 94, 146, 151

RP105 radioprotective, 105 kDa. 95

SA structured association. 27

SAT subcutaneous adipose tissue. 74, 84, 85, 112

SD standard deviation. 181, 188, 191

SE standard error. 38, 39, 41–43, 53, 58, 68, 70, 77, 107, 115, 167, 176, 177, 181–185, 199

SLC30 zinc efflux family. 89

SMS Smith-Magenis syndrome. 150, 151

SNP single-nucleotide polymorphism. v, x, 2–5, 7, 8, 10, 12, 14, 15, 17, 19–22, 25, 27, 30, 32, 34, 38, 40, 45, 47, 48, 51, 53–58, 60, 62, 64, 67–75, 77–80, 82–85, 90, 91, 93, 102, 105, 107, 109–112, 118–120, 127, 129, 132, 138–140, 142, 144–146, 148, 150, 152, 154, 156, 157, 159, 161, 162, 164, 166–170, 175, 180–185, 190, 191, 194

T1D Type 1 Diabetes. 1

T2D Type 2 Diabetes. 1, 48, 50, 64, 66, 82, 89, 90, 93, 96, 102, 115, 145, 146, 148, 166, 199

TFBS transcription factor binding site. 164

TG triglyceride level. 81, 82, 102, 152, 171

TLR toll-like receptor. 95

TSI HapMap population comprising samples of Toscani in Italia. 32–34, 181, 190

UCSC University of California Santa Cruz. 60, 85, 140

US United States. xiii

UTR untranslated region. 91

s^2 variance. 71, 107

VAT visceral adipose tissue. 112

WC waist circumference. xv, 50, 51, 53–60, 62, 63, 66–69, 80, 81, 102, 105, 109, 120, 128, 129, 131, 132, 138, 165, 166

WCadjBMI WC-adjusted-for-BMI. vi, 120, 127–129, 131, 132, 138, 140, 152, 165, 166, 171, 172

WHR $\frac{waist\ circumference}{hip\ circumference} ratio$. xvi, 50, 51, 53–56, 58, 60, 61, 64, 66–68, 70, 73, 86, 101, 102, 119, 120, 127–129, 131, 132, 138, 143, 165, 166

WHRadjBMI WHR-adjusted-for-BMI. v, vi, 66–71, 73–75, 78–87, 89–91, 93–102, 120, 127–129, 131, 132, 138, 140, 142, 144–148, 150, 165, 166, 171, 172

WTCCC Wellcome Trust Case-Control Consortium. 45, 73

WZ weighted Z-score. 40, 41, 43, 53, 176, 178, 180, 182, 188, 190, 191

YRI HapMap population comprising samples of Yoruba in Ibadan, Nigeria. 12, 32, 34

Gene List

ABCB9 ATP-binding cassette, sub-family B (MDR/TAP), member 9. 162

ABI3 ABI family, member 3. 157

ADAMTS9 ADAM metallopeptidase with thrombospondin type 1 motif, 9. 83, 93, 102, 146

ADAMTS9-AS2 ADAMTS9 antisense RNA 2 (non-protein coding). 93, 146

ADCY3 adenylate cyclase 3. 111, 112

ALPP alkaline phosphatase, placental. 160

ALPPL2 alkaline phosphatase, placental-like 2. 160

ANKRD55 ankyrin repeat domain 55. 152

APOBR apolipoprotein B receptor. 111, 112

ARL6IP4 ADP-ribosylation-like factor 6 interacting protein 4. 162

ATP5G1 ATP synthase, H+ transporting, mitochondrial Fo complex, subunit C1 (subunit 9). 157

B3GNT4 UDP-GlcNAc:betaGal beta-1,3-N-acetylglucosaminyltransferase 4. 162

B4GALNT2 beta-1,4-N-acetyl-galactosaminyl transferase 2. 157

BAP1 BRCA1 associated protein-1 (ubiquitin carboxy-terminal hydrolase). 91

BCL7A B-cell CLL/lymphoma 7A. 162

BCL7B B-cell CLL/lymphoma 7B. 162

BCL7C B-cell CLL/lymphoma 7C. 162

BDNF brain-derived neurotrophic factor. 104, 110, 111

BRCA2 breast cancer 2, early onset. 152

C12orf65 chromosome 12 open reading frame 65. 162

C14orf39 chromosome 14 open reading frame 39. 154

C1orf105 chromosome 1 open reading frame 105. 87

C2orf57 chromosome 2 open reading frame 57. 160

C3orf25 chromosome 3 open reading frame 25. 153

C3orf37 chromosome 3 open reading frame 37. 153

C3orf78 chromosome 3 open reading frame 78. 91

C4orf14 chromosome 4 open reading frame 14. 156

CALCOCO2 calcium binding and coiled-coil domain 2. 157

CBX1 chromobox homolog 1. 157

CCDC62 coiled-coil domain containing 62. 162

CDK2AP1 cyclin-dependent kinase 2 associated protein 1. 162

CDK5RAP3 CDK5 regulatory subunit associated protein 3. 157

CLIP1 CAP-GLY domain containing linker protein 1. 162

COBLL1 COBL-like 1. 90, 142, 143

COPG coatomer protein complex, subunit gamma. 153

COPS7B COP9 constitutive photomorphogenic homolog subunit 7B (Arabidopsis). 160

COPZ2 coatomer protein complex, subunit zeta 2. 157

CPEB4 cytoplasmic polyadenylation element binding protein 4. 94

DENR density-regulated protein. 162

DIABLO diablo, IAP-binding mitochondrial protein. 162

DIS3L2 DIS3 mitotic control homolog (S. cerevisiae)-like 2. 160

DMXL1 Dmx-like 1. 147

DNAH1 dynein, axonemal, heavy chain 1. 84, 91

DNM3 dynamin 3. 87

DTWD2 DTW domain containing 2. 147

ECEL1P2 endothelin converting enzyme-like 1, pseudogene 2. 160

ETV5 ets variant 5. 104

FAM117A family with sequence similarity 117, member A. 157

FAM170A family with sequence similarity 170, member A. 147

FBLL1 fibrillarin-like 1. 161

FILIP1 filamin A interacting protein 1. 155

FTO fat mass and obesity associated. 54, 79, 104, 110, 116

GDF5 growth differentiation factor 5. 117

GIP gastric inhibitory polypeptide. 157

GIPR gastric inhibitory polypeptide receptor. 111

GLT8D1 glycosyltransferase 8 domain containing 1. 91

GLYCTK glycerate kinase. 84, 91

GNGT2 guanine nucleotide binding protein (G protein), gamma transducing activity polypeptide 2. 157

GNL3 guanine nucleotide binding protein-like 3 (nucleolar). 91

GNPDA2 glucosamine-6-phosphate deaminase 2. 104

GPC5 glypican 5. 159

GPRC5B G protein-coupled receptor, family C, group 5, member B. 111, 112

GRB14 growth factor receptor-bound protein 14. 83, 84, 90, 102, 142, 143

Grb14 Mus musculus growth factor receptor bound protein 14. 90, 102, 143

GTF3A general transcription factor IIIA. 112

H1FOO H1 histone family, member O, oocyte-specific. 153

H1FX H1 histone family, member X. 153

H1FX-AS1 H1FX antisense RNA 1 (non-protein coding). 153

HCAR1 hydroxycarboxylic acid receptor 1. 162

HCAR2 hydroxycarboxylic acid receptor 2. 162

HCAR3 hydroxycarboxylic acid receptor 3. 162

HHIP hedgehog interacting protein. 117

HIP1R huntingtin interacting protein 1 related. 162

HMGA2 high mobility group AT-hook 2. 56, 116

HMGCR 3-hydroxy-3-methylglutaryl-CoA reductase. 111

HOXB1 homeobox B1. 157

HOXB13 homeobox B13. 157

HOXB13-AS1 HOXB13 antisense RNA 1 (non-protein coding). 157

HOXB2 homeobox B2. 157

HOXB3 homeobox B3. 157

HOXB4 homeobox B4. 157

HOXB5 homeobox B5. 157

HOXB6 homeobox B6. 157

HOXB7 homeobox B7. 157

HOXB8 homeobox B8. 157

HOXB9 homeobox B9. 157

HOXC10 homeobox C10. 82

HOXC11 homeobox C11. 82

HOXC12 homeobox C12. 100

HOXC13 homeobox C13. 83, 100

HOXC4 homeobox C4. 82

HOXC6 homeobox C6. 82

HPD 4-hydroxyphenylpyruvate dioxygenase. 162

HSD17B4 hydroxysteroid (17-beta) dehydrogenase 4. 147

IFT122 intraflagellar transport 122 homolog (Chlamydomonas). 153

IGF2BP1 insulin-like growth factor 2 mRNA binding protein 1. 157

IL31 interleukin 31. 162

IQCK IQ motif containing K. 112

ITIH1 inter-alpha (globulin) inhibitor H1. 91

ITIH3 inter-alpha (globulin) inhibitor H3. 91

ITIH4 inter-alpha (globulin) inhibitor H4 (plasma Kallikrein-sensitive glycoprotein). 91

ITPR2 inositol 1,4,5-trisphosphate receptor, type 2. 83, 85, 99

Itpr2 Mus musculus inositol 1,4,5-triphosphate receptor 2. 99

Itpr3 Mus musculus inositol 1,4,5-triphosphate receptor 3. 99

KAT7 K(lysine) acetyltransferase 7. 157

KCTD15 potassium channel tetramerisation domain containing 15. 104

KDM2B lysine (K)-specific demethylase 2B. 162

KNTC1 kinetochore associated 1. 162

KREMEN1 kringle containing transmembrane protein 1. 82, 83

LINC00471 long intergenic non-protein coding RNA 471. 160

LRRC43 leucine rich repeat containing 43. 162

LRRC46 leucine rich repeat containing 46. 157

LY86 lymphocyte antigen 86. 83, 95

LYPLAL1 lysophospholipase-like 1. 61, 78, 83, 89, 144

MAP3K1 mitogen-activated protein kinase kinase kinase 1. 152

MBD4 methyl-CpG binding domain protein 4. 153

MC4R melanocortin 4 receptor. 54, 104, 110, 116

MED9 mediator complex subunit 9. 150, 151

MIR103A1 microRNA 103a-1. 161

MIR103B1 microRNA 103b-1. 161

MIR10A microRNA 10a. 157

MIR1203 microRNA 1203. 157

MIR1244-1 microRNA 1244-1. 147, 160

MIR1244-2 microRNA 1244-2. 147, 160

MIR1244-3 microRNA 1244-3. 147, 160

MIR135A1 microRNA 135a-1. 91

MIR1471 microRNA 1471. 160

MIR148A microRNA 148a. 98

MIR152 microRNA 152. 157

MIR17 microRNA 17. 159

MIR17HG miR-17-92 cluster host gene (non-protein coding). 159

MIR18A microRNA 18a. 159

MIR196A1 microRNA 196a-1. 157

MIR19A microRNA 19a. 159

MIR19B1 microRNA 19b-1. 159

MIR20A microRNA 20a. 159

MIR218-2 microRNA 218-2. 161

MIR3185 microRNA 3185. 157

MIR33B microRNA 33b. 150, 151

MIR4304 microRNA 4304. 162

MIR548AN microRNA 548an. 93, 146

MIR92A1 microRNA 92a-1. 159

MIRLET7G microRNA let-7g. 91

MLXIP MLX interacting protein. 162

MORN3 MORN repeat containing 3. 162

MPHOSPH9 M-phase phosphoprotein 9. 162

MRPL10 mitochondrial ribosomal protein L10. 157

MSRA methionine sulfoxide reductase A. 60, 63

MTCH2 mitochondrial carrier 2. 104, 111, 112

MUSTN1 musculoskeletal, embryonic nuclear protein 1. 91

MYO6 myosin VI. 155

NCL nucleolin. 160

NDUFS3 NADH dehydrogenase (ubiquinone) Fe-S protein 3, 30kDa (NADH-coenzyme Q reductase). 112

NEGR1 neuronal growth regulator 1. 104, 111, 112

NEK4 NIMA (never in mitosis gene a)-related kinase 4. 91

NFE2L1 nuclear factor (erythroid-derived 2)-like 1. 157

NFE2L3 nuclear factor (erythroid-derived 2)-like 3. 98

Nfe2l3 Mus musculus nuclear factor, erythroid derived 2, like 3. 98

NGFR nerve growth factor receptor. 157

NISCH nischarin. 83, 91

NMUR1 neuromedin U receptor 1. 160

NPC1 Niemann-Pick disease, type C1. 109

NPPC natriuretic peptide C. 160

NT5DC2 5'-nucleotidase domain containing 2. 91

NXPH3 neurexophilin 3. 157

OGFOD2 2-oxoglutarate and iron-dependent oxygenase domain containing 2. 162

ORAI1 ORAI calcium release-activated calcium modulator 1. 162

OSBPL7 oxysterol binding protein-like 7. 157

PANK3 pantothenate kinase 3. 161

PBRM1 polybromo 1. 91

PDE6D phosphodiesterase 6D, cGMP-specific, rod, delta. 160

PDGFRL platelet-derived growth factor receptor-like. 56

PEMT phosphatidylethanolamine N-methyltransferase. 150, 151

PHB prohibitin. 157

PHF7 PHD finger protein 7. 91

PHOSPHO1 phosphatase, orphan 1. 157

PIGC phosphatidylinositol glycan anchor biosynthesis, class C. 83, 84, 87

PITPNM2 phosphatidylinositol transfer protein, membrane-associated 2. 162

PLXND1 plexin D1. 82, 153

PNPO pyridoxamine 5'-phosphate oxidase. 157

POC5 POC5 centriolar protein homolog (Chlamydomonas). 111

PPARG peroxisome proliferator-activated receptor gamma. 148

PPM1A protein phosphatase, Mg2+/Mn2+ dependent, 1A. 154

PPM1M protein phosphatase, Mg2+/Mn2+ dependent, 1M. 91

PRAC prostate cancer susceptibility candidate. 157

PRR15L proline rich 15-like. 157

PRR16 proline rich 16. 147

PSMD9 proteasome (prosome, macropain) 26S subunit, non-ATPase, 9. 162

PTMA prothymosin, alpha. 160

QPCTL glutaminyl-peptide cyclotransferase-like. 111

RAB32 RAB32, member RAS oncogene family. 117

RAI1 retinoic acid induced 1. 150, 151

RARS arginyl-tRNA synthetase. 161

RASAL2 RAS protein activator like 2. 164

RASD1 RAS, dexamethasone-induced 1. 150

RASGEF1B RasGEF domain family, member 1B. 117

REST RE1-silencing transcription factor. 156

RFT1 RFT1 homolog (S. cerevisiae). 91

RHO rhodopsin. 153

RHOF ras homolog gene family, member F (in filopodia). 162

RILPL2 Rab interacting lysosomal protein-like 2. 162

RPL32P3 ribosomal protein L32 pseudogene 3. 153

RSPO3 R-spondin 3. 82, 83, 85, 97

Rspo3 Mus musculus R-spondin 3 homolog. 97

RSRC2 arginine/serine-rich coiled-coil 2. 162

SBNO1 strawberry notch homolog 1 (Drosophila). 162

SCRN2 secernin 2. 157

SEC16B SEC16 homolog B (S. cerevisiae). 164

SEMA3G sema domain, immunoglobulin domain (Ig), short basic domain, secreted, (semaphorin) 3G. 91

SENP6 SUMO1/sentrin specific peptidase 6. 155

SETD1B SET domain containing 1B. 162

SETD8 SET domain containing (lysine methyltransferase) 8. 162

SFMBT1 Scm-like with four mbt domains 1. 91

SH2B1 SH2B adaptor protein 1. 104, 110–112

SIX6 SIX homeobox 6. 154

SKAP1 src kinase associated phosphoprotein 1. 157

SKOR1 SKI family transcriptional corepressor 1. 111

SLC23A2 solute carrier family 23 (nucleobase transporters), member 2. 117

SLC30A10 solute carrier family 30, member 10. 89, 144

SLC30A8 solute carrier family 30 (zinc transporter), member 8. 89

SLC35B1 solute carrier family 35, member B1. 157

SLC39A8 solute carrier family 39 (zinc transporter), member 8. 111, 112

SLIT3 slit homolog 3 (Drosophila). 161

SMCR5 Smith-Magenis syndrome chromosome region, candidate 5 (non-protein coding). 150

SNF8 SNF8, ESCRT-II complex subunit, homolog (S. cerevisiae). 157

SNORA70F small nucleolar RNA, H/ACA box 70F (retrotransposed). 142

SNORA7B small nucleolar RNA, H/ACA box 7B. 153

SNORD19 small nucleolar RNA, C/D box 19. 91

SNORD19B small nucleolar RNA, C/D box 19B. 91

SNORD69 small nucleolar RNA, C/D box 69. 91

SNORD82 small nucleolar RNA, C/D box 82. 160

SNX11 sorting nexin 11. 157

SP2 Sp2 transcription factor. 157

SP6 Sp6 transcription factor. 157

SPCS1 signal peptidase complex subunit 1 homolog (S. cerevisiae). 91

SPOP speckle-type POZ protein. 157

SREBF1 sterol regulatory element binding transcription factor 1. 150, 151

STAB1 stabilin 1. 83–85, 91

SULT1A1 sulfotransferase family, cytosolic, 1A, phenol-preferring, member 1. 112, 117

SULT1A2 sulfotransferase family, cytosolic, 1A, phenol-preferring, member 2. 111, 112

TAC4 tachykinin 4 (hemokinin). 157

TBX15 T-box 15. 82–86, 102

Tbx15 Mus musculus T-box 15. 86

TFAP2B transcription factor AP-2 beta (activating enhancer binding protein 2 beta). 59, 62

TMCC1 transmembrane and coiled-coil domain family 1. 153

TMEM110 transmembrane protein 110. 91

TMEM110-MUSTN1 TMEM110-MUSTN1 readthrough. 91

TMEM120B transmembrane protein 120B. 162

TMEM160 transmembrane protein 160. 112

TMEM18 transmembrane protein 18. 79, 104, 116

TNFAIP8 tumor necrosis factor, alpha-induced protein 8. 147

TNNC1 troponin C type 1 (slow). 91

TOM1L2 target of myb1-like 2 (chicken). 150, 151

TTLL6 tubulin tyrosine ligase-like family, member 6. 157

TUFM Tu translation elongation factor, mitochondrial. 112

TWF2 twinfilin, actin-binding protein, homolog 2 (Drosophila). 91

UBE2Z ubiquitin-conjugating enzyme E2Z. 157

VEGFA vascular endothelial growth factor A. 83, 96, 102, 145

Vegfa Mus musculus vascular endothelial growth factor A. 97

VPS33A vacuolar protein sorting 33 homolog A (S. cerevisiae). 162

VPS37B vacuolar protein sorting 37 homolog B (S. cerevisiae). 162

WARS2 tryptophanyl tRNA synthetase 2, mitochondrial. 85

WDR66 WD repeat domain 66. 162

WDR82 WD repeat domain 82. 91

ZBTB38 zinc finger and BTB domain containing 38. 117

ZC3H11B zinc finger CCCH-type containing 11B pseudogene. 61, 89, 144

ZC3H4 zinc finger CCCH-type containing 4. 112

ZCCHC8 zinc finger, CCHC domain containing 8. 162

ZNF652 zinc finger protein 652. 157

ZNRF3 zinc and ring finger 3. 84, 101

Chapter 1

Introduction

In the past few years, genome-wide association studies (GWAS) have been performed on over 80 diseases and traits and have resulted in identifying hundreds of common genetic variants that are significantly associated with a variety of diseases and traits[1]. In addition to a large increase in the number of implicated loci for many diseases, including Type 1 Diabetes (T1D), Type 2 Diabetes (T2D), inflammatory bowel disease (IBD), prostate cancer, and breast cancer[2]; some common variants have also been identified that influence quantitative traits such as lipid levels, height, fat mass, and fat distribution[2–5]. Still, a large proportion of heritability in the vast majority of traits remains unexplained[2, 6–8]. For example, the known common variants for obesity explain only $\approx 1\%$ of the genetic variation expected to be present in the population based on heritability estimates[9]. The lack of an understanding of the unexplained variance is one motivation supporting continued research into the genetics of complex quantitative traits, the primary goal of which is to identify regions of the genome that are significantly associated with a particular trait in order to move towards identifying causal genetic variants and/or to implicate genes and elucidate pathways that are involved in trait regulation, with the ultimate goal of improving the understanding of the underlying biology behind each trait.

1.1 Genes and genotypes

A gene is a discrete unit of inheritance, while a genotype describes which gene variants (alleles) an individual has for a given trait or molecular phenotype[10]. In the early days of genetics, an individual's genotype was not directly measured but rather inferred from the phenotypes of a pedigree of related individuals, typically as the result of explicit hybridisation experiments in plants and animals[11]. Later, the discovery of deoxyribonucleic acid (DNA) as the material carrier of genetic information[12, 13], followed by the elucidation of the structure of DNA as a double-stranded helix formed by two nucleotide strands with complementary base pairs[14, 15], has led to the field of molecular genetics.

In studying the genetics of quantitative traits, we are ultimately searching for the causal molecular variants that effect a continuously variable trait through some underlying biological process. Single-nucleotide molecular variants that causally effect a quantitative trait are referred to as quantitative trait nucleotides (QTNs)[16], although other variation such as insertions, deletions, or copy number variants (CNVs) could also be causal molecular variants.

An effort to identify the causal genes for a quantitative trait typically begins with a study of a population of individuals across which both the phenotype and genotypes vary, in an effort to locate quantitative trait loci (QTLs) within regions of the genome (known as QTL mapping)[17]. Phenotype and genotype data are collected and analyzed, resulting in some genetic loci (regions of the genome) being more likely to harbor a causal gene (a QTN or other causal molecular variant) within them than others, and these regions are considered QTLs[16, 17]. The average size of a quantitative trait locus (QTL) implicated by a particular method can be thought of in terms of the resolution of that method to detect causal genes – methods that can find significantly associated QTLs of smaller size are of higher resolution.

While QTL mapping was first made possible by the completion of restriction fragment length polymorphism (RFLP) linkage maps[18], early work was somewhat limited (at least in humans) by the necessity of family-based study designs and the relatively low resolution of RFLP mapping[16]. However, developments in inexpensive array-based single-nucleotide polymorphism (SNP) genotyping[19] paired with the revelation that the extensive patterns of linkage disequilibrium (LD) found in human populations can be used to develop SNP genotyping panels that provide "genome-wide" coverage of most common variants with only a few hundred thousand marker SNPs[20, 21]. For example, arrays of 500,000 SNPs can cover up to 88% of all International HapMap Project (HapMap) phase 2[20]SNPs with $r^2 > 0.8$[20].

1.1.1 Linkage disequilibrium

LD is the non-random association between alleles at two (or more) positions in the genome within a population[21–23]. Under no selective pressure and without mutations, alleles tend towards a state of linkage equilibrium after a number of generations depending on initial allele frequencies[22], but under selective pressure or due to recent mutations alleles can be in a state of disequilibrium (and therefore also correlation). From this correlation between alleles some haplotypes can arise that are much more common than others[21, 23]. However, prior to the HapMap[20, 24] it was not clear to what extent LD in humans would segment the genome into haplotypes, or how common those haplotypes might be.

The HapMap project[20, 24] characterized the patterns of LD in several representative human populations, and found that there are substantial differences in recombination rates across the genome, such that most recombination occurs in localized hotspots giving rise to LD patterns in humans that appear as haplotype blocks[24] within which genetic variants (such as SNPs) can to some extent be used as proxies (tagging markers) for each other. Measures of pairwise LD (between two variants) can be used to evaluate the usefulness of a potential tag SNP. Two different measures are

commonly used, $|D'|$ and r^2. The r-squared (r^2) measure is the simply square of the correlation between the two SNPs, and $r^2 = 1$ when two variants arose on the same haplotypic background and have not since been disrupted by recombination events and will be less than 1 if the variants initially arose on different haplotypes or if they arose on the same haplotype but have since been disrupted by recombination. The D-prime ($|D'|$) measure is similar to r^2 in that ranges from 0–1 and $|D'| = 1$ when there has been no recombination (or recurrent mutation) between the two variants[24]. However, $|D'|$ is adjusted such that it is independent of allele frequency differences between the two variants and, and declines exponentially towards 0 in the presence of recombination events[24, 25]. While $|D'|$ is the more direct measure of LD, the independence of $|D'|$ from allele frequency differences means that $|D'|$ can be high when one of the variants is rare, resulting in r^2 being the measure more typically used when selecting tagging markers[21], perhaps because it can be useful in genetic association studies for a tag SNP to have a similar allele frequency to the SNP it is tagging in order to maintain power to detect effects of similar size.

These developments in understanding of LD in the human genome has made it possible for high-density SNP genotyping arrays to be developed that represent the majority of all common variants within a population[19–21, 23, 26]. At present, it is economically feasible to perform studies that use fairly high density SNP arrays to simultaneously assay 300k-2.5M SNPs across the genome in thousands of individuals. This has the potential to yield a genomic coverage of 65-91% of common variation (with a minor allele frequency (MAF) of at least 5%) in European populations after combining study genotype data with genotypes from HapMap phase 2[20][26, 27] (see Section 1.2.3). SNP array panels now on the market include up to 4.3M SNPs in a single array. The increased numbers of marker SNPs simultaneously lower the MAF at which the panel can usefully be used as proxies for underlying causal variants, and also increase the resolution at which the most likely causal variants can be mapped.

At the same time, large-scale sequencing efforts to increase our understanding of the

diversity of human populations such as 1000 Genomes Project (1000G)[28] will further lower the MAF and increase the resolution at which panels of marker SNPs can be useful by generating higher resolution recombination and haplotype maps than HapMap phase 2[20] by including more rare variants as well as other types of (non-SNP) variants such as insertions, deletions, and structural variants, which can be represented by tag SNPs once LD patterns between those variants and potential tag SNPs has been established by typing them in reference populations such as 1000G.

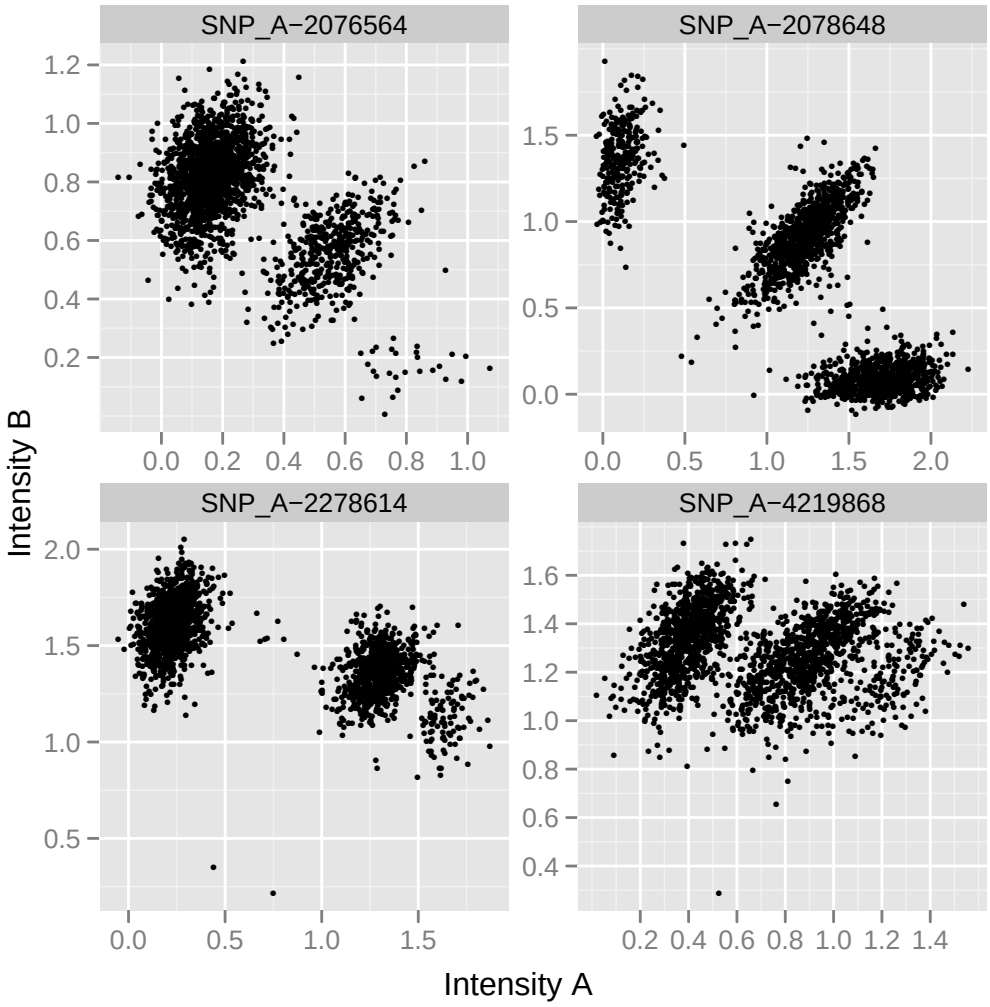
1.1.2 Genotyping methods

Data on an individual's genotype can be generated using a wide variety of laboratory techniques. In the discovery phase of genome-wide association studies, data are typically generated using microarrays consisting of probes that simultaneously assay $300k - 2.5M$ common SNPs from a sample of DNA extracted from an easily obtainable source such as blood or saliva. The raw data from these microarrays typically consist of points along two intensity axes and some calibration data for normalisation[29]. Figure 1.1 is an example of a plot of normalised intensity data for a SNP. The data format of these intensity files is dependent on the technology used, and due to differences in microarray design the calling algorithms used to process the raw intensity data into genotypes are typically specialised for a particular technology as well.

1.1.3 Genotype calling algorithms

Two of the largest commercial vendors of genome-wide association microarrays, AffyMetrix and Illumina, both have in-house propreitary calling algorithms (BRLMM and GenCall), though a number of third-party software packages for calling genotypes have also been developed[30–35]. In general, each calling algorithm works by defining genotype clusters based on the 2-dimensional intensity data and then either assigning individual points to a particular cluster or determining the posterior probability that an individual point belongs to each of the clusters. The output from any of these

Figure 1.1: Normalised intensity plots for some example SNPs.



calling algorithms typically results in post-processed data that falls into one of two main categories. The first (called genotype data) preserves only the most likely genotype at each SNP, while the second (genotype uncertainty data) carries forward some measure of uncertainty in the genotype calls.

1.1.4 Called genotype data

In studies of SNPs, data passed into downstream analyses are typically of the called genotype type. In a format of this type, data at each SNP are reduced down to one of the three possible values for the genotype, or alternatively a missing value code is used if the calling algorithm could not determine the most likely genotype beyond a confidence threshold.

For instance, in the Pedigree file for PLINK (PED) format used by the PLINK software package[36], genotype values for a given SNP are coded using either the single letter code for each of the two alleles at that SNP, or another letter (such as '0') to indicate missing data. For example, the four possible codes for a genotype at a SNP for which the two alleles are 'A' and 'C' would be 'A A', 'A C', 'C C', or '0 0', where the first three codes correspond to the three genotype classes ('A' homozygote, heterozygote, and 'C' homozygote), and the fourth code ('0 0') is the flag indicating missing data at that SNP[37]. Listing 1 shows an example of a very basic PED file.

Listing 1: An example PED file showing 4 individuals and 6 SNPs.

```
F1  I1  0  0  1  0    A A      G G      A C      C T      C C      0  0
F2  I2  0  0  1  1    C C      A G      0  0      C T      T T      A  T
F3  I3  0  0  1  1    A C      A G      0  0      C T      C T      A  A
F4  I4  0  0  1  1    C C      G G      A C      T T      C T      A  A
```

However, since there are only 4 possible values given at each SNP, equivalent information can be represented using only 2 bits of data for each SNP-individual combination (with the addition of a small amount of meta information indicating the allele labels for each SNP). Therefore, to reduce storage space and processing time, alternative

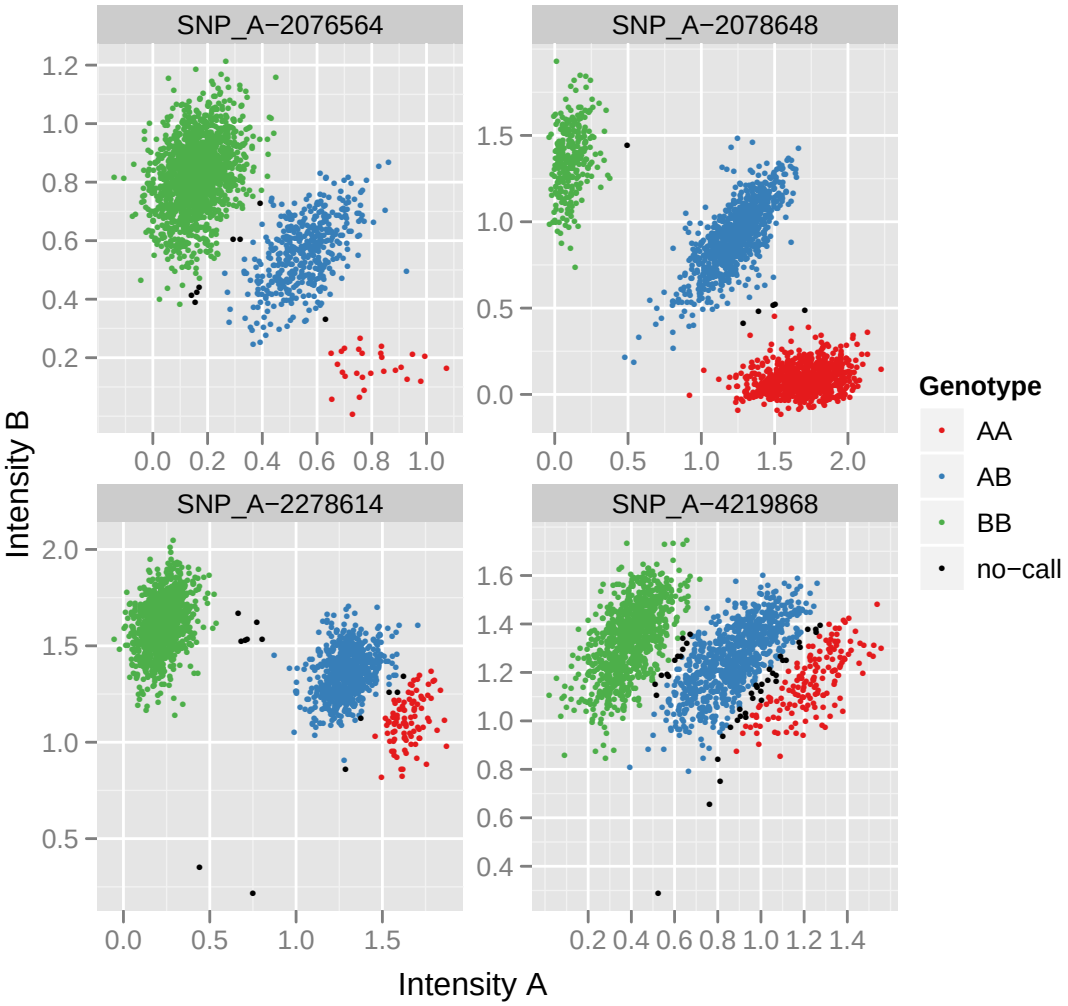
formats have been developed which store the data in a packed binary format. The PLINK Binary PED for PLINK (BED) format is an example of a more compact binary representation of the 'called genotype' format[37, 38]. In the BED format, allele labels are stored in a separate file that lists the first and second alleles, and the binary BED file uses 2 bit codes to store the genotype ('00' for a homozygote for the first allele, '10' for a heterozygote, '11' for a homozygote of the second allele, and '01' for a missing genotype). Listing 2 shows an example of a BED file.

Listing 2: A representation of a 9-byte BED file containing the same genotype data as the PED file in Listing 1. Each byte in the file is displayed as 8 binary digits, with spaces after each 2-bit pair in order to assist in reading genotypes. The first three bytes contain version information and can be ignored, but the remaining bytes contain the genotype data. Data are in 'SNP-major' mode, meaning that all individuals for the first SNP go first, followed by individuals for the second SNP, and so on. However, each byte is read backwards, so the first genotype in the file is represented by the last two (7th and 8th) bits of the 4th byte and the second by the 5th and 6th bits of the 4th byte, the third by the 3rd and 4th bits of the 4th byte, the fourth by the 1st and 2nd bits of the 4th byte, and then the fifth by the last two bits of the 5th byte and so on until the last individual at the 1st and 2nd bit of the 9th byte.

```
01 10 11 00      00 01 10 11      00 00 00 01
11 10 11 00      11 10 10 11      10 01 01 10
11 10 10 10      10 10 11 00      00 00 10 01
```

One limitation of the called genotype format is that it requires the establishment of a threshold for the confidence in genotype calls under which the data will be marked missing. However, in most cases when a point does not lie within one of the clusters, a calling algorithm is likely to have higher confidence that the point belongs to one or two genotype clusters and lower confidence in the other cluster or clusters, so to simply call the genotype missing is to lose potentially informative data, resulting in a potential loss of power. While it may make sense to exclude (as missing) points that lie far away from all clusters, such as those near the 'A' intensity axis in the bottom two panels of Figure 1.2; it is unfortunate to have to exclude points that lie between two clusters, such as those between the red and blue cluster in the lower right panel of Figure 1.2 which obviously are much more likely to belong to either the red or the

Figure 1.2: Normalised intensity plots for some example SNPs, showing the results of calling with CHIAMO[31] with a threshold of 0.8. Black points (labelled 'no-call') indicate missing calls, indicating that no genotype had a probability greater than 0.8 for that point).



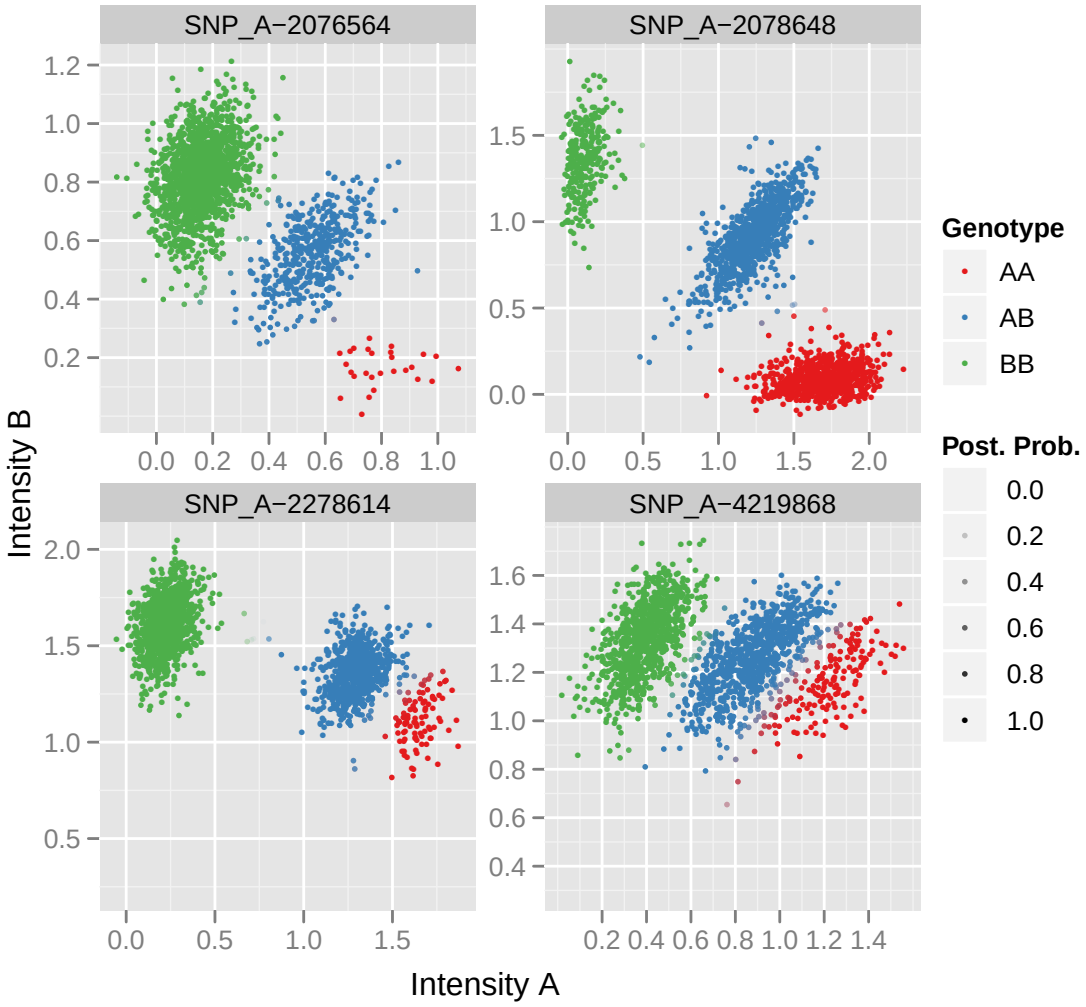
blue cluster than to the green cluster. The genotype uncertainty data format could represent that by setting a very low probability on belonging to the genotype class represented by the green cluster, and perhaps distributing the rest of the probability equally between the genotype classes represented by the red and blue clusters.

1.1.5 Genotype uncertainty data

With genotype uncertainty data, the genotype at a given SNP is typically represented as a set of probabilities for each genotype—individual combination and therefore carries information about the uncertainty of the genotype calls. In these formats, one probability for each of the possible values of the genotype is given – homozygous for the first allele, heterozygous, and homozygous for the second allele. If the calling algorithm were 100% certain that the genotype was heterozygous, it could be coded as something like '0.0 1.0 0.0' (this is the format output by the CHIAMO[31] genotype calling software and the IMPUTE[27] imputation software, as well as that used by the SNPTEST[39] and QUICKTEST[40] genetic association testing software).

Unlike the called genotype format, missing values do not typically exist in the genotype uncertaintly format, but rather overall uncertainty can be represented by the three probabilities summing to a value less than 1. In Figure 1.3, the data are plotted on the same intensity axes as in Figure 1.2, but in this case with each point plotted as a blend of the three genotype colours and the probability of each genotype represented by the opacity of that genotype's colour, with fully opaque points indicating individuals with a probability of 1 for the genotype corresponding to its colour and fully transparent (not visible) points representing individuals with a probability of 0 for all three genotype classes (100% missing data). As a result, points between clusters that had been called missing in Figure 1.2 are now light shades of magenta (when between red and blue clusters) or cyan (when between green and blue clusters), and points that are very far away from clusters are almost entirely transparent and thus not visible.

Figure 1.3: Normalised intensity plots for some example SNPs, showing the results of calling with CHIAMO[31], displaying the genotype probabilities in the alpha channel (opaque indicates a probability of 1, transparent indicates a probability of 0).

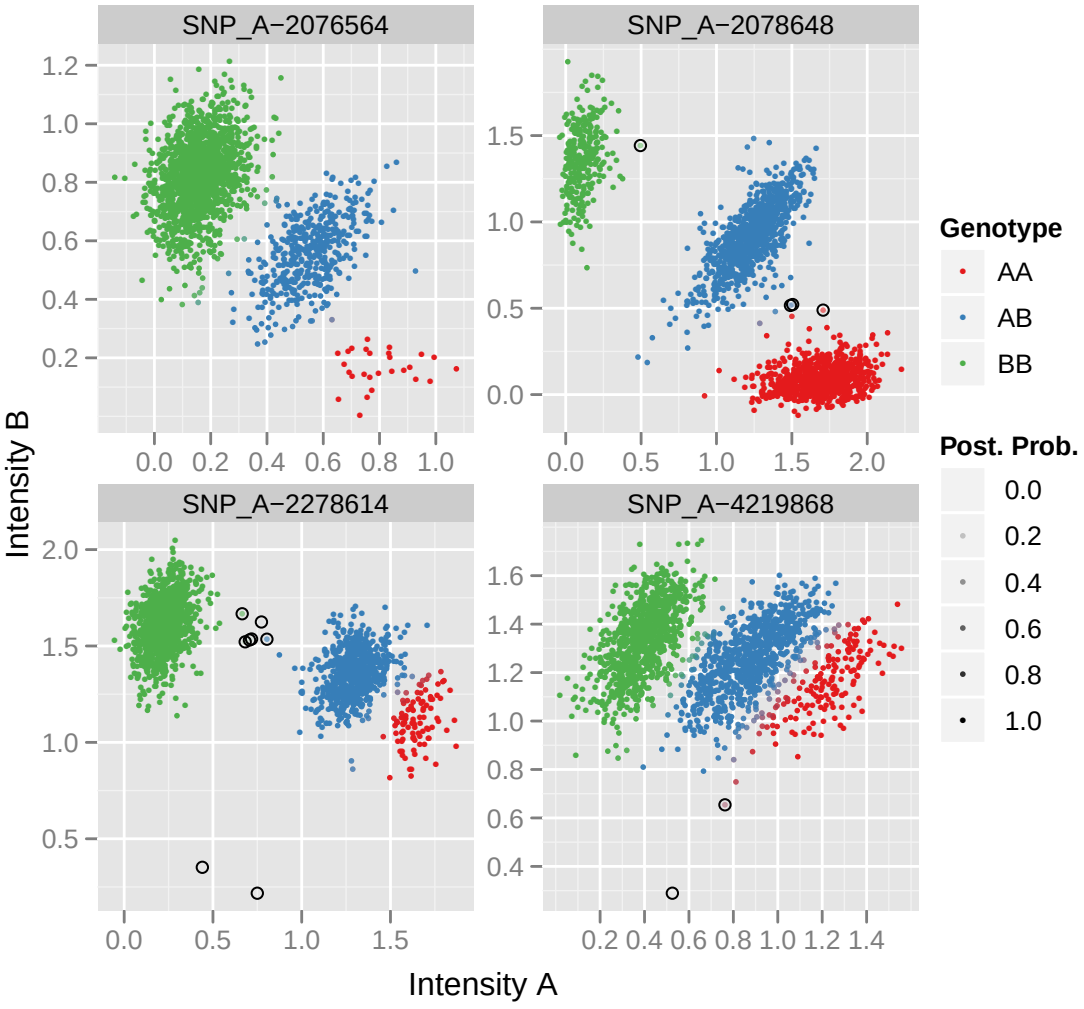


Some calling algorithms, such as CHIAMO[31], handle missing data by allowing for the identification of additional clusters (called null clusters), beyond the three genotype clusters, to represent genotyping failure (e.g. because of the failure of one or both of the probes represented by the two intensity signals). Increased probability of membership in a null cluster reduces the certainty of the other three genotype probabilities, because the sum of the probabilities across all possible clusters considered by the calling algorithm must be equal to 1. Thus, belonging in part to a null cluster effectively makes the genotype partially missing. An example of individuals that have partially missing data is shown in Figure 1.4, in which individuals wherein more than 20% of the probability has been assigned to a null cluster are circled. Note that within this set, there is a large variation between those which are almost certainly an error (such as those near the 'A' axis in the lower two panels) that tend to appear almost, if not entirely, transparent and those which are in between the clusters, such as the magenta shaded points between the red and blue clusters in the upper right panel, and the cyan shaded points between the green and blue clusters in the lower-left panel.

1.2 Genetic association studies

Early genetic association studies were typically carried out in small regions of the genome, such as the regions surrounding the coding region of a gene that researchers hypothesized was a good candidate for association (e.g. because of known biological function or association with monogenic disorders). Following the discovery by the HapMap project[20, 24] that LD patterns across the human genome were organised into local recombination hotspots[20], high-coverage SNP genotyping arrays have made possible the GWAS, in which a dense panel of marker SNPs is used to tag the majority of common genetic variation across the genome. Coverage of these arrays varies, both by product and depending on the population being studied. Early products were based on LD patterns found in early releases of HapMap data and were best suited

Figure 1.4: Normalised intensity plots as in Figure 1.3 but in which individual points with greater than 20% missing are circled (for which the sum of the probabilities of all three genotypes is less than 0.8 ($AA + AB + BB < 0.8$). This plot; as well as the plots in Figure 1.1, Figure 1.2, & Figure 1.3; were created in R[41] using the ggplot2[42] and reshape[43] packages with colour palettes from ColorBrewer[44]. Source data are from the WTCCC T2D genome-wide scan[45].



to analysis of populations similar to the HapMap population of utah residents with ancestry from northern and western Europe (CEU)[24]. More recent products have added many times more markers and have based marker selection on the additional populations of HapMap phase 3[46][46] as well as SNPs detected as part of 1000G. Modern whole-genome arrays, such as the Illumina HumanOmni5-Quad BeadChip, can now capture 91% (at an $r^2 > 0.8$) of common variation (MAF $> 5\%$) in the CEU population; 89% of the common variation in the combined HapMap population comprising samples of Han Chinese in Beijing, China (CHB) & HapMap population comprising samples of Japanese in Tokyo, Japan (JPT) HapMap populations; and 70% of the common variation in the HapMap population comprising samples of Yoruba in Ibadan, Nigeria (YRI) HapMap population[46].

1.2.1 Quality control

1.2.1.1 Experimental replicates and basic genotype QC

It may also be a good idea to include a few samples with "known" genotypes (such as HapMap samples) along with study samples in order to validate that the genotyping platform is operating as expected[47]. After genotyping, there are a number of checks that should be performed in order to exclude samples and/or SNPs that did not perform well in genotyping, including checks of concordance rate, missingness rates for both SNPs and samples (after base calling), and an investigation into batch effects.

1.2.1.2 Departure from Hardy-Weinberg Equilibrium

The Hardy-Weinberg theorem states that in a large enough population undergoing random mating, the frequencies of the three genotypes at a biallelic genetic variant will tend to remain in equilibrium over time when not under the influence of selection or mutation, following the relationship p^2, $2pq$, and q^2; where p and q are

the frequencies of the two alleles[48]. Departures from the expected genotype frequency distribution predicted under Hardy-Weinberg equilibrium (HWE) can be due to inbreeding, population stratification, selection pressure, or problems in genotyping (including non-random missingness, miscalled heterozygotes, or unobserved common variation that disrupts primer sites)[21, 49, 50].

Testing for deviation from HWE is typically carried out using either a Pearson goodness-of-fit test (χ^2 test) or a Fisher exact test, both of which yield a p-value indicating the significance of departure from the null hypothesis (H_0) that the genotype frequencies are in HWE[21]. Extremely significant departures from HWE are likely to indicate quality issues and such SNPs can be safely discarded[45], though the choice of an appropriate threshold to use for this criterion will vary with both sample size and data quality and should not be set too liberally, since moderately significant departure from HWE can also result from real associations under the influence of selection[2, 21, 49].

Some studies have based their HWE quality control (QC) criterion on the observed distribution of the p-values for HWE disequilibrium in relation to call rate failures[45], while others have used a relatively stringent threshold, such as $P_{HWE} < 10^{-6}$, as a cut-off to exclude SNPs[47]. However, setting such a stringent threshold may exclude SNPs that are under significant selective pressure that could potentially be due to the trait or disease under study (though that is less of an issue for quantitative traits than for case-control studies)[21].

When considering excluding a SNP based on departure from HWE, manual examination of its intensity plot (see Section 1.2.6.1) may provide insight into why the issue is occuring if it is actually due to genotyping failure or poor genotype calling. SNPs that have obvious genotyping or calling problems should be excluded from the analysis if the issues can't be resolved by performing the genotype calling again with different parameters or alternative software. SNPs that do not appear to have any genotype calling problems but that nonetheless display significant departures from

HWE should be retained, as they could be important markers of latent population stratification that needs to be controlled for in the analysis (see Section 1.2.5.1), and removing them may limit the effectiveness of empirically based controls for population stratification (such as PCA; see Section 1.2.5.3). Another approach for handling such a SNP would be to initially exclude it, subsequently fill in its genotypes using imputation (see Section 1.2.3), and finally compare the imputed genotypes with the initial genotypes[21].

1.2.2 Phenotype

1.2.2.1 Quantitative traits

Quantitative phenotypes vary continuously across a range of values, though limited precision of measurements may tend to discretise the values to some degree. QC of quantitative traits is important in order to verify that phenotype data are in the correct units, which can be checked by applying liberal thresholds to phenotype data, imposing limits on the reasonable range for the trait in the specified units and for the expected population. It can also be a good idea to check for outliers and take extra steps to ensure that data with extreme values are correct, such as going back to clinicians who supplied the measurements to verify the extreme values observed agree with their records and are not simply a typographical or data entry error.

Association testing can be performed directly on raw quantitative traits, though it is important to test the extent to which the trait is normally distributed in the population as many of the statistical tests to evaluate significance of an association require a normally distributed trait in order to be valid[21]. To determine whether a trait is approximately normally distributed, one could simply plot a histogram of the data and visually inspect the shape for deviations from normality in terms of skewness (asymmetry of the distribution in either direction) and kurtosis (being more or less peaked than normal)[51]. However, in a large sample it is possible that even small

deviations that are not obvious to the eye could result in significant non-normality, so in most cases it is a good idea to perform a formal test for non-normality (see Figure 1.5 for examples of significantly non-normal distributions that may not be obvious by eye). Several tests for normality can be performed computationally, such as the Shapiro-Wilk Test, for sample sizes up to 5000[52], or the Lilliefors test[53]. When normality does not hold, a transformation can potentially be used if it makes the data appear more normally distributed (see Section 1.2.2.2), or samples could potentially be stratified into subgroups for analysis if the non-normality were due to a bimodal or multimodal distribution across multiple subgroups of the samples and stratifying them by those subgroups would therefore result in a normal distribution for each.

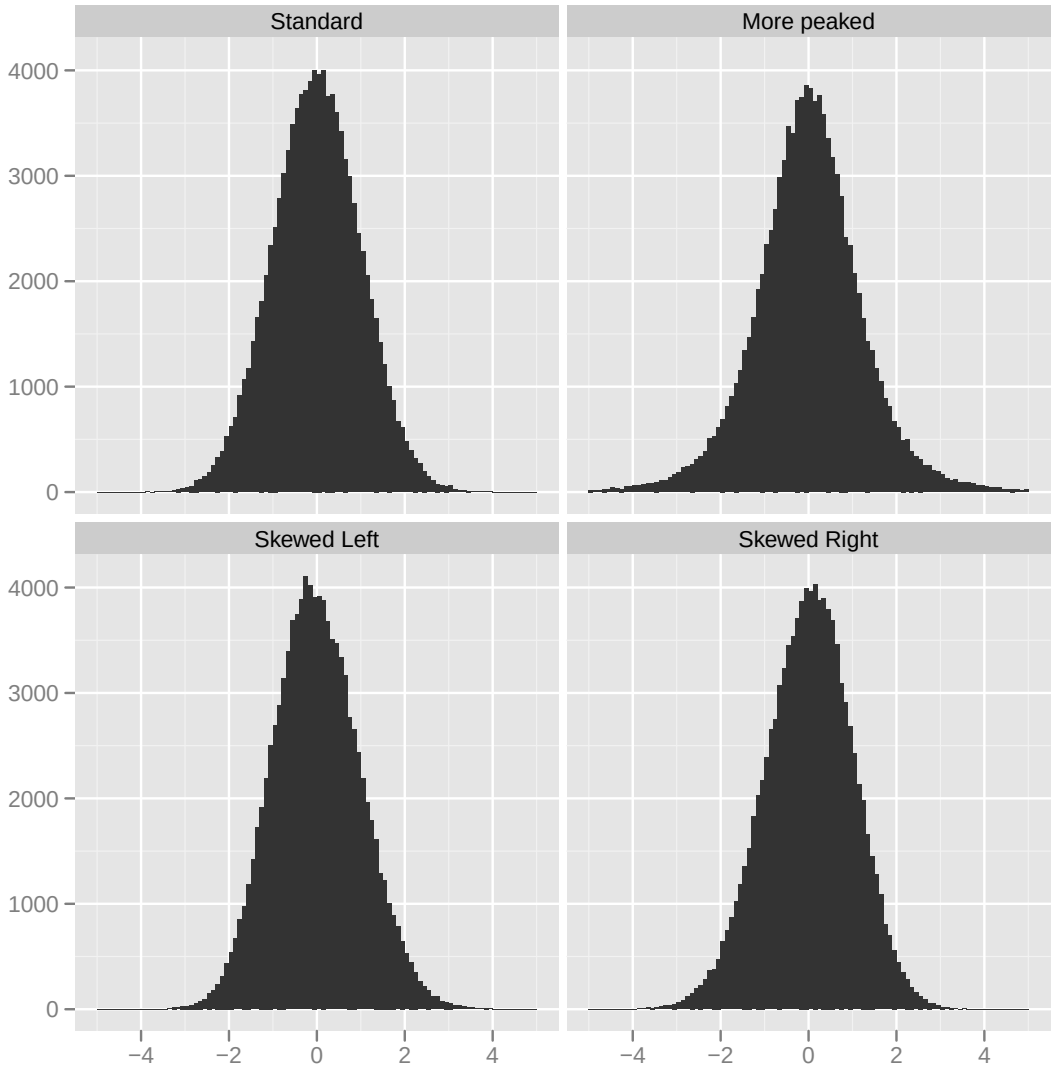
1.2.2.2 Phenotype transformation

One particularly useful transform to force phenotype data to be ly distributed and thereby satisfy the requirements of tests requiring normality is the inverse-normal transformation. The inverse-normal transform is a rank-based transformation that simply orders the phenotype data from lowest to highest and then fits that data to a standard normal distribution[3, 5, 54, 55]. While this will make the assumptions of statistical tests that require normal data valid, the drawbacks are a potential loss in power (e.g. due to amelioration of extreme values) as well as a lack of useful units in the resulting effect estimates (the units of the effect estimate will be in standard deviations).

1.2.2.3 Treating quantitative traits as case-control phenotypes

A simple phenotype is one in which the sampled populations are divided into two groups: cases and controls. Case-control phenotypes are commonly used in studies of disease, but a quantitative trait can also be treated as case-control data by taking the extremes of the distribution or by grouping into cases and controls according to

Figure 1.5: A standard normal distribution compared to three distributions that are significantly non-normal. All four distributions are based on 100,000 simulated data points. On the upper left ("Standard") is a standard normal distribution. To its right is a distribution that displays excess kurtosis (high peak and wide tails). On the bottom row are distributions skewed to the left or the right. Estimates of skewness, kurtosis, and a p-value for non-normality based on the Lilliefors normality test for each of the distributions were calculated. The upper left distribution had skewness=0.15, kurtosis=-0.01, and $p_{Lilliefors} = 0.9$, indicating it is not significantly different from a normal distribution. The upper right distribution had skewness=-0.22 and kurtosis=8.1, the lower left distribution had skewness=0.15 and kurtosis=0.004, and the lower right distribution had skewness=-0.14 and kurtosis=0.02. The results of the Lilliefors test for all three distributions was significant $p_{Lilliefors} < 2.2 \times 10^{-16}$, indicating the distributions are all significantly different than a normal distribution and thus it would not be safe to use tests that require normality with any of these distributions.



absolute thresholds, such as comparing "normal" and "obese" ranges of body mass index (BMI))[56, 57]. In the most basic analysis of case-control data, only the counts of cases and controls of each genotype are important for the test, and this data can be organised into a 2×3 contingency table (see Table 1.1) for each SNP, which is then tested for association between the rows (case-control status) and columns (genotype), using a test such as a Pearson test with 2 degree of freedom (DF), a Fisher exact test, or a Cochran-Armitage trend test (which may be the best option when assuming additive genotype risk)[21].

Table 1.1: An example 2×3 contingency table for a case-control phenotype.

		Genotype		
		AA	*AB*	*BB*
Status	*case*	49	42	9
	control	20	50	30

1.2.3 Genotype imputation

Genotype imputation is a method by which missing genotype data can be estimated by utilizing information about correlation patterns in non-missing data[58]. A number of different imputation methods have been developed for use with genotype data[27, 58–64], most of which employ hidden markov models (HMMs) to estimate missing genotypes while simultaneously modeling the uncertainty regarding haplotype phase[63]. A typical use of imputation in GWAS is to include data from an appropriate reference population, such as the CEU population from HapMap phase 2[20][20] for studies of individuals of European descent, in order both to increase coverage and to have results at the full set of SNPs across multiple studies. Without imputation, meta-analyses of studies that used different platforms could easily result in a rather small set of overlapping SNPs that would not be considered as having genome-wide coverage (depending on the overlap between the genotyping platforms used), so the use of imputation to obtain comparable sets of genome-wide SNPs genotypes has become common in GWA meta-analysis (GWAMA)[3–5, 65].

1.2.3.1 Imputation accuracy

An aspect of imputation of particular importance is that the accuracy of imputation can vary substantially from SNP to SNP for reasons such as local LD structure and/or allele frequency differences between the reference panel(s) and genotyped population(s) being imputed[66]. It is therefore a good idea as part of QC to filter the output of imputation by some measure of the imputation accuracy in order to exclude poorly imputed SNPs from analyses, or alternatively to perform association testing that takes the genotype uncertainty into account (see Section 1.2.5.4).

Several different measures of imputation accuracy are reported by various imputation software. One that is typically reported is simply the probability that an imputed genotype is correct[64]. While intuitive to understand, a threshold based on probability is not very useful for comparing SNPs of different allele frequencies because a reasonable threshold for common SNPs (such as $probability > 0.9$) would not as useful for rare SNPs because no basis on surrounding haplotype structure would be necessary to exceed the threshold[66]. For example, for a 5% SNP, simply assigning the most frequently observed genotype to every individual should exceed a $p > 0.9$ threshold, but would not represent any of the actual genotypic variation observed in the sample and would therefore simply be adding noise to an analysis of that SNP[64].

The most commonly used metrics of imputation accuracy for a SNP are estimates of the squared correlation between observed allele dosage and the expected allele dosage under HWE[62, 67]. In MACH[64] the estimate of this is referred to as $\hat{r^2}$, while in BEAGLE[62] it is referred to as "Allelic R^2." The two measures are similar but are estimated in different ways, with MACH using a ratio of the empirically observed variance in the imputed data to the expected variance of the allele dosage under HWE, and BEAGLE estimating the squared correlation between the most likely genotype and the true genotype[66]. The information content reported by PLINK[37] imputation and referred to as "INFO" is also equivalent to the $\hat{r^2}$ measure reported

by MACH[67].

When using imputation software that does not provide an accuracy measure (such as BIMBAM[58]), $\hat{r^2}$ can be calculated after imputation given an expected allele dosage (e) and the MAF of the SNP in the reference population. The ratio of observed variance in allele dosage in the imputed data ($var_{observed}$) to the theoretical expected variance in allele dosage under HWE ($var_{expected}$) yields $\hat{r^2}$ (Equation 1.1). Under HWE, the expected variance of allele dosage ($var_{expected}$) can be estimated by $2p(1-p)$, where p is the population allele frequency[62]. Since the true population allele frequency is unknown, the observed MAF in the reference panel can be used as an estimate for p (Equation 1.2). The variance of the observed allele dosages ($var_{observed}$) can be calculated from the observed allele dosage (e_i), where i is a sample index, using the standard formula for sample variance across all samples (Equation 1.3). This yields an expanded formula for $\hat{r^2}$ in terms of MAF, e, and N (the number of samples) that is shown in Equation 1.4[66].

$$\hat{r^2} = \frac{var_{observed}}{var_{expected}} \tag{1.1}$$

$$var_{expected} = 2\text{MAF}(1-\text{MAF}) \tag{1.2}$$

$$var_{observed} = \frac{1}{N}\sum_{i=0}^{N} e_i^2 - (\frac{1}{N}\sum_{i=0}^{N} e_i)^2 \tag{1.3}$$

$$\hat{r^2} = \frac{\frac{1}{N}\sum_{i=0}^{N} e_i^2 - (\frac{1}{N}\sum_{i=0}^{N} e_i)^2}{2\text{MAF}(1-\text{MAF})} \tag{1.4}$$

The SNPTEST[27] software, which is designed to perform association tests using imputation that has been carried out using IMPUTE[27], takes into account the uncertainty in each of the reported genotype calls (see Section 1.1.5) and reports an additional information measure (I_S) which in the output files is typically called either

"proper_info" (SNPTEST version 1) or "info" (SNPTEST version 2). This metric is an estimate of the ratio of the observed information ($i_{observed}$) to the complete information ($i_{complete}$) (Equation 1.5) for the SNP[66] and is directly related to the power to detect the genetic effect estimated by the association test[67], in that multiplying the number of individuals in the imputed sample (N) by I_S will result in the effective sample size ($N_{effective}$) if all of the samples had genotypes known with complete certainty (Equation 1.6)[68]. The I_S metric is typically used as a threshold for exclusion of SNPs after association testing for GWAS that use IMPUTE and SNPTEST, with a typical threshold of $I_S > 0.4$, or in some cases $I_S > 0.5$[3, 5, 54, 55, 65, 68–71]. Version 2 of IMPUTE also includes a similar information measure (I_A) which also represents the ratio of observed to complete information. While I_A only represents the uncertainty in the genotypes resulting from imputation, the I_S measure also takes into account the genetic model being tested and therefore under some models the two measures can be quite different, although they are highly correlated under the additive model[66].

$$I_S = \frac{i_{observed}}{(i_{complete})} \tag{1.5}$$

$$N_{effective} = I_S \times N \tag{1.6}$$

1.2.4 Genetic association analysis

While more complex methods exist, such as those that analyse a large set of SNPs simultaneously[72], most GWAS currently use more straightforward methods that test a single SNP at a time. For genetic association testing, the choice of model is important as choosing a model that does not accurately capture the true variation in the data can drastically limit the power of a study to detect those variants[73].

1.2.4.1 Additive, dominant, and recessive models

A variety of models can be used to test for association between genotype and phenotype. The additive, dominant, and recessive models are all 1-DF tests and correspond directly to the additive, dominant, and recessive modes of mendelian inheritance[73]. In the additive model, the number of copies of an allele (0,1,2) at a particular SNP is used as the value to be tested for association with the phenotype. Thus, genotype AA has a value of 0, AB has a value of 1, and BB has a value of 2 (using B as the effect allele). In the dominant model, AA would have a value of 0, while AB and BB would each have a value of 1. In the recessive model, AA and AB would each have a value of 0, while BB would have a value of 1.

Typically, only the additive model is used in a GWAS analysis. This is perhaps because it is the most straightforward to explain, but it also has fairly good power to detect signals that are actually dominant or those which are actually involved in certain kinds of multiple locus interactions, though it is not well powered to detect signals that are actually recessive or which exhibit many kinds of multiple locus interactions[73].

Simulations have shown better power can be gained by performing all three tests, although that only applies when using empirical means such as permutation to correct p-values, rather than more conservative means such as bonferroni corrections (see Section 1.3.3.1), or alternatively just performing a single 2-DF test of a co-dominant model[73].

1.2.5 Confounding factors and methods to control confounding

In a genetic association study, confounding factors (confounders) are variables that are to some degree correlated with both the genotype and phenotype under test but

that are not on the causal pathway between genotype and phenotype. If not controlled for, confounders can result in spurious (false-positive) associations.

Methods to control for confounders includes case-control matching, use of cohort studies, stratification-by-confounders, and adjusting for confounders by including them as covariates in a multivariate statistical test. In a case-control study, case-control matching can control for confounding by selecting for each case a control that closely matches for all confounding factors. However, unless a quantitative trait has been converted to case-control status (e.g. by taking extremes) and is being treated as a case-control study, this method will not be possible. In that instance, study design could have taken confounders into consideration from the outset by being organised as a cohort study in which all potential confounders have been matched at the outset. For example, some genetic studies control for age, birth year, and some environmental factors by employing a birth cohort in which all subjects were born in the same period within a localized area.

1.2.5.1 Population stratification and cryptic relatedness

It is important to be aware of the potential for population stratification, which occurs when the population being studied includes sub-populations that are more closely related to each other than other members of the population as a whole[21]. If those sub-populations also have a different distribution of the quantitative trait under study, then genetic variation that happens to have a different allele frequency in that sub-population will appear to be associated to the trait, when in fact it is likely to just be an indicator of membership in that subgroup[23]. It may be obvious that these corrections are necessary when gathering data from different populations across continents or countries, but even for a study that draws only from a well-defined population in one country or city, the potential problems of population stratification can't necessarily be safely ignored and should instead be examined and, if necessary, corrected[74].

1.2.5.2 Genomic control

Genomic control (GC) is a method that can be used to control for inflation in a test statistic such as could be due to population stratification due to admixture. GC relies upon having data on a large number of markers, the vast majority of which have no real association with the trait being studied[75–78]. For each of these markers, association with the trait is tested. The median value is then divided by the theoretical median underH_0 to yield λ_{GC}, which is a measure of the inflation present in the data and which, if $\lambda_{\mathrm{GC}} > 1$, can then be used to adjust the test statistic[77]. The method to perform GC correction on a χ^2 statistic is to simply multiply it by $1/\lambda_{\mathrm{GC}}$ (Equation 1.7)[75, 77], and a standard error (σ) can also be easily corrected by multiplying by $\sqrt{\lambda_{\mathrm{GC}}}$ (Equation 1.8)[67].

$$\chi^2_{GC} = \chi^2 \times \frac{1}{\lambda_{\mathrm{GC}}} \tag{1.7}$$

$$\sigma_{GC} = \sigma \times \sqrt{\lambda_{\mathrm{GC}}} \tag{1.8}$$

In practice, genomic control λ_{GC} for a GWAS is typically calculated using data from all SNPs, even though some SNPs may be known to have association with the disease[3]. Though it would be possible to remove loci that have already been established to have an association with the trait from the λ_{GC} calculation, leaving those loci in the calculation only make the GC correction more conservative, so it is safe to do so while still getting the full benefit from the correction, though it will likely result in a loss in power.

Since the median is a robust estimator and is therefore not overly affected by individual outliers, a few real associations would have a negligible affect on the correction. If, however, a large number of SNPs have some amount of real association, overcorrecting could become a potential issue in that it could result in substantial power loss. It is therefore a good idea, especially when studying complex traits that may

have hundreds or thousands of real effects (such as height[70]), to consider removing markers from the calculation that have been established to be associated with the trait being studied, as well as other markers in close LD with those markers, in order to avoid overcorrection.

It is important to note that λ_{GC} would be expected to scale with sample size given stratification, so in order to compare inflation across studies or traits with different sample sizes, it can be useful to compare $\lambda_{1000,1000}$, which is the inflation factor for an equivalent study of 1000 cases and 1000 controls (Equation 1.9) or λ_{1000} for quantitative trait analysis of 1000 samples (Equation 1.10)[67, 79].

$$\lambda_{1000,1000} = 1 + (\lambda - 1) \times \frac{(\frac{1}{n_{cases}} + \frac{1}{n_{controls}})}{(1/1000 + 1/1000)} \tag{1.9}$$

$$\lambda_{1000} = 1 + (\lambda - 1) \times \frac{1000}{n} \tag{1.10}$$

1.2.5.3 Principal components analysis

Principal components analysis (PCA) or eigenanalysis is a technique used to reduce the dimensionality of a data set by creating a set of representative axes of variation based on eigenvectors calculated from the data[80]. The first axis, or first principal component (PC), is made up of the combination of measurements that accounts more the largest amount of variability in the data[81]. Each of the subsequent PCs account for the most variability remaining after all the previous ones have been taken into account. The full set of PCs would fully represent the original data, and could be used to reconstruct it, but an interesting property of PCA when applied to genome-wide genetic data is that when using it to reduce the dimensionality of genotype data taken from a population, the first few PCs typically vary along with differences in ancestry and can therefore be used to separate a population into subpopulations[82].

As a result, PCA can also be used as a tool to detect population structure and to correct for population stratification[82]. Since the first PC will explain a larger proportion of variance in the presence of a larger amount of population structure, performing a test on the proportion of variance explained by the first PC yields a p-value that is related to the amount of population structure present[81]. Setting a threshold for this p-value and applying it to the first few PCs until one is found that is not significant is an empircal way to decide how many (N_{PC}) PCs to use to correct for population stratitification[83], though this number can also be chosen based on a prior belief regarding the number of subpopulations (N_{POPS}) likely to be present in the population, in which case $N_{PC} = N_{POPS} - 1$[84]. Alternatively, PCs can be calculated including known reference populations (such as HapMap populations) that are thought to be informative subpopulation differentiators for the population under study and an appropriate N_{PC} determined by examining each PC's ability to differentiate the known populations, such as by examining plots of the PCs including labelled reference populations (see Section 1.2.6.2).

However, it is important not to use too many PCs, as at some point the PCs will stop representing population structure and begin to represent variation that is likely be related to the trait being studied, resulting in a loss of power. This is made especially clear when considering the case when N_{PC} is equal to the full set of PCs, which would represent all genetic variation and thus correcting for them would result in no variation for association testing.

PCA algorithms can be quite computationally intensive, though less so than other methods such as structured association (SA)[84]. To reduce the computational burden, especially on memory requirements, a reduced set of genotypes is sometimes used to compute the PCs used for population stratification correction in a GWAS. Software, such as eigensoft[83], which are designed to perform PCA on genetic data are typically limited in the number of samples that can be processed due to memory constraints (issues with current versions of the software tend to begin when the number of samples

exceeds 10,000)[82], though new methods are being developed that can perform PCA on large data sets such as full GWAS scans, one of which is a multistage approach in which SNPs are divided into into subsets, PCA is run on each subset, and the top PCs resulting from each of those PCAs are carried forward into a final PCA, the results of which are used in the usual way[84].

1.2.5.4 Statistical methods for genetic association testing

The statistical methods used depends on the type of trait being analysed, but for quantitative traits, the frequentist test employed is typically a simple as linear regression or an analysis of variance (ANOVA)[21]. However, more advanced methods are needed in order to take genotype uncertainty into account. Several software packages that implement such tests are available, including SNPTEST[27, 39], MACH2QTL[85], and QUICKTEST[40].

In order to take uncertain genotypes into account (other than simply thresholding on some value and calling the most likely genotype), several methods can be used. One straightforward approach under an additive genetic model is to convert the probabilities of the three genotype classes into an expected value for the allele dosage at one of the alleles[86, 87] and to test that expected allele dosage for association with the quantitative phenotype in a simple linear regression[40].

More advanced methods use missing data likelihood tests[66] or mixture models[88], which are more computationally demanding but that more fully incorporate the probabilities of the three genotype classes. Bayesian approaches for including the genotype uncertainty into the model have also been developed[66].

Simulations comparing thresholded, dosage, and mixture models have shown that the dosage model is comparable in power to the mixture model when analysing modest effects in a large sample size when there is relatively little uncertainty in the genotypes (based on imputation), although the mixture models had substantially more power

with large effects in a small sample size, especially when imputation accuracy or MAF was low[88]. In all cases, both the dosage method and mixture models had more power than using thresholded genotypes[88].

1.2.5.5 Automation of association analysis

Performing association analysis for use in meta-analyses of summary statistics can be prone to error if performed manually – especially when numerous analyses on different traits are being performed at the same time. Each study must perform the analyses in a compatible manner, and effectively communicating the analysis methods and even the file transfer formats required for the meta-analysis to individual study analysts can be difficult in itself.

To help reduce errors in association analysis, I have developed a pipeline that takes genotype, phenotype, and analysis driver files as input and performs all steps of the analysis that can be performed automatically through to providing the output in a standardised format. Through the use of this automated pipeline, the opportunities for making errors are reduced. While the pipeline itself could of course have an error, such an error would tend to occur systematically rather than sporadically, such that, should an error be suspected or detected, the source of the error can be tracked down and resolved once, and all affected results can be re-run using the updated pipeline code. In contrast, when running analyses manually, there are more opportunities to introduce errors through increased human interaction, but more importantly when errors are detected it is very difficult to know whether the error was due to a systematic problem with the procedure the analyst was following or with one of the software packages, or whether it was just a one-time error.

The core of the association analysis pipeline is based on GNU Make, with a variety of components written for Bash, Perl, GNU R, and Maxima.

The pipeline supports phenotype transformations, stratification, and covariates that

can be quantitative or categorical. Categorical covariates are handled by converting to a set of contrasts.

It is important to note that such a pipeline is not intended to replace the primary analyst on a study or to remove requirements for QC, as those steps still require a great deal of manual inspection. It is, however, well suited to the task of performing additional analyses on a study that has already been prepared for another analysis, such as analysing the covariates from a disease study as primary traits.

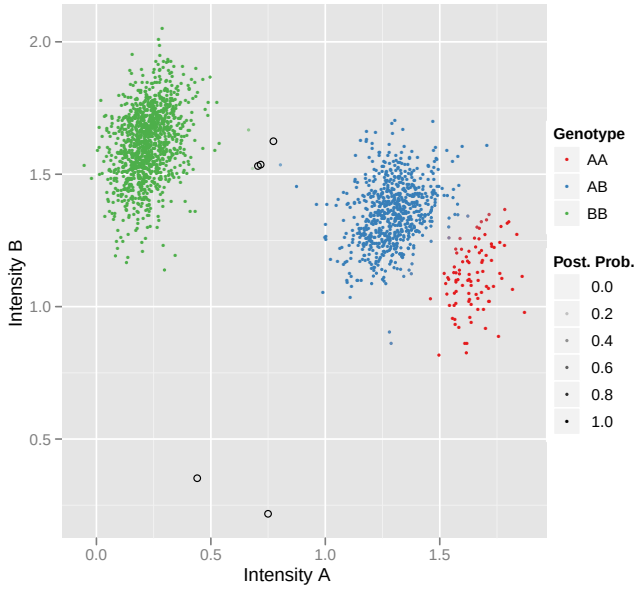
The pipeline allows the analyst to specify all parameters for a set of analyses in a driver file, and for those analyses to be run automatically on the data according to the contents of that file. As a result, the opportunities to introduce error have been reduced by encapsulating the human interaction into a single driver file.

1.2.6 Interpretation and visualisation of data

1.2.6.1 Intensity plots

An important tool for validating genotype calling is to make plots of individual marker intensity data along with an indication of the called genotype (see Figure 1.2). When genotype uncertainty data is used, it more useful to provide an indication of the confidence in each genotype call. A particularly nice way to do that is to make a scatterplot in which three points are overplotted for each SNP, one for each of the three genotypes, using a different colour for each genotype and using the alpha channel (transparency) to represent the uncertainty of that genotype call (ranging from 0 to 1).

Figure 1.6: A normalised intensity plot for a single SNP, showing the results of calling with CHIAMO[31], displaying the genotype probabilities in the alpha channel (opaque indicates a probability of 1, transparent indicates a probability of 0), and with nearly transparent points circled (those in which the sum of the probabilities of all three genotypes is less than 0.1 ($AA + AB + BB < 0.1$). This plot was created in R[41] using the ggplot2[42] and reshape[43] packages with colour palettes from ColorBrewer[44]. Source data are from the WTCCC T2D genome-wide scan[45].



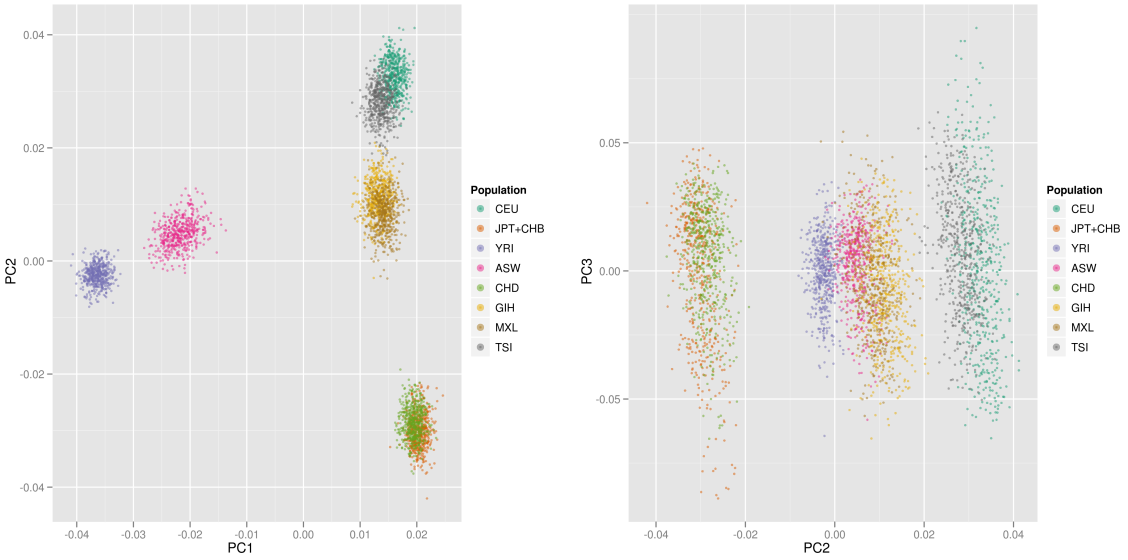
The result is that points that could equally belong to either of two genotype classes are displayed in a colour which is the result of the colour of the two classes mixed together (for example, a point between red and blue clusters would appear magenta). Since some points may appear almost completely transparent in these plots, it can be helpful to add a circle around points with low probabilities across all three genotype classes. Using these plots, performance of the calling algorithms can be evaluated by manual inspection (see Figure 1.6 for an example).

1.2.6.2 Plotting PCs with reference populations

To get an idea of the extent of population stratification in a particular sample, PCA can be used to calculate PCs based on reference populations such as those from HapMap, and and then the first few PCs plotted against each other along with the same

PCs evaluated in the population under study (or vice-versa, where PCs are calculated based on the population under study and then projected onto the reference samples). An alternative method is to run the PCA on the full combined data set including both the sample and reference populations, although the computational complexity of PCA increases significantly with sample size and, in particular, memory constraints may make it difficult to analyse extremely large sets of samples in one PCA.

Figure 1.7: The first three principal components calculated from the HapMap phase 2 populations (CEU, JPT+CHB, and YRI) and projected onto samples simulated from several HapMap phase 3 populations (ASW, CHD, GIH, MXL, and TSI). Eight populations of 500 samples were each simulated based on reference haplotypes for chromosome 21 (19,306 SNPs total) using HAPGEN2[89]. EIGENSOFT[82] was then used to calculate PCs based only on the three HapMap phase 2[20] populations, but projecting the results onto all eight populations.



(a) First and second principal components. **(b)** Second and third principal components.

Calculating PCs based on reference populations that include the population that will be used as a reference panel for imputation is particularly useful, since individuals not clustering with the reference population are good candidates for exclusion from the study[45]. Figure 1.7 shows an example of a scatterplot of the first three PCs resulting from a PCA calculated using only the three HapMap phase 2[20] populations, and then projected onto a number of other (in this case HapMap phase 3[46]) populations. In Figure 1.7a we see that the TSI population clusters very near the CEU cluster (and

indeed they are both of European origin), the CHD population clusters very near the JPT+CHB cluster (and both are of East Asian origin), the ASW population appears to fall along a line between the YRI cluster and the CEU cluster (which might be what we expect considering the ASW population is made up of African-Americans), and the GIH and MXL populations appear to lie on a line between the JPT+CHB cluster the CEU cluster. Note that some of the other populations have clustered very near the CEU HapMap population, while others are part way between the European (CEU) cluster and the East Asian (CHB+JPT) clusters. In Figure 1.7b, we see that the third PC does not appear to do very much to differentiate between the populations, which is what we'd expect given that only three populations were used to in the PCA to calculate PCs. If a study population to be imputed using the CEU reference panel was plotted along these axes, we might exclude any individuals who lie far from the CEU cluster in the 2-dimensional space of PCs 1 & 2.

Relative to the principal variation between the three HapMap phase 2[20] populations, some of the other five populations appear largely indistinguishable from each other using any of the PCs calculated based only on those three populations. For example, the GIH and MXL populations appear in Figure 1.7 to belong to a largely overlapping cluster, as do the CEU and TSI populations. If the population under study included individuals from these two populations, using only the PCs calculated from the three HapMap phase 2[20] populations would not do much to correct for population stratification at the level of the differences between CEU and TSI or GIH and MXL.

If the PCs are instead calculated from a larger set of HapMap phase 3[46] populations, more PCs can be used to differentiate between populations at a more fine-grained level. One way to visualize many PCs is to plot every PC against each other in all combinations, labeling known populations with colour in order to observe which PC can differentiate between which populations (see Figure 1.8). If samples of unknown population were plotted along with these reference populations, it could provide insight

into which individuals were outliers in terms of the variation observed between all of the reference populations (for the purposes of sample exclusion), or alternatively the plots can be used to decide which PCs are able to differentiate between population differences in the population under study in order to decide how many PCs to use as covariates in an association model when PCs were calculated along with these reference populations.

In Figure 1.8, all 100 combinations of the first ten PCs for simulated populations based on eight HapMap phase 3[46] populations are plotted against each other in a 10×10 matrix. For the first PC, looking down the first column one can see that it differentiates well between three clusters, one of ASW samples, one of YRI samples, and one containing all the other samples. Looking down the second column shows that PC2 also differentiates between three clusters, one of CEU and TSI samples, one of CHD and JPT+CHB, and one of the rest of the samples. The third PC appears to separate MXL samples from the rest of the samples, while the fourth is better at differentiating GIH samples from the rest of the samples. The fifth PC separates ASW from YRI more significantly than PC1, while the sixth separates TSI from CEU and the seventh separates CHD from JPT+CHB. Looking down the eighth, ninth, and tenth columns, it is clear that these PCs do not do much to differentiate between any of the populations. If using PCs based on these populations to correct for population stratification, based on this it would make sense to use the first seven PCs, as they are likely to be informative for variation along the axes of these eight populations, which again is what is expected given the eight populations used to calculate PCs.

1.2.6.3 Genome-wide association plots

Genome-wide association (GWA) plots, also called "Manhattan" plots[71], represent the association of a particular trait across all markers tested. These are typically plotted with log_{10} p-values on the y-axis and genomic position (chromosome by chromosome) on the x-axis. See Figure 1.9 for an example.

Figure 1.8: The first ten principal components calculated from eight HapMap phase 3 populations (ASW, CEU, CHD, GIH, JPT+CHB, MXL, TSI, and YRI). Each of the ten PCs is plotted against itself and the nine other PCs to form a 10×10 matrix of all combinations in order to show the relationships between them. Eight populations of 500 samples were each simulated based on reference haplotypes for chromosome 21 (19,306 SNPs total) using HAPGEN2[89]. EIGENSOFT[82] was then used to calculate PCs based on all eight populations.

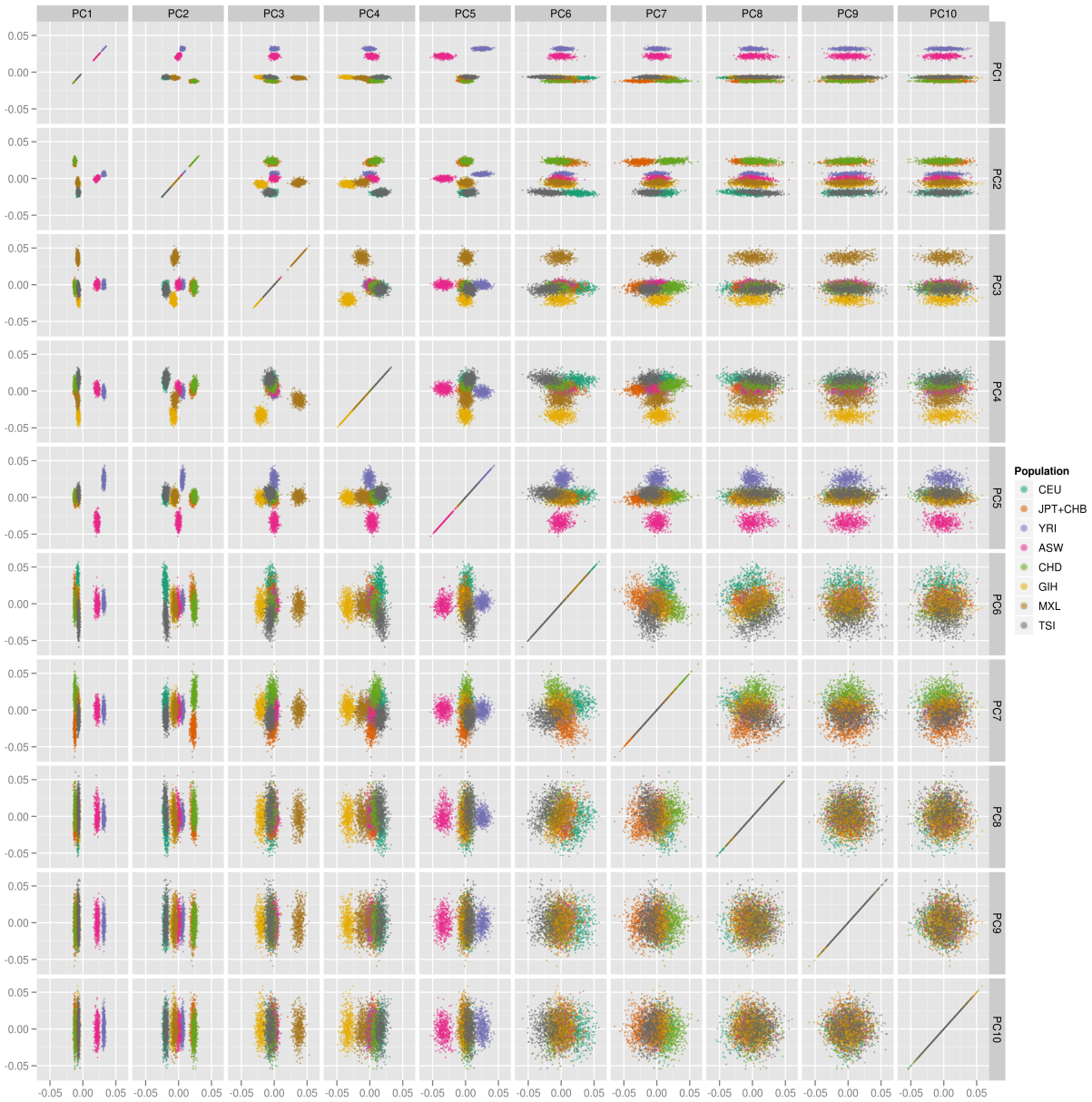
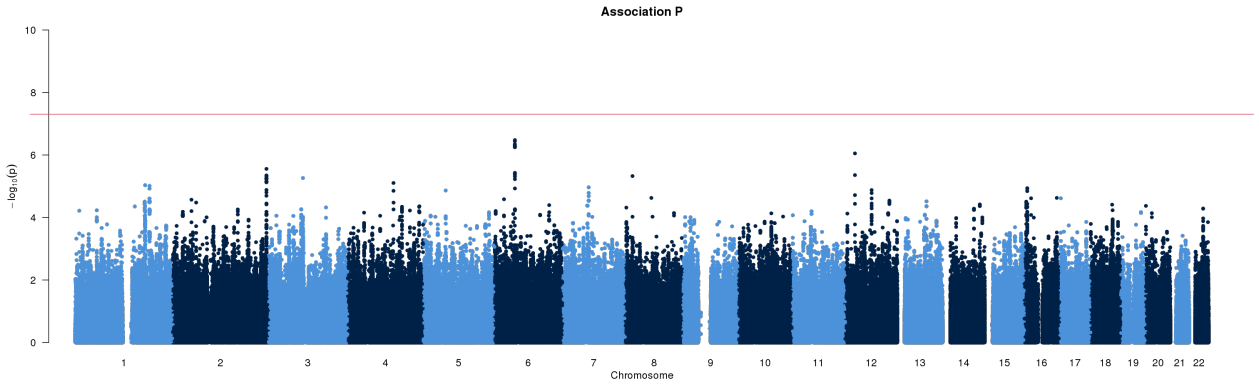


Figure 1.9: An example genome-wide association plot, showing some marginally significant results but nothing beyond the threshold for genome-wide significance.

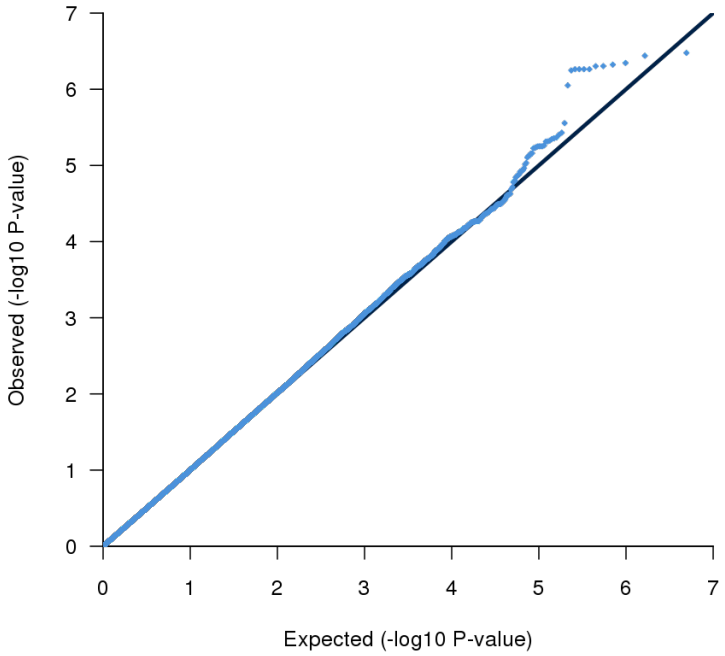


1.2.6.4 Genome-wide quantile-quantile plots

Quantile-quantile (QQ) plots plot the observed distribution versus the expected distribution under H_0 for a statistic. They typically plot the quantiles of the observed data on the y-axis and the quantiles of the expected distribution on the x-axis. The expected distribution for all markers under H_0 is a straight line from the origin with slope 1 ($-log_{10}(p_{expected}) = -log_{10}(p_{observed})$). These plots are typically used to show the extent of inflation above the expected distribution of p-values under H_0. Sometimes it is also useful to plot several QQ plots over each other, first showing the full set of markers and then showing sets of markers with those markers which have already been established to be associated with a particular trait (or those in close LD with such markers) having been excluded. These 'exclusive' QQ plots show the extent to which the distribution of p-values observed deviates from those that would be expected under H_0. Observing a distribution of points that begins near the origin and falls fairly near the expected line until the tail, where there is a significant departure above the line, represents a strong association signal with little overall inflation. Observing a line that deviates early from the expected line, having an increasing slope throughout the plot, can indicate systematic inflation of the test statistic – though it could also be an indication that there are a large number of real effects of small size which the study is underpowered to detect. Systematic departure below the ex-

pected line could indicate that the data has been overcorrected or perhaps that there is some problem with the calculations. See Figure 1.10 for an example QQ plot for a quantitative trait.

Figure 1.10: An example genome-wide quantile-quantile plot, based on the same analysis as Figure 1.9.



1.3 Meta-analysis of genome-wide association studies

Once multiple independent GWAS have been performed that each have collected data on the same phenotype, it would be advantages to combine the samples in order to increase power. One way of doing this combination would be to simply collect all genotype and phenotype information together in one combined mega-analysis in which a single statistical test is used across all samples simultanteously. Indeed, even considering additional corrections for population stratification necessitated by the multiple studies being drawn from distinct populations, this method is likely to be

the most powerful way to analyse the data (see Chapter 3 for more information). However, due primarily to concerns regarding privacy, it is in practice difficult to coordinate the exchange of this individual-level data between many groups around the world, and as an alternative a meta-analysis using summary statistics can be carried out (although see Chapter 4 for possible alternatives to summary statistics meta-analyses). Such a meta-analysis is typically based on the results of association analyses carried out within the individual studies using similar if not identical analysis plans.

Imputation to a reference population, such as the HapMap samples[20], enables GWAMA of studies that were based on different genotyping platforms, as the intersection of SNPs from different platforms could easily be too small to be considered a genome-wide representation. Imputing to a common reference population allows estimates of genotypes not present on the genotyping platforms to be made and for each study to therefore have a common set of genome-wide SNPs[67].

The analyses in each study are typically performed on the same phenotype and use the same covariates, although this depends on the study design. Whenever possible, phenotypes and covariates should be in the same units across studies, although when they are not, transformations can be used to convert into compatible units, and population-specific covariates can also be included to correct for confounding factors specific to a particular population, or to correct for population stratification within studies (see Section 1.2.5.3)[67].

Typically, data that each study will contribute to a meta-analysis for each SNP includes: an identifier for that SNP, the effect allele, the other (non-effect) allele, strand, sample size, allele frequency, effect (typically β from regression), standard error (SE), $p - value$, and software-specific metrics describing the information content on which the results were based (see Section 1.2.3.1).

However, a minimum set of data from each study would simply consist of four columns

that identify the marker, give the strength of association, the effect allele, and a sample size for weighting. The strength of association can be represented either as a Z-score (the deviation of the test statistic as the number of standard deviations above or below zero) or p-value. An example of such a data set would consist of: identifier, positive forward-strand effect allele, sample size (N), and p-value; or alternatively: identifier, effect estimate (β),N, and Z-score. Such a data set would allow for a weighted Z-score meta-analysis but would not permit any filtering to occur centrally by meta-analysis analysts—all filtering would have had to occur within the study, and it would be very difficult to do any QC at the meta-analysis stage.

One of the problems with such a minimalist approach is that when combining data from many studies prepared by different analysts, there are a lot of opportunities for errors to be introduced into the many submitted data sets. When collecting a full set of data including both alleles, strand, allele frequency, and information metrics, a number of checks can be performed to verify that the data is as expected[67]. These include: range checks on all metrics, a check that both alleles are consistent for the marker (taking into account the strand), and a check that the frequency of the effect alleles are, on average, reasonably close to the frequency of the same alleles in other studies (and/or the HapMap). In practice, the cleaning stage, where the data is submitted from each study and then systematically checked for potential issues in this manner, can yield a high percentage (in practice, in some cases it has been more than 50%) of studies in which serious issues are detected that need to be addressed before the meta-analysis can take place. The fact that errors are routinely detected is an argument in favour of collecting enough data so that these sorts of checks can be carried out, as without the ability to correct these errors, the analysis could easily be based on data that is largely incorrect.

In practice, we also routinely include effect estimates (such as β and SE) along with p-values, as that provides redundancy and allows the results from different meta-analysis techniques to be compared. The observation of significant differences in

results between a method based on p-values or Z-score (such as weighted Z-score, described in Section 1.3.1.2) and those from a method based on β and SE (such as inverse-variance, described in Section 1.3.1.3) can indicate that some studies may have used different units or an incorrect transformation.

It is also important to filter SNPs that have low information content (see Section 1.2.3.1) or when they have made too few observations of a particular class of data. Under the additive model, observing at least 20 copies of the minor allele (MAC $\geq$ 20) might be a good threshold to use[47].

1.3.1 Statistical methods for meta-analysis

1.3.1.1 Fisher's combined probability test

For much of the 20th century, the most popular test used to combine p-values from independent experiments that tested the same H_0 was Fisher's combined probability test[90]. The method pools p-values from k experiments by summing the log of the p-value and multiplying by 2, which results in a χ^2 distribution with $2 \times k$ DF (Equation 1.11)[91].

$$\chi^2 = -2 \times \sum_{i}^{k} log(p_i) \tag{1.11}$$

This is only valid when combining tests that test the same H_0, so if the direction of effect is important, Fisher suggests using a one-sided test to calculate a single-tailedp-value based on an effect in one direction, such that the p-value for effects in the opposite direction will always be $p > 0.5$ and therefore, after taking into account the additional DF added by the combination formula, the χ^2 should result in a higher p-value when adding in terms in which effect directions are not aligned[92].

1.3.1.2 Weighted Z-score

A weighted Z-score (WZ) meta-analysis[93] is based on Z-scores, which can be reported directly or calculated from the p-value and effect direction along with a weighting term (Equation 1.13) which is the square root of the ratio of the sample size in the i^{th} study to the total sample size across all studies, and sums the product of the Z-score and the weight for each study to yield the overall meta-analysis Z-score (Equation 1.12)[67]. That Z-score can then be converted to a p-value using a normal approximation.

$$z_{meta} = \sum_i (z_i \times w_i) \tag{1.12}$$

$$w_i = \sqrt{\frac{N_i}{N_{total}}} \tag{1.13}$$

This method is perhaps the most straightforward and works well if only p-values (or Z-scores) are available from some or all studies, or if some studies (may) have used different units (or transformations) for their effects, because in weighted Z-score (WZ) only the significance and direction of the effect and not the effect itself is used. However, this also means the results cannot include an effect estimate (β).

1.3.1.3 Inverse-variance

Inverse-variance meta-analysis operates on the effect estimates (β and SE), with the weighting for each study proportional to the inverse of the variance within that study. The fixed effects model[94] assumes that the effect is common across all studies, while the random effects model provides a probability model that treats individual study effects as if they were sampled from a common distribution[95].

In the fixed effects model, the formula for the weight in the i^{th} study is given by Equation 1.14.

$$w_i = \frac{1}{\mathrm{SE}_i^2} \tag{1.14}$$

Calculation of the overall effect estimates is then given by the inverse-variance equations (1.15) and (1.16)[67].

$$\beta_{meta} = \frac{\sum_i (\beta_i \times w_i)}{\sum_i w_i} \tag{1.15}$$

$$\mathrm{SE}_{meta} = \sqrt{\frac{1}{\sum_i w_i}} \tag{1.16}$$

In the random effects model[96], the weighting term adds an estimate of between-study variance to the variance used to calculate the weighting for each study. A side-effect of this is that less weighting is given to more precise studies (those with less within-study variance), resulting in a more conservative pooled effect estimate than the fixed effects model[95]. The revised formula for the weight in a random-effects meta-analysis including τ^2, which represents an estimate of the between-study variance, is given by Equation 1.17[95]. In this method, the inverse-variance equations remain as in Equations 1.15 and 1.16.

$$w_i = \frac{1}{\mathrm{SE}_i^2 + \tau^2} \tag{1.17}$$

1.3.1.4 Testing the difference in effect between two analysis strata

Testing for the difference in effect between two meta-analysis results can be achieved by using Welch's T-test, which has no prerequisites on sample size or variance of the studies. The SE of the difference is calculated from the SE in each group (1 and 2) as Equation 1.18[97].

$$\mathrm{SE}_{diff} = \sqrt{\mathrm{SE}_1^2 + \mathrm{SE}_2^2} \tag{1.18}$$

And then the T-test statistic formula is calculated from the SE of the difference and the two effect estimate (β_1 and β_2) as Equation 1.19[97].

$$T = \frac{\beta_1 - \beta_2}{\mathrm{SE}_{diff}} \tag{1.19}$$

A normal approximation can then be used to calculate the p-value, which should be sufficient for most meta-analyses, though for small studies it can also be calculated using student's distribution, using DF calculated the using the Welch-Satterthwaite equation (1.20)[98].

$$DF = \frac{(\mathrm{SE}_1^2 + \mathrm{SE}_2^2)^2}{(\frac{(\mathrm{SE}_1^2)^2}{N_1 - 1} + \frac{(\mathrm{SE}_2^2)^2}{N_2 - 1})} \tag{1.20}$$

However, in practice we have found that the results of these tests can be somewhat deflated (as evaluated using a QQ plot), which may be due to correlation between men and women due to relatedness and/or population stratification. We used the assumption that most markers genome-wide do not have a real effect to correct the SE of the difference based on the genome-wide correlation between β (somewhat akin to GC correction). We first calculated a Spearman correlation coefficient ($r_{1,2}$) between the effect estimate for each group and then used a modified formula for the SE of the difference (1.21).

$$\mathrm{SE}_{diff\_corr} = \sqrt{\mathrm{SE}_1^2 + \mathrm{SE}_2^2 - (2 \times r_{1,2} \times \mathrm{SE}_1 \times \mathrm{SE}_2)} \tag{1.21}$$

1.3.2 Practical issues

The quality of meta-analysis results is not only dependent on the set of studies included in the analysis or the choice of statistical model, but also on a number of practical concerns that arise when combining data from multiple studies from many different sources into a single analysis.

In order to automate analysis, I developed an automated pipeline for meta-analysis. In this pipeline, input files for each meta-analysis to be performed are placed into a directory with the same name as the analysis. The input files are checked for errors and cleaned versions are generated. Each input file is then GC corrected and the lambda values logged for future reference. A driver file for the Metal meta-analysis program[99] is then generated and then metal is run to perform the meta-analysis, once using the WZ method, and once using the inverse-variance (IV) fixed-

effects method. Afterward, the output from both methods are merged into one data set.

The pipeline also includes a GNU R program for meta-analysis which implements the inverse-variance fixed effects and inverse-variance random effects meta-analysis methods and which can also generate forest plots for each marker. Typically, metal is run on the genome-wide data set and then the R script is used to confirm results and to produce plots of the meta-analysis results for top hits.

It also includes tools for automatically generating GWA plots and QQ plots for each analysis, filtering on values such as p-value, and annotating the output with chromosome, position, and nearest gene.

1.3.2.1 Pruning into independent loci

Finally, the results sorted by p-value can be pruned into independent loci based on criteria based on HapMap r^2 values, genetic distance, or genomic position. In any case, the procedure begins by taking the first marker (with the lowest p-value) and recording that marker as the lead marker of a new locus. Then, each subsequent marker of increasing p-value is taken in turn and the criteria for independent loci used to determine whether the marker belongs to an existing recorded locus (e.g. if the pairwise LD between that marker and the lead marker for any recorded locus is within the chosen r^2 threshold). If the marker does belong to an existing locus, the rank of that marker at the previously defined locus is recorded. If it does not belong to an existing locus, the marker is recorded as the lead marker for a new locus. This procedure continues until all markers have been processed and assigned to loci. The output of this procedure is a list of markers with two data columns: the lead marker of the locus that this marker belongs to and the rank of this marker at the locus.

Another component of the pipeline can test for heterogeneity between effect estimates in the output of two separate meta-analyses. For example, separate meta-analysis can

be performed on men and women and the results tested for heterogeneity to test the extent of gender-specific effects.

1.3.3 Significance levels in GWAS meta-analysis

The significance of a GWAS meta-analysis needs to be corrected for the multiple tests that were performed (across all SNPs). Several methods for this correction exist, including Bonferroni correction, use of a genome-wide significance (GWS) threshold, or false-discovery rate (FDR) control.

1.3.3.1 Bonferroni correction

Bonferroni correction treats all tests as if they were completely independent, so it is likely to overcorrect an analysis of a genome-wide set of markers with LD between them. For this reason, Bonferroni correction is not typically used in GWAMA.

1.3.3.2 Genome-wide significance

The GWS level for p-values are generally accepted to be somewhere in the range from 5×10^{-7} and 5×10^{-8}[45, 100]. These thresholds are meant to roughly correct for multiple testing of a genome-wide set of markers that are, due to LD, representing an unknown number of real signals. At a significance level of 5%, a GWS threshold of 5×10^{-7} is equivalent to a Bonferroni correction of 100,000 independent tests and a threshold of 5×10^{-8} is equivalent to a Bonferroni correction of 1,000,000 independent tests. Use of a GWS threshold to determine would be less conservative than full Bonferroni correction when a larger number of markers than that are being tested, but the argument for genome-wide thresholds is that simply correcting for the full set of tests performed will tend to pick up the same real signals many times over because of the high degree of LD between the markers being tested, and there is evidence that the effective number of independent tests in a GWAS is within this range[101]. The

Wellcome Trust Case-Control Consortium (WTCCC)[45] used a threshold of 5×10^{-7} rather than what would have been their Bonferroni corrected 5% level of 1×10^{-7} and while it is possible that this threshold could also be on the conservative side in some instances, thus far no findings from GWAS that have exceeded the 5×10^{-8} significance level have been shown to be false positives[102].

1.3.3.3 False-discovery rate control

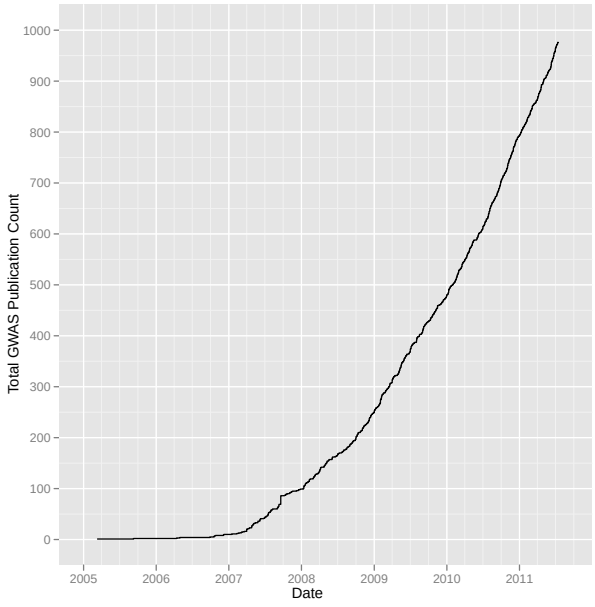
Rather than using a p-value threshold based on the number of actual or effective independent tests that were performed, methods based on the distribution of p-values can be used to control for the false-discovery rate (FDR)[103–106].

Chapter 2

Results of genome-wide association meta-analyses

Over the past six years since the earliest GWAS were conducted in 2005, the number of studies performed has increased exponentially (see Figure 2.1), likely due to a decrease in the cost of genotyping using whole-genome arrays and the success of early studies in reproducibly identifying associations between common genetic variants and a variety of diseases and traits[1].

Figure 2.1: Total GWAS Publications listed in the NHGRI GWAS catalog[107], March 2005 - August 2011.



In the first three years of GWAS combined, only ≈ 100 studies were published, while in the following three years a further ≈ 700 were published, raising the total to ≈ 800 at the beginning of 2011 and continuing to rise at a rate of $\approx 300/year$ to date. As of September 2011, the National Human Genome Research Institute (NHGRI) GWAS catalog[107] lists 976 studies that have assayed at least 100,000 SNPs in the initial stage.

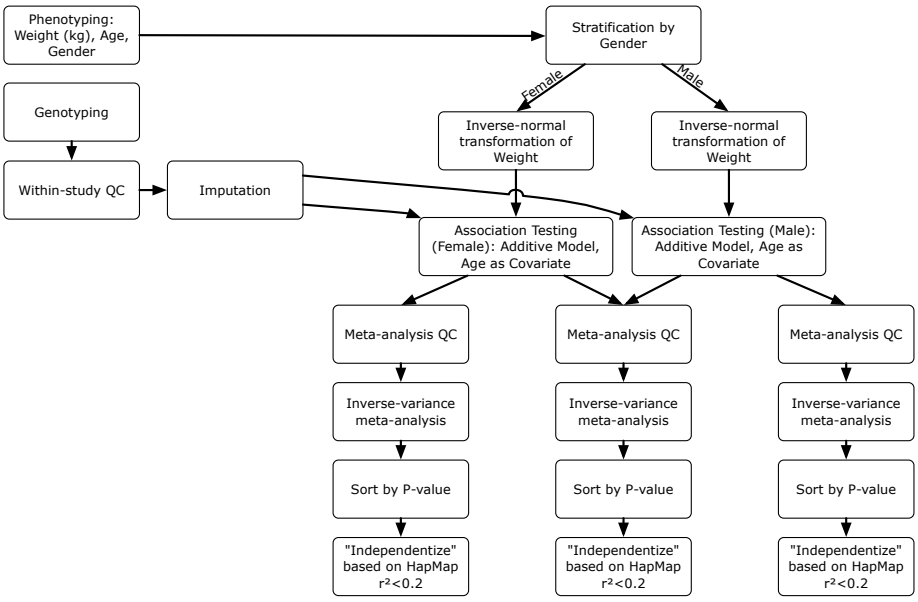
The result of this is that there are a large number of samples that have been genotyped on large-scale arrays, and after the primary trait analyses have been completed, many of these studies are left with a set of quantitative phenotypes that had been collected (in some cases for use as covariates in the primary disease/trait analysis), but that have not yet been analysed for genetic association. However, a few of the larger studies did perform secondary analyses of some of these covariates in order to detect associations between the quantitative traits and genetic loci, yielding the first insights into common genetic variation that contributes to anthropometric traits such as BMI[57, 108] (a common covariate used in many disease GWAS such as those investigating T2D).

The Genome-wide Investigation of ANThropemtric measures (GIANT) consortium (`http://www.broadinstitute.org/collaboration/giant/`) was formed in order to capitalize on the many extant studies that have collected anthropometric trait phenotypes and have genome-wide genotype data available. Under the hypothesis that common genetic variants of modest effect size exist for these traits but that previous studies have been underpowered to detect them, the expectation of GIANT is that the increased sample size made possible by meta-analysis would increase power sufficiently to detect common variants of modest effect sizes.

The effort required to perform these secondary analyses within each study is reduced from that of a primary analysis, largely because most, if not all, of the QC steps performed in the primary analysis can be directly applied to the secondary analysis without having to be re-run. In particular, SNP and sample exclusion lists generated

for the primary analysis can usually be reused for the secondary analysis, and if the primary analysis had detected population stratification and determined a method to control for it (such as using PCA PCs as covariates), the parameters of those methods (such as the PCs) can also be reused. This typically results in few (in any) manual steps that need to be performed by analysts for these analyses, so they can be easily automated (such as by using the pipeline described in Section 1.2.5.5). The typical analysis workflow of a meta-analysis within GIANT is shown in Figure 2.2.

Figure 2.2: Flowchart of a typical meta-analysis pipeline for the GIANT consortium.



2.1 Meta-analysis of waist traits, 2008-2009

This analysis of waist circumference (WC) and $\frac{waist\ circumference}{hip\ circumference} ratio$ (WHR) was undertaken by the GIANT consortium in 2008-2009, and was originally published in Lindgren et al. [3].

2.1.1 Introduction

Obesity is an increasing public health issue, but not all forms of obesity carry the same risk. Those who possess a high degree of central obesity have increased risk of adverse metabolic and cardiovascular outcomes, including T2D and atherosclerotic heart disease[109], relative to those with fat more evenly distributed or primarily deposited elsewhere in the body. Because of this, measures of overall obesity such as BMI do not represent the risk of central obesity as well as more direct measures of the trait such as WC or WHR[110, 111].

Measures of overall obesity and measures of central obesity are highly-correlated (BMI has $r^2 \approx 0.9$ with WC and $r^2 \approx 0.6$ with WHR), but central adiposity measures are themselves highly heritable[112], and that heritability remains high even after correcting for BMI ($\approx 60\%$ for WC and $\approx 45\%$ for WHR)[113]. Further, the existence of monogenic syndromes (partial lipodystrophies) which dramatically effect the formation and maintenance of specific regional fat depots indicate that at least some genetic variants can act on central obesity independently of overall obesity.

Efforts to identify variants influencing BMI and risk of overall obesity have emphasized the role of neuronal (hypothalamic) regulation of overall adiposity[4, 57, 108, 114–117] but have provided little to indicate processes responsible for individual variation in fat distribution such as those responsible for central obesity. Discovery of the mechanisms involved in the regulation of fat distribution in general is therefore important to developing an understanding of the morbidity associated with obesity.

Furthermore, due to the difficulties and complications involved in pharmacological manipulation of hypothalamic processes associated with overall obesity, the identification of genes or pathways associated specifically with central obesity may present more easily manipulated targets and present opportunities for therapeutic development to treat the highest risk forms of obesity.

2.1.2 Results & Discussion

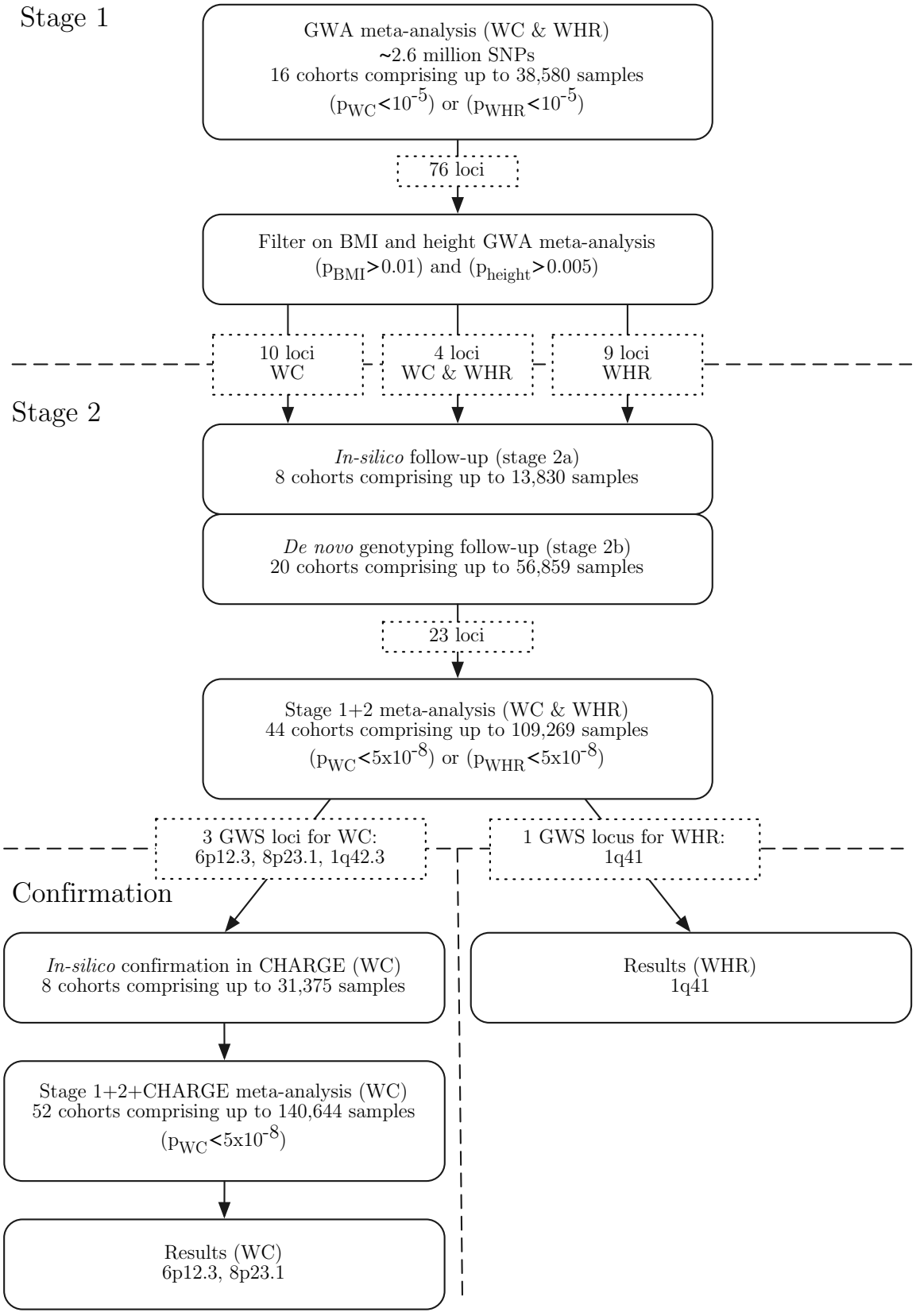
Our strategy for identifying common variants influencing central adiposity is summarized in Figure 2.3. The study was based on an initial (stage 1) GWAS meta-analysis to identify SNPs strongly-associated with measures of central adiposity.We then focused our follow-up (stage 2) efforts on the subset of those signals for which the strength of the evidence of association for measures of central adiposity appeared to be substantially stronger than that observed for overall adiposity or height. We reasoned that this subset of signals would be enriched for variants with preferential influences on central fat accumulation.

2.1.2.1 Stage 1 meta-analysis

The stage 1 meta-analysis combined data from 16 GWAS scans (N=38,580), all of European ancestry and which included measures of anthropometric phenotypes.We selected two complementary but related measures of central adiposity for analysis: WC and WHR.

Each study performed genotyping using study specific methods followed by imputation using the HapMap CEU phase 2 reference panel to yield a common set of 2,573,738 SNPs for association testing. Phenotypes were prepared for association testing by first adjusting the raw phenotype for covariates including *age* and age^2 as well as other study-specific covariates such as principal components or population identifiers, and then taking an inverse-normal transform of the residuals in order to yield a standard

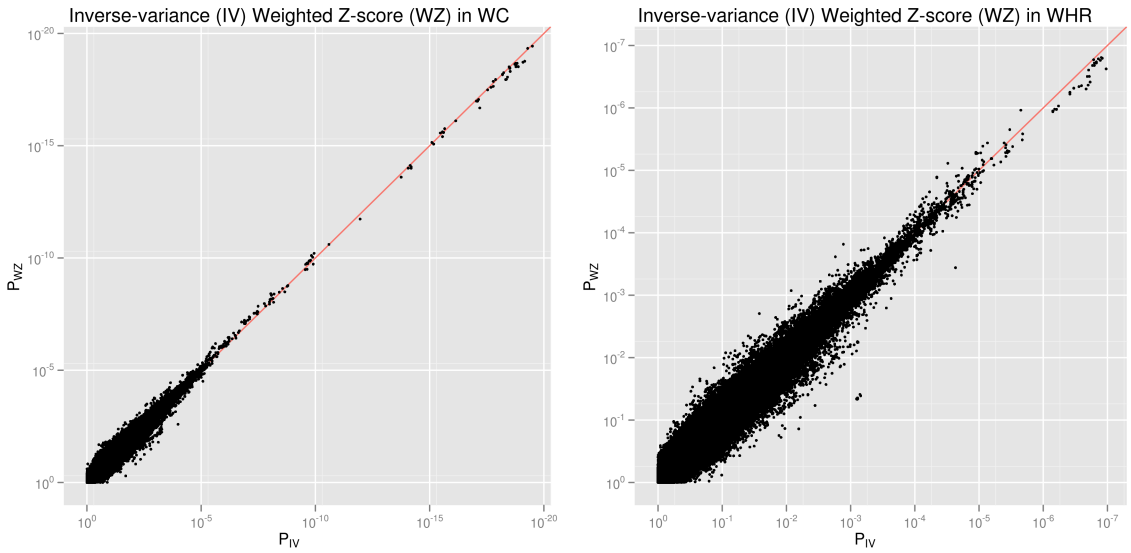
Figure 2.3: Overall diagram of analysis of waist traits.



normal distribution. Each study then performed sex-stratified association testing under an additive genetic model to test SNP genotype against the inverse-normal transformed phenotype, again using study-specific methods.

We then conducted a WZ meta-analysis (Section 1.3.1.2) based on p-value (p) and sample size (N) reported by each contributing study for each SNP in each sex. In addition, we also performed an IV meta-analysis (Section 1.3.1.3) based on the effect estimate (β) and standard error (SE) reported by each contributing study and compared the resulting p-values. The results of the two meta-analyses were highly concordant for both WC (Figure 2.4a) and WHR (Figure 2.4b).

Figure 2.4: Comparison of genome-wide p-values from inverse-variance (IV) and weighted Z-score (WZ) methods for (a) WC and (b) WHR. Both the x- and y-axes represent $-log_{10}(p)$.



(a) **(b)**

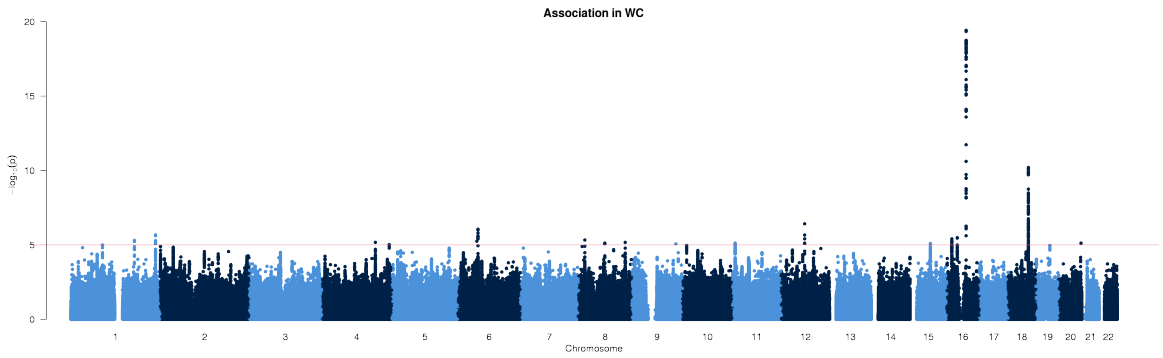
We used the WZ meta-analysis to select SNPs for follow-up genotyping, so we report p-values based on that analysis but have also reported effect estimates based on the IV meta-analysis. We observed a deviation from the expected p-value distribution under H_0 for both WC and WHR (Figure 2.6).

For each study, before performing the meta-analysis, we corrected for inflation of

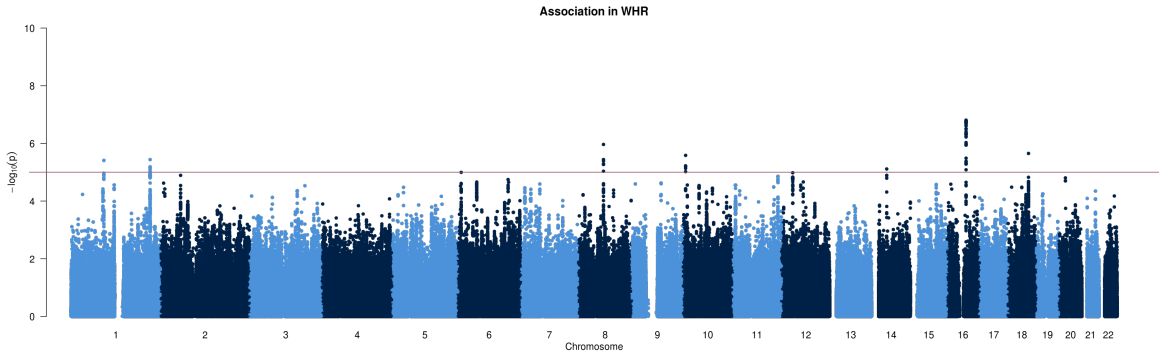
both p and SE (e.g. due to population stratification within the study) using genomic control methods (Section 1.2.5.2). We again corrected for inflation of p and SE that resulted from the meta-analyses (e.g. due to population stratification across studies) using GC correction. Values of the inflation factor for the overall meta-analysis, standardized for comparative purposes to a sample size of 1000 (λ_{1000}) (Section 1.2.5.2 and Equation 1.10), ranged from 1.002 (WHR) to 1.003 (WC).

From these data, we identified a set of 76 loci that showed evidence of association with one or both traits (Figure 2.5). We selected one SNP per independent region of association, based on a threshold of $p \leq 10^{-5}$ in preliminary pre-GC corrected analyses.

Figure 2.5: Genome-wide association plots in stage 1 for (a) WC and (b) WHR. The x-axis represents genomic position and the y-axis represents $-log_{10}(p)$.



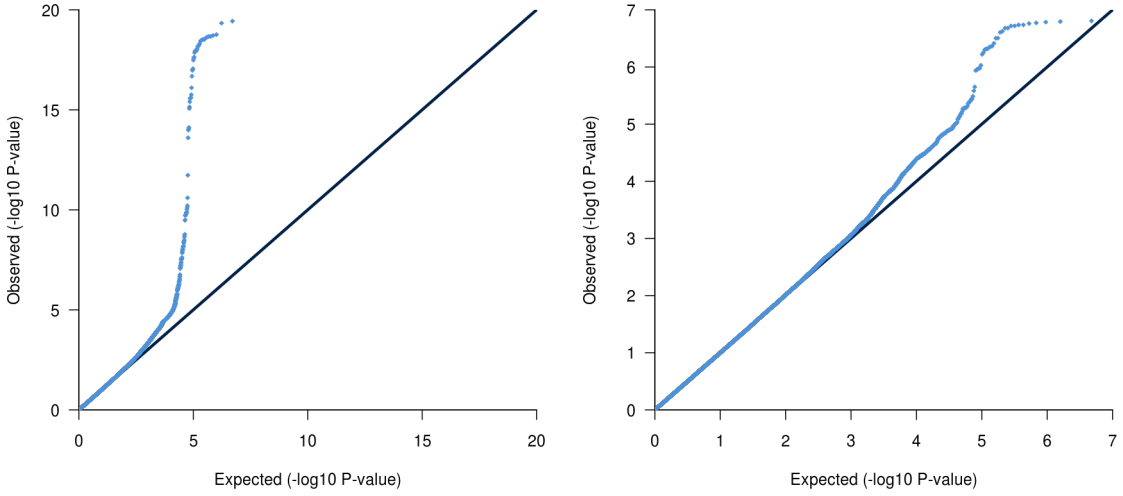
(a)



(b)

In the meta-analysis of WC, we strongly confirmed fat mass and obesity associated (*FTO*) (rs1421085: $p_{\mathrm{WC}} = 3.7 \times 10^{-20}$) and melanocortin 4 receptor (*MC4R*)

Figure 2.6: QQ plots of genome-wide association in stage 1 for (a) WC and (b) WHR. The x-axis represents expected $-log_{10}(p)$ under H_0 while the y-axis represents the observed $-log_{10}(p)$.



(a) **(b)**

(rs17700144: WC, $p = 6.2 \times 10^{-11}$) as obesity loci with an effect on WC, but did not find any other genome-wide significant ($p \leq 5 \times 10^{-8}$) loci in stage 1. Variants at *FTO* and *MC4R* are common variants with two of the largest effect sizes reported for both BMI and risk of obesity[4, 57, 108, 114, 116].

2.1.2.2 In silico and de novo follow-up

From this initial set of 76 strongly-associated signals, we sought to enrich for variants with specific impacts on central adiposity, by identifying a subset of 23 SNPs for which there was the greatest evidence for a disproportionate effect on central adiposity, as opposed to overall obesity or height. These variants were selected as having strong associations ($p \leq 10^{-5}$) with WC and/or WHR, while displaying weak evidence of an association with overall adiposity (BMI, $p \geq 0.01$) or adult height ($p \geq 0.005$) in separate stage 1 GWAS meta-analysis results performed on a largely overlapping set of studies[5, 71].

For these 23 SNPs, we obtained *in silico* follow-up data from another 8 studies with GWAS data (stage 2a) comprising a maximum of 13,830 subjects, all of European-ancestry. We also performed *de novo* genotyping in 20 additional studies (Stage 2b) comprising a maximum of 56,859 subjects, also all of European-ancestry. Follow-up analyses were restricted to the same phenotype(s) (WC and/or WHR) for which the SNP had been selected in stage 1, resulting in a total of 30 SNP-phenotype combinations under test in stage 2.

Combining stage 1 and 2 studies into a combined meta-analysis with a maximum of 109,269 subjects, we identified three signals reaching genome-wide levels of significance ($p < 5 \times 10^{-8}$): *6p12* (rs987237: $p = 4.5 \times 10^{-9}$), *8p23.1* (rs545854: $p = 1.2 \times 10^{-8}$), and *1q42.3* (rs6429082: $p = 2.6 \times 10^{-8}$). These three loci were selected for further examination in results of the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium. Two additional loci showed strong but less compelling evidence for association ($p < 5 \times 10^{-7}$) with waist phenotypes. These included SNPs mapping near genes encoding high mobility group AT-hook 2 (*HMGA2)* on chromosome *12q14.3* (rs7970350: $p_{\mathrm{WC}} = 2.2 \times 10^{-7}$) and platelet-derived growth factor receptor-like (*PDGFRL)* on chromosome *8p22-p21.3* (rs2245667: $p_{\mathrm{WC}} = 3.1 \times 10^{-7}$). The role of these genes on central obesity phenotypes will require further study.

2.1.2.3 Sex-specific association analyses

Given the clear sexual dimorphism of central obesity, as well as evidence that some genetic effects on fat distribution may be gender-specific[118], we performed additional meta-analysis of stage 1+2 data, this time stratifying the meta-analysis by sex in order to detect possible effects that were specific to one sex or the other. These analyses revealed a single additional locus of interest on *1q41* (rs2605100: $p_{women} = 1.3 \times 10^{-8}$) (Table 2.1).

2.1.2.4 Confirmation in CHARGE consortium GWAS data

As a final stage of confirmation, we analysed genotype data for rs987237, rs545854 and rs6429082 made available to us by the CHARGE consortium, members of which had recently completed a GWAS meta-analysis of WC in non-overlapping samples from 31,375 individuals. As the CHARGE GWAS analyses were restricted to WC, we were unable to follow-up the WHR signal at *1q41* in these data (Table 2.1).

At *6p12.3*, CHARGE analyses revealed directionally-consistent association with WC (rs987237: $N = 31,372$, $p_{\text{CHARGE}} = 3.6 \times 10^{-4}$) resulting in an overall combined p-value of 1.9×10^{-11} ($N = 118,691$). At *8p23.1*, genotypes for rs545854 could only be imputed in a subset ($N = 8,097$) of CHARGE samples, but the effect in CHARGE was directionally-consistent ($p = 0.28$), and in the overall results ($N = 80,210$) for this SNP, the evidence for association with WC was improved ($p = 8.9 \times 10^{-9}$) (Table 2.1).

In contrast, rs6429082 at *1q42.3* showed no evidence of association with WC in CHARGE ($N = 31,373$, $p = 0.12$). Since analysis of the combined data set no longer reached genome-wide significance ($p = 2.9 \times 10^{-7}$), further studies will be required to investigate association at this locus.

Locus	SNP	EA	EAF	Phenotype	Sex	Stage	N	β	SE	p-value
6p12	rs987237	G	16.4%	WC	Both	Stage 1	38635	0.038	0.010	1.10×10^{-04}
						Stage 2a	12369	0.019	0.017	0.26
						Stage 2b	43016	0.037	0.009	2.22×10^{-05}
						Stage 1+2	94021	0.035	0.006	4.54×10^{-09}
						CHARGE	31372	–	–	3.64×10^{-04}
						Overall	118691	–	–	1.87×10^{-11}
8p23.1	rs7826222	G	18.3%	WC	Both	Stage 1	36865	0.045	0.011	1.32×10^{-05}
						Stage 2a	3406	0.023	0.033	0.46
						Stage 2b	31841	0.036	0.011	5.31×10^{-04}
						Stage 1+2	72113	0.040	0.007	1.20×10^{-08}
						CHARGE	8097	–	–	0.28
						Overall	80210	–	–	5.89×10^{-09}
1q41	rs2605100	G	69.2%	WHR	Women	Stage 1	21397	0.062	0.011	1.30×10^{-08}
						Stage 2a	6021	0.035	0.019	8.17×10^{-02}
						Stage 2b	20213	0.018	0.011	9.06×10^{-02}
						Stage 1+2	47633	0.040	0.007	2.55×10^{-08}
						CHARGE	–	–	–	–
						Overall	–	–	–	–

Table 2.1: The three loci with genome-wide significant evidence for association with WC or WHR.

2.1.2.5 Disentangling effects on overall and central adiposity

This study was designed to be complementary to equivalent analyses of overall adiposity (as measured by BMI) conducted on a largely overlapping set of samples[5]. By focusing on widely-available anthropometric proxies of central adiposity, and targeting follow-up analyses to those signals which, in the GWAS data, had the most compelling evidence for disproportionate effects on central adiposity, our aim was to enrich for variants influencing regional rather than overall obesity.

We were interested in determining the extent to which the genome-wide associations identified at *6p12.3*, *8p23.1* and *1q41* were specific to central fat accumulation as opposed to being driven by other highly-correlated anthropometric traits. To evaluate this, we used data from the stage 2 samples for which we ran additional meta-analyses of BMI, dual energy X-ray absorptiometry (DXA), and bioimpedance tratis.

For *6p12.3*, stage 2 data indicated that rs987237 showed strong associations with overall adiposity ($p_{BMI} = 7.0 \times 10^{-12}$ in stage 2 alone), which was surprising given that in stage 1 the locus had been selected based in part on a lack of strong association with BMI ($p_{\mathrm{BMI}} > 0.01$). The association with WC remained only nominally significant in stage 2 (p=0.02) after adjustment for BMI. In addition, rs987237 was weakly associated (0.15% difference per-allele, $p = 0.02$) with overall fat mass in 29,316 individuals with bioimpedance data, and also weakly associated (0.25% difference per-allele, $p = 0.02$) with DXA measures in 13,039 additional individuals. In the 7,346 individuals for which we had DXA information specifically on fat distribution, there was no apparent association with percent central fat mass ($p = 0.98$), although this analysis is likely to be underpowered.

These data suggest that the *6p12.3* signal exerts its predominant effect on fat accumulation at multiple sites, a finding consistent with the known biology of transcription factor AP-2 beta (activating enhancer binding protein 2 beta) (*TFAP2B)*, which is

the most obvious candidate gene in the locus (see Section 2.1.2.6 for a more detailed description of the role of *TFAP2B*).

In contrast, while the signal at *1q41* did show modest associations with overall obesity (stage 2, women only, $p = 1.9 \times 10^{-4}$ for BMI) and WC ($p = 0.01$), the strength of the association with WHR was greater after adjustment for BMI (stage 2, women only, $p = 4.3 \times 10^{-6}$). In the limited subset of women ($N = 7,228$) for whom measures of hip circumference (HC) were available, and in whom we observed a proportionate signal for WHR ($p = 5.2 \times 10^{-4}$), we found no association with HC ($p = 0.7$) and a directionally consistent trend of association with WC ($p = 0.06$). Whilst these data would suggest that the *1q41* signal does indeed have a specific effect on fat distribution, large-scale clinical studies would be required to substantiate this.

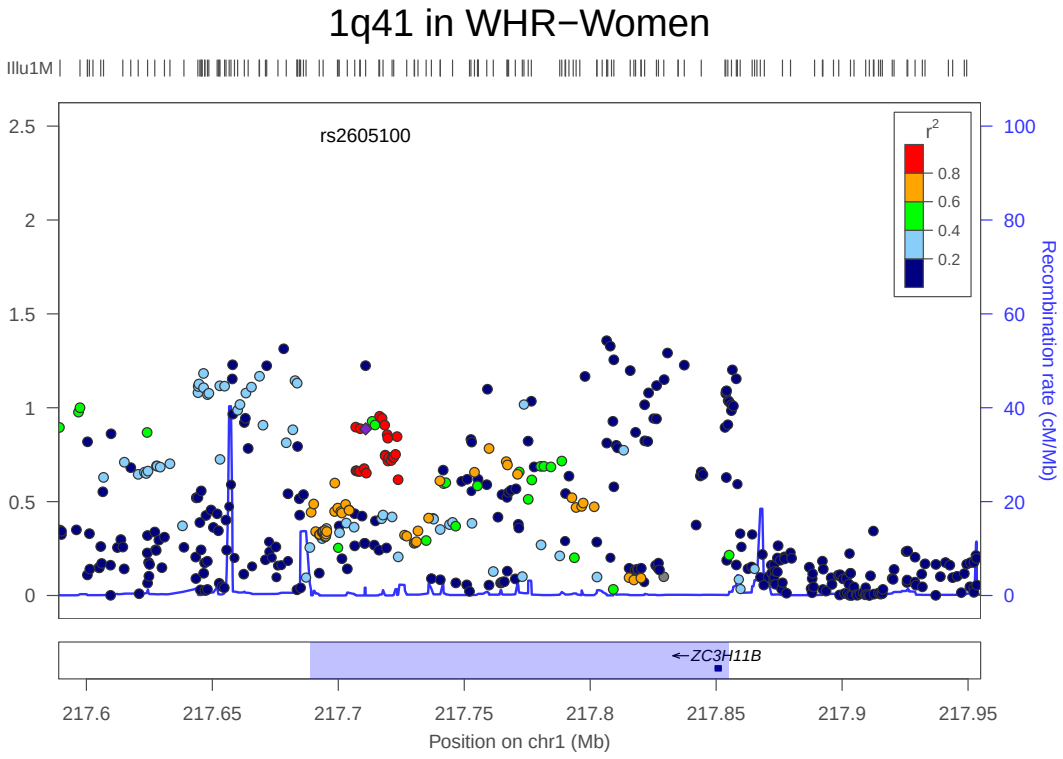
Biological connections between the *8p23.1* locus and adiposity phenotypes are unclear at this stage. The signal at *8p3.1* showed only weak association with overall adiposity (p=2.2×10^{-3} for BMI in stage 2), but the strong association with WC in stage 2 samples became non-significant after adjustment for BMI($p = 0.11$). The main proposed function of the closest gene, methionine sulfoxide reductase A (*MSRA)*, is to repair oxidative damage to proteins by enzymatic reduction of methionine sulfoxide (see Section 2.1.2.6 for more detail on this region).

2.1.2.6 Description of the three loci

The extents of each association region were determined by first taking all SNPs within a genetic distance of $1.0cM$ of the lead marker (based on HapMap phase 2[20] fine-scale recombination rate), then filtering out all SNPs with a p-value within 2 orders of magnitude of the lead marker ($p_{\text{SNP}} > p_{lead\ marker} \times 100$), and finally taking the positions of the first and last SNPs within that set to be the extents of the associated region. This yields a prediction of the associated region that is empirically based on the association signal we observed in our discovery stage data, rather than being based more crudely on HapMap genetic distance or base position criteria.

In the locus association plots, each point represents a SNP in the region, with color to indicate LD (r^2) with the lead SNP. Point position along the y-axis indicates $-log_{10}(p)$ for association (left-hand scale). Underneath the points lies the HAPMAP recombination rate (right-hand scale) traced in blue. Genes from the University of California Santa Cruz (UCSC) genome browser NCBI reference sequence database (REFSEQ) genes database[119] are shown as annotation tracks under the plot. The x-axis shows genomic position (in NCBI genome build 36 (NCBI36) coordinates), and the region highlighted in blue is the associated region for the locus. The locus plots were produced using the standalone version of the LocusZoom software[120].

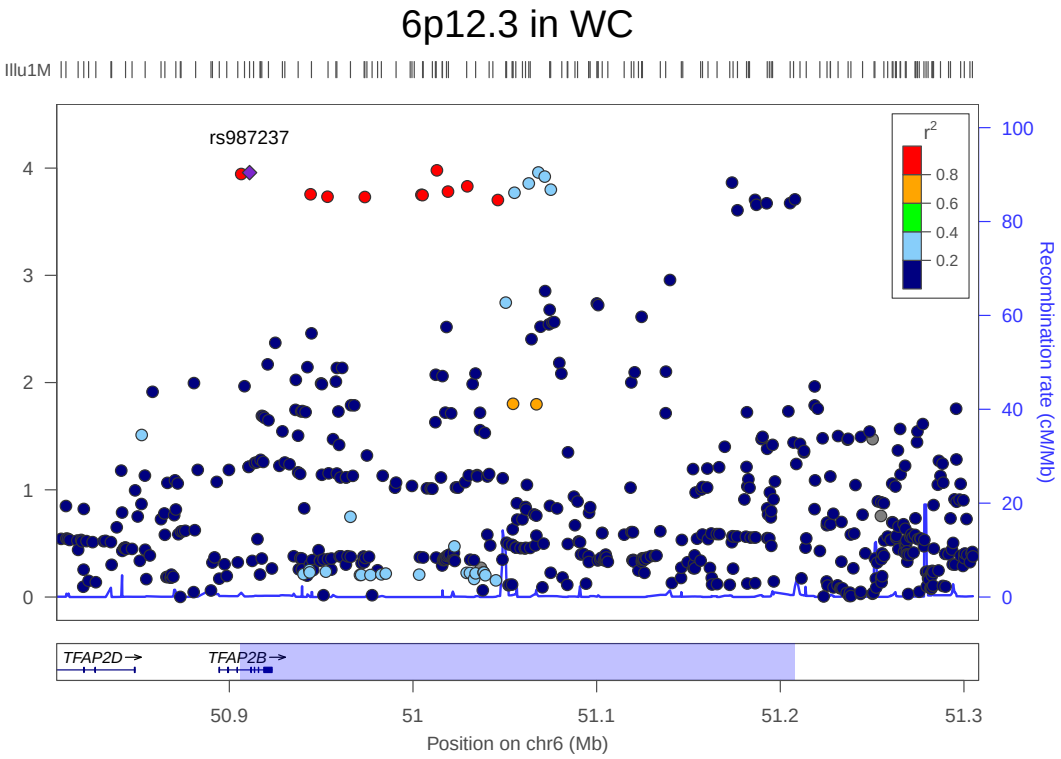
Figure 2.7: Discovery stage association at *1q41*.



1q41 The *1q41* locus is represented by lead marker rs2605100, with association signal for WHR in women extending across $\approx 166kb$ of chromosome 1, ranging from $217689kb - 217855kb$ (see Figure 2.7). One gene (*ZC3H11B*) overlaps this signal region. Zinc finger CCCH-type containing 11B pseudogene (*ZC3H11B)*

is a pseudogene with no known function. The nearest gene outside the signal region is lysophospholipase-like 1 (*LYPLAL1)*. The edge of the signal region lies 236.4kb downstream of *LYPLAL1*. *LYPLAL1* is thought to act as a triglyceride lipase that has been reported to be up-regulated in subcutaneous and visceral adipose tissue of obese subjects[121].

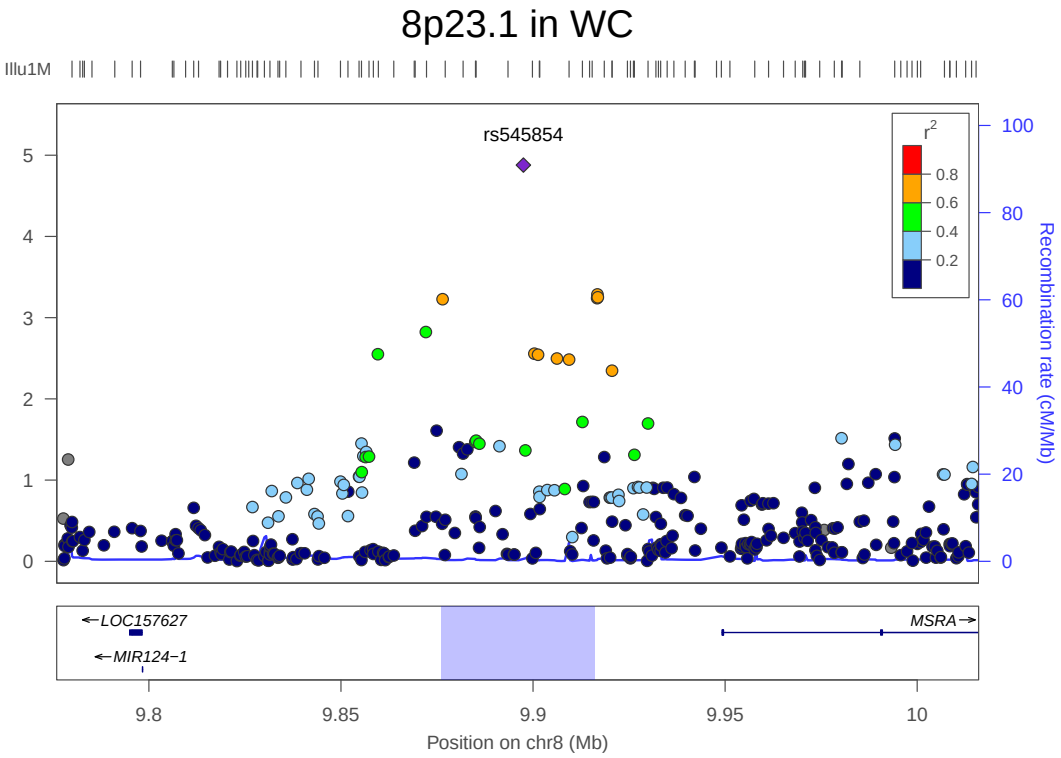
Figure 2.8: Discovery stage association at *6p12.3*.



6p12.3 The *6p12.3* locus is represented by lead marker rs987237, with association signal for WC extending across $\approx 302kb$ of chromosome 6, ranging from $50906kb - 51208kb$ (see Figure 2.8). One gene (*TFAP2B*) overlaps the signal region, as does a SNP with suggestive association with P-wave duration (rs283566: $\approx 256kb$ & $\approx 0.0627cM$ from lead marker with $r^2 = 0.12$ & $D' = 1.0$)[122]. The lead marker rs987237 is located within the third intron of *TFAP2B* and exons 4-7 of *TFAP2B* overlap the signal region. *TFAP2B* encodes a member of the AP-2 family of transcription factors and functions as both a transcrip-

tional activator and repressor of AP-2 proteins, which are thought to stimulate cell proliferation and suppress terminal differentiation of specific cell types during embryonic development[123]. An association between variants not present in HapMap (and therefore not imputed into this study) within intron 1 of *TFAP2B* and T2D has previously been reported in Japanese[124]. *TFAP2B* is reported to be preferentially expressed in adipose tissue[125], and over-expression in 3T3L1 adipocytes leads to decreased insulin sensitivity via enhanced glucose transport and increased lipid accumulation[125]. Over-expression of *TFAP2B* also down-regulates expression of the insulin-sensitizing hormone adiponectin by direct transcriptional repression[126]. Genetic variants within *TFAP2B* have recently been reported to be functional and to positively correlate with *TFAP2B* transcript levels in adipose tissue[125].

Figure 2.9: Discovery stage association at *8p23.1*.



8p23.1 The *8p23.1* locus is represented by lead marker rs545854, with associa-

tion signal for WC extending across $\approx 40kb$ of chromosome 8, ranging from $9876kb - 9916kb$ (see Figure 2.9). No known genes overlap this region, but it does lie $\approx 32kb$ upstream of *MSRA*. *MSRA* encodes an antioxidant repair enzyme that reduces oxidised methionine to methionine[123]. Moreover, the oxidation of methionine residues in proteins is considered to be an important consequence of oxidative damage to cells[127]. Oxidation of proteins by reactive oxygen species is associated with aging, oxidative stress, and many diseases and a mutant mouse that lacks the Msra gene, compared with the wild type, exhibited enhanced sensitivity to oxidative stress[128]. Excess lipid accumulation leads to increased endoplasmic reticulum (ER) activity, which ultimately can overwhelm the capacity of the ER to properly fold nascent proteins. If this process proceeds unchecked, apoptosis may result. ER stress can lead to oxidative stress in the mitochondrion, as does the presence of excess free fatty acids (FFA) which through indirect mechanisms can contribute to cellular insulin resistance[129].

2.1.2.7 Associations with other phenotypes

The accumulation of central adiposity has serious adverse health consequences including hyperlipidemia and increased risks of T2D. We examined the relationships between adiposity-related SNPs and these clinical phenotypes using available GWAS meta-analysis data. We found an association between the WHR-increasing G-allele of rs2605100 (*1q41*) and increased fasting triglycerides ($p = 3.9 \times 10^{-4}$) in data from a recent GWAS meta-analysis of 14,343 European samples[130]. This is further supported by a parallel GWAS meta-analysis effort in 19,840 samples where the G allele is similarly associated with increased triglycerides ($p = 0.02$). Using T2D case-control data from the DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) consortium[68], we found directionally-consistent but weak associations with T2D-risk for all three loci, but most obviously at *6p12.3* ($p = 0.09$).

2.1.2.8 Conclusion

By focusing on anthropometric measures of central obesity, we have identified three loci harboring common genetic variants implicated in the regulation of human adiposity and fat distribution. Determining the influence of these signals on the development and maintenance of specific fat depots would require analyses that relate genetic variation to more detailed measurements of distribution (such as DXA imaging data) in large numbers of subjects, but the loci identified appear to highlight a variety of novel mechanisms involved in the regulation of adiposity. The data are consistent with a model whereby fat mass and distribution are determined through the action of processes acting at the level of both the hypothalamus and peripheral fat depots.

2.2 Meta-analysis of WHRadjBMI

This analysis of WHR-adjusted-for-BMI (WHRadjBMI) was undertaken by the GIANT consortium in 2009-2010, and was originally published inHeid et al. [54].

2.2.1 Introduction

Central obesity and body fat distribution, as measured by WC and WHR, are associated with individual risk of T2D[131, 132] and Coronary Artery Disease (CAD)[133] as well as with mortality from all causes[109]. These effects are independent of overall adiposity as measured by BMI. WHR is of particular interest as a measure of body fat distribution because it is representative of both the adverse metabolic risk associated with increasing WC and the more protective role of gluteal fat deposition with respect to diabetes, hypertension, and dyslipidemia[134, 135].

There is evidence that body fat distribution is influenced by genetic variants distinct from those regulating BMI and overall adiposity. After accounting for BMI, individual variation in WHR remains heritable[136, 137], with estimates of heritability ranging from 22%–61%[113, 136–138]. Further, abnormalities in regional fat deposition associated with lipodystrophic syndromes demonstrate that genetic variation can have dramatic effects on the development and maintenance of specific fat depots[139, 140]. Finally, in a previous GWA analysis of waist traits (see Section 2.1), a locus on *1q41* was found to be significantly associated with WHR independent of any effects on BMI[3], providing a first example of a common genetic variant that influences control of body fat distribution distinct from that of overall adiposity.

Within the GIANT consortium, we performed a large-scale GWAMA of WHR, using an adjustment for BMI (WHRadjBMI) to bias our findings toward loci associated with body fat distribution alone rather than those that also affect overall obesity

2.2.2 Methods

We conducted a two-stage study among individuals of European descent, first examining GWA in a discovery stage and then following-up promising candidates in a second follow-up stage in additional studies consisting of non-overlapping samples.

2.2.2.1 Contributing studies

This GWAS on WHRadjBMI involved 32 discovery stage studies comprising up to 77,167 individuals (34,601 men and 42,735 women) to identify potentially interesting common genetic variants associated with central obesity. Given the sample size in the discovery stage, we calculated that we have $\approx 80\%$ power to detect SNP associations that explained as little as 0.025% of trait variance.

In the follow-up stage, we included 11 studies which extracted genotypes from *in silico* genome-wide genotype data and 18 studies with *de novo* genotyping performed specifically for the set of follow-up SNPs to provide genotypes for up to 113,636 individuals (up to 47,882 men and 66,244 women) in the follow-up stage. A description of the constituent studies of both the discovery and follow-up stages can be found in Supplementary Table 1 ofHeid et al. [54] and descriptive statistics for all studies are given in Supplementary Table 8 of Heid et al. [54]. To avoid stratification issues, studies that had originally been designed as case-control studies were stratified into separate case and control cohorts and all studies except for family-based studies were also stratified by sex.

2.2.2.2 Phenotype definition

Each study calculated BMI based on weight and height data according to Equation (2.1) and WHR based on WC and HC according to Equation (2.2).

$$\text{BMI}(kg/m^2) = \frac{weight(kg)}{height(m)^2} \tag{2.1}$$

$$\text{WHR}(cm/cm) = \frac{\text{WC}(cm)}{\text{HC}(cm)} \tag{2.2}$$

Within each strata, residuals of the WHR phenotype were then calculated by fitting a linear model with BMI, *age*, and age^2 as covariates. Some studies included additional covariates to control for population stratification or other confounding factors (see Section 1.2.5), depending on the individual study design. The resulting residuals were then inverse-normal transformed (see Section 1.2.2.2) within each strata to yield a standardised WHRadjBMI phenotype with a standard normal distribution.

2.2.2.3 Genotypes and association testing

Each discovery stage study and *in silico* follow-up study used genotypes from a genome-wide SNP genotyping array along with the HapMap phase 2[20] CEU reference panel to impute up to ≈2.85 million SNPs. The *de novo* follow-up studies performed genotyping using study-specific methods, in most cases genotyping the actual SNP selected for follow-up but in instances where genotyping that SNP was not possible, a tag SNP in high LD was used as a proxy.

In each study, the inverse-normal transformed WHRadjBMI phenotype was tested for association with SNP genotype using an additive genetic model. Methods for association testing were study-specific. Some studies used association testing methods that accounted for genotype uncertainty (see Section 1.1.5) while others used methods that used pedigree information to account for relatedness between subjects, and still others used more straightforward tests based on thresholded genotype calls. All studies provided results for each SNP that included p-value, β, SE, effect allele frequency (EAF), and N.

For comparison with other anthropometric measures, additional analyses were carried out for inverse-normal transformed BMI, WC, and HC phenotypes, and in order to get an effect estimate in the same units as the phenotype an additional analysis was performed in the follow-up studies on the WHR phenotype without inverse-normal transformation.

Details of genotyping platform, genotype calling, imputation software for each study and association testing software are given in Supplementary Table 9 ofHeid et al. [54].

2.2.2.4 Discovery stage GWAMA

In the discovery stage, up to 2,850,269 imputed and genotyped SNPs were examined in 32 GWAS comprising up to 77,167 participants and which collected anthropometric measures informative for body fat distribution (including HC, WC, height, weight, and age).

We collected association testing results separately for each study strata and performed centralised meta-analysis QC. Each data file was checked for completeness, the distribution of β estimates was compared between studies to check for outliers that could indicate phenotype transformation issues, and allele frequencies were checked against HapMap and against other studies. Analysts responsible for studies with potential issues were asked to recheck their analyses to ensure they were correct, and in many cases errors were identified and corrected. We performed SNP QC by filtering on imputation quality ($proper\_info \geq 0.4$ or $\hat{r^2} \geq 0.3$, depending on the type of information measure reported; see Section 1.2.3.1) and minor allele count (MAC > 3).

We then performed a genome-wide fixed-effects meta-analysis of WHRadjBMI using the METAL software[99], including performing a GC correction on each study. Finally, the meta-analysis results were again corrected using GC correction.

2.2.2.5 Follow-up meta-analysis

In the follow-up stage, we evaluated 16 candidate SNPs in 29 additional, independent studies including samples of European descent (comprising up to 113,636 individuals) and using a combination of *in silico* GWAS data and *de novo* genotyping. For *in silico* studies, the methods for phenotype transformation, genotyping, imputation, and association testing were the same as in the discovery stage (see Section 2.2.2.4). For *de novo* genotyped studies, no imputation was performed but within-study testing was otherwise the same as described in Section 2.2.2.4.

Likewise, the meta-analysis QC was identical, and the meta-analysis itself was again performed using the METAL software[99], but in this instance no GC correction was performed on either the study results or the overall meta-analysis results, because GC correction requires a large set of markers that are not associated with a trait and would therefore not be appropriate with a set of 16 candidate SNPs (see Section 1.2.5.2).

2.2.2.6 Testing for sex-difference

To test for differences in the effect estimate between men and women, we used the method for testing the difference in effect between two analysis strata (given in Section 1.3.1.4), using the correction for correlation between strata given in (1.21) to correct the SE estimate based on the correlation of genome-wide effect estimates between men and women. However, for follow-up stage data we did not use the correlation-based correction, as that is likely to be anticonservative when performed on a high proportion of true variants (like GC correction, this correction requires a large set of unassociated markers).

2.2.2.7 Percentage of variance explained

We computed the percentage of the variance of WHRadjBMI that is explained by a particular SNP based on the effect size of the SNP by first performing a separate analysis based on the raw (untransformed) WHR phenotype, still using BMI, age, and age^2 as covariates and otherwise using the same methods described in Section 2.2.2.4, but resulting in effect estimates in units of WHR rather than units of standard deviations. We then used the formula for percentage of variance explained given byRosner [141] and shown in Equation (2.3). We performed this analysis in one study (KORA-S3; $N = 3996$).

$$2 \times \text{EAF} \times (1 - \text{EAF}) \times (\beta^2/s^2) \tag{2.3}$$

2.2.2.8 Pathway analysis

For the pathway analysis we selected 680 SNPs from the discovery stage WHRadjBMI meta-analysis with $p < 1 \times 10^{-5}$ and that had genotype information (typed or imputed) in more than 50% of the discovery stage individuals, which resulted in 48 independent SNPs (based on HapMap CEU $r^2 < 0.2$), including the 16 loci we evaluated in the follow-up stage.

We defined a $0.2cM$ interval centered on each of the 48 SNPs based on the HapMap phase 2[20] CEU fine-scale recombination map and selected all genes that overlapped with the interval. Genes mapping to the major histocompatibility complex (MHC) region were excluded due to high density of related genes in that region. For intervals with no overlapping genes we included the nearest gene to the index SNP regardless of the distance.

This resulted in a set of 95 genes, 89 of which were found in the Protein ANalysis THrough Evolutionary Relationships (PANTHER) database of 25,431 genes. This subset of 89 genes was tested for correlation with the 240 biological processes classified

in the database[142]. For each biological process, the difference between the observed fraction of genes in the WHRadjBMI associated regions and what was expected by chance was tested using a Fisher exact test. To adjust for potential bias in the selection of genes and for testing of the many biological processes, we constructed 2,500 random sets of 48 autosomal SNPs that were matched to the original SNP set on frequency, distance to the nearest gene, and size of the nearest gene. For each random set of SNPs, we repeated the process of selecting genes and testing for correlation with all the 240 biological processes. Based on the results from the 2,500 simulated sets, we adjusted the original p-value associated with each biological process.

We calculated two separate adjusted p-values for significance of each biological process. In one (P_{adj-I}), we treated each biological process separately and adjusted based on the fraction of randomly simulated sets that produced a p-value for the same process that was equal or lower than the original p-value. In another (P_{adj-II}), we adjusted for the 240 biological processes tested by adjusting based on the fraction of randomly simulated sets that produced a p-value for any process that was equal or lower than the original p-value.

2.2.2.9 Copy number variant analyses

We examined SNPs known to be in high LD with CNVs found in samples of European descent by combining four catalogues of CNV tag SNPs (CTSs):

- 261 CTSs ($r^2 > 0.8$) generated at the Broad Institute by typing HapMap samples on the Affymetrix 6.0 array[143].
- $2,174$ multiethnic CTSs ($r^2 > 0.8$) made available by the Genomic Structural Variation (GSV) consortium and based largely on typing 450 HapMap samples on a custom-made Agilent 105k array capable of genotyping $\approx 3,320$ CNVs in CEU[144].

- 3,113 SNPs selected to tag each of the 856 CNVs in the HAPMAP phase 3[46] catalog across all HAPMAP phase 3[46] populations (where the CNV was present) with the 856 CNVs generated using the Affymetrix 6.0 and Illumina 1M arrays[46].
- 2,905 CTSs generated on a custom-made Agilent 105K array[144], but using $\approx 19,000$ samples (all of European descent, $\approx 3,000$ controls and $\approx 2,000$ cases for each of 8 diseases) typed by the WTCCC[145].

Taken together, these lists comprise a total of 6,018 CTSs for which we had WHRADJBMI discovery results in our meta-analysis, but the list was not further cropped to exclude SNPs that are tagging the same CNV.

2.2.2.10 eQTL analyses

It was our aim to identify cis-expression QTLs (eQTLs) that were coincident with the WHRADJBMI signal.

We looked for interesting cis-eQTL signals (unadjusted WHR SNP association with transcript expression $p < 1.0 \times 10^{-5}$) for transcripts that mapped within $1Mb$ of the lead WHRADJBMI SNPs and were expressed in $> 5\%$ of the samples.

For each transcript with an interesting cis-eQTL, we checked whether there was a nearby SNP with a more convincing association with expression of that transcript (transcript peak SNP). If the WHRADJBMI signal and the cis-eQTL were coincident, one would expect the p-value for the WHRADJBMI SNP association with the transcript expression to be equally small as the p-valuefor the transcript peak SNP.

If there was a transcript peak SNP different from the WHRADJBMI SNP, but in high LD ($r^2 > 0.7$) with the WHRADJBMI SNP, this was another indication for the WHRADJBMI signal and the cis-eQTL signal being coincident.

We performed mutual conditional analyses computing the association of the WHRADJBMI SNP adjusted for the transcript peak SNP and vice versa. If the association

of the transcript peak SNP disappeared ($p > 0.05$) by adjusting on the WHRadjBMI SNP, this was the final indication for the WHRadjBMI signal and the cis-eQTL signal being coincident, i.e. that the WHRadjBMI SNP signal was mediated through the respective gene and that the transcript was likely to be implicated in WHRadjBMI modulation.

Lymphocytes: Publicly available eQTL GWA data from lymphocytes fromDixon et al. [146] were utilized. Briefly, peripheral blood lymphocytes were transformed into lymphoblastoid cell lines for 206 families of European descent, totaling 830 parents and offspring. Using extracted ribonucleic acid (RNA), gene expression was assessed with the Affymetrix HG-U133 Plus 2.0 chip and the Illumina Sentrix Human Expression BeadChip. Genotyping was conducted using the Illumina Human1M Beadchip and Illumina HumanHap300K Beadchip and imputation performed based on HapMap and 1000G data. SNPs were tested for cis associations with genes within $1Mb$ of the lead SNP.

Liver, subcutaneous fat, and omental fat from the Massachusetts General Hospital in Boston, Massachusetts, USA (MGH): Liver tissue ($N = 955$), abdominal subcutaneous adipose tissue (SAT) ($N = 610$), and omental fat tissue ($N = 740$) were available from 518 caucasian bariatric surgery patients and 437 patients post mortem at MGH as described previously[147]. RNA was isolated from the tissues and gene expression was measured using a custom Agilent 44,000 microarray composed of 39,280 oligonucleotide probes. Genotyping was conducted using the Affymetrix 500K and Illumina 650K platforms followed by imputation based on HapMap phase 2[20] CEU. Each SNP was tested for association with genes within $1Mb$ of lead WHRadjBMI SNP (cis associations) using an ANOVA analysis. p-values were adjusted for the 10,000 tests performed using bonferroni correction (see Section 1.3.3.1).

Blood and subcutaneous adipose tissue from DECODE: DeCODE genetics, Reykjavik, Iceland (DECODE) eQTL analyses were performed on 23,720 transcripts in abdominal SAT ($N = 603$) and whole blood ($N = 745$) as described inEmilsson et al. [148]. SAT samples were removed through a 3 cm incision at the bikini line. Analyses were based on genotypes from the Illumina 317K or 370K chip and HAPMAP phase 2[20] based imputation, and was done by regressing the mean logarithm (log_{10}) expression ratio on the number of effect alleles per person controlling for age and sex (and for whole blood also for differential cell count) for each SNP. p-values were adjusted for multiple testing using permutation and adjusted for relatedness of the individuals using GC correction, dividing the χ^2 statistics by the λ_{GC} adjustment factors 1.063 and 1.078 for adipose tissue and blood, respectively.

2.2.3 Results

2.2.3.1 Genome-wide significance association of WHRADJBMI with 14 SNPs

In the discovery stage, the GC inflation factor of the meta-analysis results was $\lambda_{GC} = 1.09$ for the WHRADJBMI analysis, and after applying the overall GC correction, the discovery stage meta-analysis revealed a substantial excess of low p-values (see Figures 2.10 & 2.11).

From the discovery stage analysis, we selected SNPs representing the top 16 independent (defined as being located $> 1Mb$ apart and using the procedure described in Section 1.3.2.1) regions of association ($p_{discovery} < 1.4 \times 10^{-6}$) (see Table 2.2).

In these follow-up studies, 14 of the 16 SNPs analyzed showed strong directionally consistent evidence for replication ($p < 10^{-3}$) and ten SNPs reached GWS ($p < 5 \times 10^{-8}$) in follow-up studies alone. Joint analysis of the discovery and follow-up results revealed GWS associations for 14 signals ($1.9 \times 10^{-9} \geq p \geq 1.8 \times 10^{-40}$). Between-study heterogeneity was low ($I^2 < 30\%$) for all but two signals (*2q24.3* and *1q41*).

Figure 2.10: QQ plot of genome-wide association in the discovery stage for WHRadjBMI. The x-axis represents expected $-log_{10}(p)$ under H_0 while the y-axis represents the observed $-log_{10}(p)$.

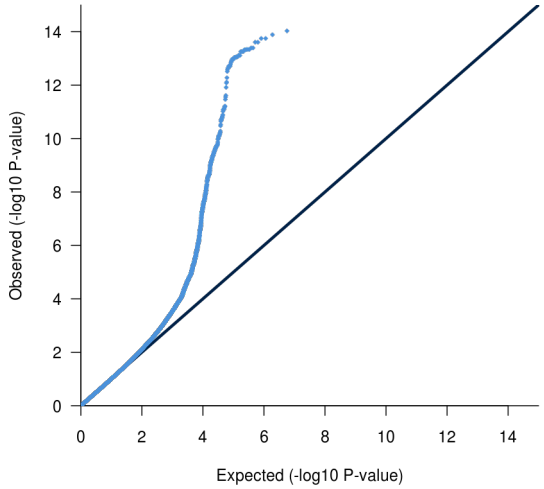
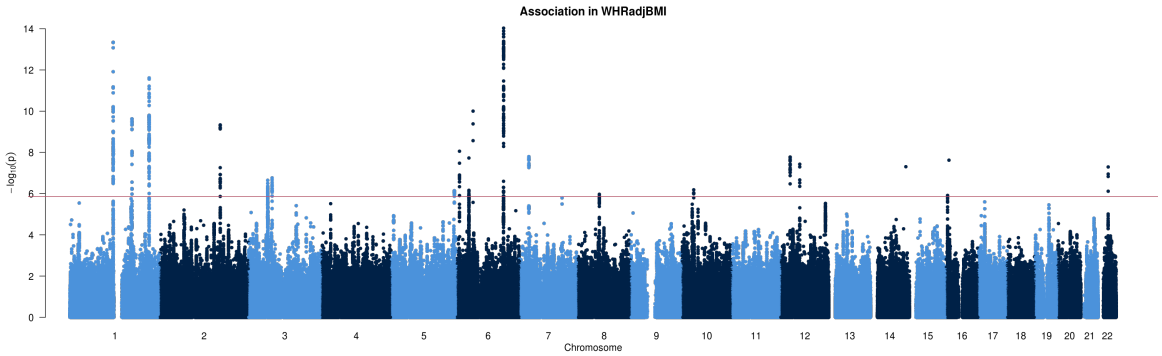


Figure 2.11: Genome-wide association plot in the discovery stage for WHRadjBMI. The x-axis represents genomic position and the y-axis is $-log_{10}(p)$. The red line at $p = 1.4 \times 10^{-6}$ indicates the threshold for bringing loci forward into the follow-up stage.



				Discovery			Follow-up			Discovery+Follow-up		
Locus	**SNP**	**EA**	**EAF**	**p**	β	**SE**	**p**	β	**SE**	**p**	β	**SE**
6q22.33	rs9491696	C	52.0%	2.10×10^{-14}	-0.037	0.005	3.26×10^{-28}	-0.045	0.004	1.83×10^{-40}	-0.042	0.004
6p21.1	rs6905288	A	56.2%	4.72×10^{-10}	0.033	0.005	1.18×10^{-16}	0.039	0.005	5.88×10^{-25}	0.036	0.005
1p12	rs984222	C	36.5%	3.81×10^{-14}	-0.037	0.005	1.56×10^{-12}	-0.031	0.004	8.69×10^{-25}	-0.034	0.004
7p15.2	rs1055144	T	21.0%	1.49×10^{-08}	0.034	0.006	3.26×10^{-18}	0.043	0.005	9.97×10^{-25}	0.040	0.005
2q24.3	rs10195252	T	59.9%	3.23×10^{-10}	0.031	0.005	3.18×10^{-16}	0.036	0.004	2.09×10^{-24}	0.033	0.004
1q41	rs4846567	T	28.3%	2.37×10^{-12}	-0.037	0.005	3.15×10^{-10}	-0.032	0.005	6.89×10^{-21}	-0.034	0.005
1q24.3	rs1011731	A	57.2%	1.72×10^{-10}	-0.030	0.005	7.47×10^{-09}	-0.026	0.004	9.51×10^{-18}	-0.028	0.004
12p11.23	rs718314	A	74.1%	2.41×10^{-08}	-0.030	0.005	1.49×10^{-10}	-0.030	0.005	1.14×10^{-17}	-0.030	0.005
6p25.1	rs1294421	T	38.7%	6.31×10^{-09}	-0.029	0.005	2.69×10^{-10}	-0.028	0.004	1.75×10^{-17}	-0.028	0.004
12q13.13	rs1443512	A	23.9%	3.33×10^{-08}	0.031	0.005	2.92×10^{-10}	0.030	0.005	6.38×10^{-17}	0.031	0.005
3p14.1	rs6795735	T	40.6%	2.47×10^{-07}	-0.025	0.005	6.75×10^{-08}	-0.026	0.005	9.79×10^{-14}	-0.025	0.005
22q12.1	rs4823006	A	56.9%	4.47×10^{-08}	0.027	0.005	2.41×10^{-05}	0.019	0.004	1.10×10^{-11}	0.023	0.004
3p21.1	rs6784615	T	94.1%	3.18×10^{-07}	0.052	0.010	1.56×10^{-04}	0.036	0.009	3.84×10^{-10}	0.043	0.009
5q35.2	rs6861681	A	34.0%	1.40×10^{-06}	0.026	0.005	2.13×10^{-04}	0.018	0.005	1.91×10^{-09}	0.022	0.005
6p21.32	rs2076529	T	57.0%	2.22×10^{-08}	-0.041	0.007	1.24×10^{-02}	-0.011	0.004	3.71×10^{-07}	-0.019	0.004
10p11.22	rs7081678	A	8.5%	5.76×10^{-07}	0.045	0.009	9.39×10^{-02}	0.013	0.008	5.57×10^{-06}	0.027	0.008

Table 2.2: The sixteen loci evaluated in the follow-up stage, sorted by combined discovery+follow-up stage p-values.

One of these SNPs, rs4846567 at *1q41*, is in moderate LD with the previously reported WHRadjBMI-associated variant near *LYPLAL1* (rs2605100: $\approx$106.5*kb* & $\approx$0.041*cM* from rs4846567 with $r^2 = 0.6440$ & $D' = 0.8350$)[3]. The remaining 13 loci were in or near genes not previously associated with WHRadjBMI or with other measures of adiposity: *1p12*, *1q24.3*, *2q24.3*, *3p14.1*, *3p21.1*, *5q35.2*, *6p21.1*, *6p25.1*, *6q22.33*, *7p15.2*, *12p11.23*, *12q13.13*, and *22q12.1* (see Section 2.2.3.10). All 14 loci together explain 1.03% of the variance in WHRadjBMI (after adjustment for BMI, age, and sex), with each locus contributing between 0.02% (*22q12.1*) to 0.14% (*6q22.33*) of the variance based on effect estimates in the follow-up stage.

2.2.3.2 Sexual dimorphism at several of the WHRadjBMI loci

Given the known sexual dimorphism of WHRadjBMI and the evidence from variance decomposition studies that this may reflect sex-specific genetic effects[118], we performed sex-specific meta-analyses for the 14 WHRadjBMI-associated SNPs. These analyses included up to 108,979 women (42,735 in the discovery stage and 66,244 in follow up) and 82,483 men (34,601 in the discovery stage and 47,882 in follow up). In a joint analysis of discovery and follow-up data, 12 of the 14 SNPs reached GWS in women, but only three SNPs reached GWS in men. At all but one locus (*1p12*), effect-size estimates were numerically greater in women. At seven of the loci (those near *6q22.33*, *6p21.1*, *2q24.3*, *1q41*, *12q13.13*, *12p11.23*, and *3p14.1*), there were marked differences in sex-specific β coefficients ($1.9 \times 10^{-3} \geq p_{sexdiff} \geq 1.2 \times 10^{-13}$). All loci displayed consistent patterns of sex-specific differences in both the discovery and follow-up studies. These 14 loci explain 1.34% of the variance in WHRadjBMI (after adjustment for BMI and age) in women but only 0.46% of the variance in WHRadjBMI in men.

2.2.3.3 Between-study heterogeneity

We found low between-study heterogeneity for 12 of the 14 SNP associations with WHRadjBMI ($I^2 < 30\%$) but moderate heterogeneity at both the *2q24.3* ($I^2 = 41\%$) and the *1q41* ($I^2 = 53\%$) loci. The heterogeneity dissipated in analyses of men-only ($I^2 = 4\%$ for *2q24.3* and $I^2 = 22\%$ for *1q41*), which is what we'd expect given that these two SNPs showed association mainly in women. However, the I^2 values only slightly reduced when restricting to women ($I^2 = 37\%$ for *2q24.3* and $I^2 = 43\%$ for *1q41*), indicating that sex does not fully explain the heterogeneity we observed. In order to investigate the source(s) of this heterogeneity, we checked whether the I^2 values differed between studies with mean age of subjects by splitting up studies into older (> 50 *years*) and younger (≤ 50 *years*) sets and evaluating the heterogeneity within each set. We also checked between studies of Northern European origin (including studies from Finland, Sweden, United Kingdom, and Estonia) versus those of Central and Southern European origin (including studies from the Netherlands, France, Germany, Austria, Croatia, and Italy), and excluding studies from the United States. We found no significant differences in I^2 values when dividing into these strata, so the source of heterogeneity in the association with WHRadjBMI at these two loci remains elusive.

2.2.3.4 Association with other anthropometric measures

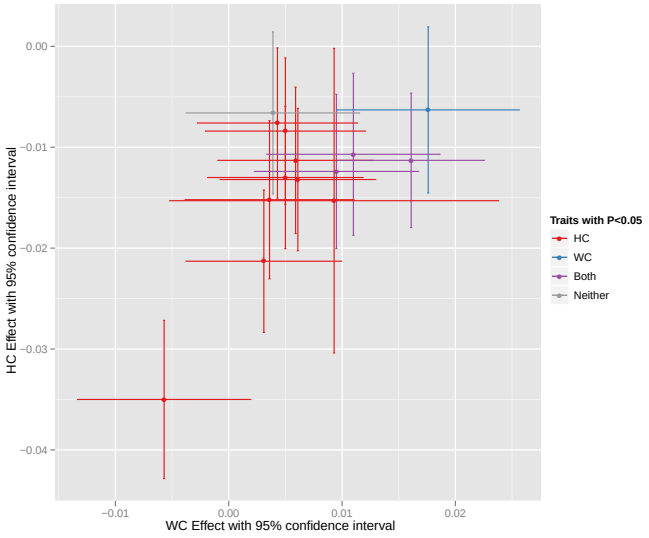
By focusing on WHRadjBMI, our goal was to detect effects on body fat distribution independent of those influencing overall adiposity. As expected, we found very little evidence that known BMI-associated variants were detected in our WHRadjBMI analysis. Of the ten loci shown to be associated with BMI in previous GWAS[5, 57, 149], only two showed nominally significant ($p < 0.05$) associations for WHRadjBMI in the discovery stage analysis (rs8050136 near *FTO*: $p = 0.03$, $N = 77,074$; and rs6548238 near *TMEM18*: $p = 3.0 \times 10^{-3}$, $N = 77,016$).

We also tested the 14 WHRadjBMI-associated SNPs for their effect on BMI using data from up to 242,530 participants available from the GIANT consortium (including most of the studies available for WHRadjBMI association as well as additional studies that did not have waist traits available). Of the 14 WHRadjBMI loci, four (near *1p12*, *5q35.2*, *1q41*, and *2q24.3*) also showed evidence of association with BMI ($4.1 \times 10^{-3} \geq p \geq 3.2 \times 10^{-6}$), with the WHRadjBMI-increasing allele associated with decreased BMI in all cases.

After adding an interaction term of SNP with BMI into the model, we observed that BMI modified the WHRadjBMI association at the *6p25.1* locus ($p_{interaction} = 9.5 \times 10^{-5}$), with a larger WHRadjBMI effect among obese individuals compared to non-obese individuals.

To determine whether the WHRadjBMI-associated signals exert their effects primarily through an effect on WC or HC, we performed meta-analyses for these specific phenotypes in the discovery and follow-up studies. Overall, we observed stronger associations for HC than for WC. β estimates were numerically greater for HC than for WC at 11 of the 14 loci, and there were nominal associations ($p < 0.05$) with HC for 12 of the WHRadjBMI-associated loci but there were only four associations with WC (see Figure 2.12).

Figure 2.12: Effect estimates in HC and WC for the fourteen loci associated with WHRadjBMI. Point estimates of effect along with 95% confidence intervals are shown with WC on the x-axis and HC on the y-axis.



In both sexes, the WHRadjBMI-associated loci with nominal association with HC always had the WHRadjBMI-increasing allele associated with reduced HC. In contrast, we observed sexual dimorphism in the pattern of WC associations. In women, the WHRadjBMI-increasing allele at all 14 loci was associated with increased WC, whereas this was only true for six of these loci in men. At *2q24.3*, for example, the WHRadjBMI-increasing allele was associated with increased WC in women ($p = 3.6 \times 10^{-4}$) but with decreased WC in men ($p = 6.8 \times 10^{3}$). These differences in the relationships between WC, HC and WHRadjBMI underlie some of the sexual dimorphism in the patterns of WHRadjBMI association.

2.2.3.5 Enrichment of association with metabolic traits

We evaluated the 14 WHRadjBMI-associated loci for their relationships with related metabolic traits using GWAS data provided by trait-specific consortia[69, 150, 151] as well as our *de novo* genotyped follow-up studies. As expected given the overlap between samples in these metabolic trait GWAS data and our WHRadjBMI GWAS data

as well as information on known trait correlations, we observed directionally consistent enrichment of associations ($p < 0.05$) between the 14 WHRadjBMI-associated alleles and increased triglyceride levels (TGs), low-density lipoprotein (LDL) cholesterol, fasting insulin (FI) and homeostasis model assessment (HOMA) derived measures of insulin resistance (binomial p-values from $3.2 \times 10^4 \geq p \geq 1.8 \times 10^{-8}$).

The WHRadjBMI-increasing allele at *2q24.3* showed strong associations with increased TGs ($p = 7.4 \times 10^{-9}$), FI ($p = 5.0 \times 10^{-6}$) and IR ($p = 1.9 \times 10^{-6}$). Eleven of the 14 WHRadjBMI-associated loci showed directionally consistent associations with T2D, and three of these loci (at *3p14.1*, *3p21.1*, and *12p11.23*) reached nominal significance ($p < 0.05$). Because the association signals for correlated traits in this analysis were likely to be overestimated given the overlap in the GWAS samples examined, we repeated these analyses and restricted the samples included to those from our de novo genotyped follow-up studies. Although this also resulted in a lower sample size, similar patterns of enrichment and directional consistency were still observed.

2.2.3.6 Pathway analysis and potential biological roles

To identify potential functional connections and pathway relationships between genes mapping at the WHRadjBMI-associated loci, we focused on the 95 genes located in a $2Mb$ interval centered around each of the 48 independent SNPs that attained $p < 10^{-5}$ in the WHRadjBMI discovery studies.

First, we performed a survey of the published literature using GRAIL[152] to search for connectivity between the genes and specific keywords that describe these functional connections. Although there was no evidence after correcting for multiple testing that the connectivity between these genes was greater than chance, we identified eight genes with nominal significance ($p < 0.05$) for potential functional connectivity: plexin D1 (*PLXND1)*, homeobox C10 (*HOXC10)*, T-box 15 (*TBX15)*, R-spondin 3 (*RSPO3)*, homeobox C4 (*HOXC4)*, homeobox C6 (*HOXC6)*, kringle containing

transmembrane protein 1 (*KREMEN1)*, and homeobox C11 (*HOXC11).* The keywords associated with these connections included "vegf," "homeobox," "patterning," "mesenchyme," "embryonic," "development," and "angiogenesis."

Additionally, we performed pathway analyses using the PANTHER database[142] based on the same set of 95 genes. This analysis generated some evidence for overrepresentation of "developmental processes" ($p = 5.8 \times 10^{-8}$) and "mRNA transcription regulation" ($p = 2.7 \times 10^{-6}$) but neither of these factors had even nominal significance after adjustment for bias (P_{adj-I}) and the number of biological processes tested (P_{adj-II}).

Finally, we examined the described functional roles of some of the most compelling candidates based on either proximity to the signal or the other analyses described in this paper (see Section 2.2.3.10 for a description of all fourteen associated regions). These analyses uncovered possible genetic roles in adipocyte development (*TBX15*), pattern formation during embryonic development (*HOXC13*), angiogenesis (*VEGFA*, *RSPO3*, and *STAB1*), Wnt and β-catenin signaling (*RSPO3* and *KREMEN1*), insulin signaling (*ADAMTS9*, *GRB14*, and *NISCH*), lipase activity (*LYPLAL1*), lipid biosynthesis (*PIGC*), and intracellular calcium signaling (*ITPR2*).

2.2.3.7 Evaluation of CNVs and non-synonymous changes

Both common and rare CNVs have been reported to be associated with overall adiposity[5, 149, 153, 154], but the impact of CNVs on fat distribution has not been evaluated previously. To examine the potential contribution of common CNVs to variation in WHRadjBMI, we looked for evidence of association in our discovery GWAMA using a set of 6,018 CNV-tagging SNPs which collectively capture $> 40\%$ of common CNVs that are greater than $1kb$ in length[144, 145].

One CNV-tagging SNP (rs1294421 at *6p25.1*) was observed among our 14 WHRadjBMI-associated loci. This SNP is in strong LD ($r^2 = 0.98$) with a $2,832bp$ du-

plication variant (CNVR2760.1)[145] located 12 kb from an expressed sequence tag (BC039678) and $\approx 87kb$ from lymphocyte antigen 86 (*LY86)* such that the duplication allele is associated with reduced WHRadjBMI. The duplicated region consists entirely of noncoding sequence but includes part of a predicted enhancer sequence (E.5552.1)[155].

To identify other putative causal variants in our associated regions, we searched for non-synonymous coding SNPs in strong LD (defined as $r^2 > 0.7$) with the most strongly associated SNPs at each locus using data from both HapMap and 1000G (April and August 2009 releases). In this search, one lead SNP (rs6784615, at the *3p21.1* locus) was correlated with non-synonymous changes in two nearby genes, dynein, axonemal, heavy chain 1 (*DNAH1)* (p.Val441Leu, p.Arg1285Trp and p.Arg3809Cys) and glycerate kinase (*GLYCTK)* (p.Leu170Val). Fine-mapping and functional studies will be required to determine whether the *DNAH1* and/or *GLYCTK* variants or the *6p25.1* CNV are causal for the WHRadjBMI associations at these loci.

2.2.3.8 Effect of WHRadjBMI associations on expression in relevant tissues

eQTL data can implicate regional transcripts that mediate trait associations, and we therefore examined the 14 WHRadjBMI-associated loci using eQTL data from human SAT[148] (two separate sample sets, $N = 610$ and $N = 603$), omental fat ($N = 740$), liver[147] ($N = 518$), blood[148] ($N = 745$), and lymphocytes[146] ($N = 830$).

At six of the loci, the WHRadjBMI-associated SNP was either the strongest SNP associated with significant ($p < 1 \times 10^{-5}$) expression of a local (within $1Mb$) gene transcript or explained the majority of the association between the most significant eQTL SNP and the gene transcript in conditional analyses ($P_{adj} > 0.05$). For example, the WHRadjBMI-associated SNP rs1011731 at *1q24.3* was strongly associated with expression of phosphatidylinositol glycan anchor biosynthesis, class C (*PIGC)* in lymphocytes ($p = 5.9 \times 10^{-10}$); furthermore, rs1011731 is in high LD ($r^2 = 1.00$,

$D' = 1.00$) with the SNP with the strongest effect on *PIGC* expression (rs991790), and this cis eQTL association was eliminated by conditioning on rs1011731. These analyses therefore indicate that these two signals are coincident and that *PIGC* is a strong candidate for mediating the WHRadjBMI association at rs1011731. We found similar evidence for coincidence of the WHRadjBMI signal with expression for rs984222 (*TBX15* in omental fat), rs1055144 (EST AA553656 in SAT), rs10195252 (*GRB14* in SAT), rs4823006 (*ZNRF3* in SAT and omental fat) and rs6784615 (*STAB1* in blood). Taken together, the overlap between trait association and gene expression at these loci suggests that the WHRadjBMI associations may be driven through altered expression of growth factor receptor-bound protein 14 (*GRB14*), *PIGC*, stabilin 1 (*STAB1*), *TBX15*, zinc and ring finger 3 (*ZNRF3*), and expressed sequence tag (EST) AA553656.

2.2.3.9 RNA expression in gluteal and abdominal fat tissue

To determine whether genes within the WHRadjBMI-associated loci showed evidence of differential transcription in distinct fat depots, we compared expression levels in gluteal or abdominal SAT in 49 individuals. We focused on the 15 genes with the strongest evidence for causal involvement (on the basis of proximity to the lead SNP and/or other biological or functional data) for which expression data were available. Five of these genes (*RSPO3*, *TBX15*, *ITPR2*, *WARS2*, and *STAB1*) were differentially expressed between the two tissues (using an F-test, controlling for FDR across the 15 genes at 5%, $Q_{FDR} < 0.05$). This supports the hypothesis that the association with WHRadjBMI reflects depot-specific differences in expression patterns, at least at some loci.

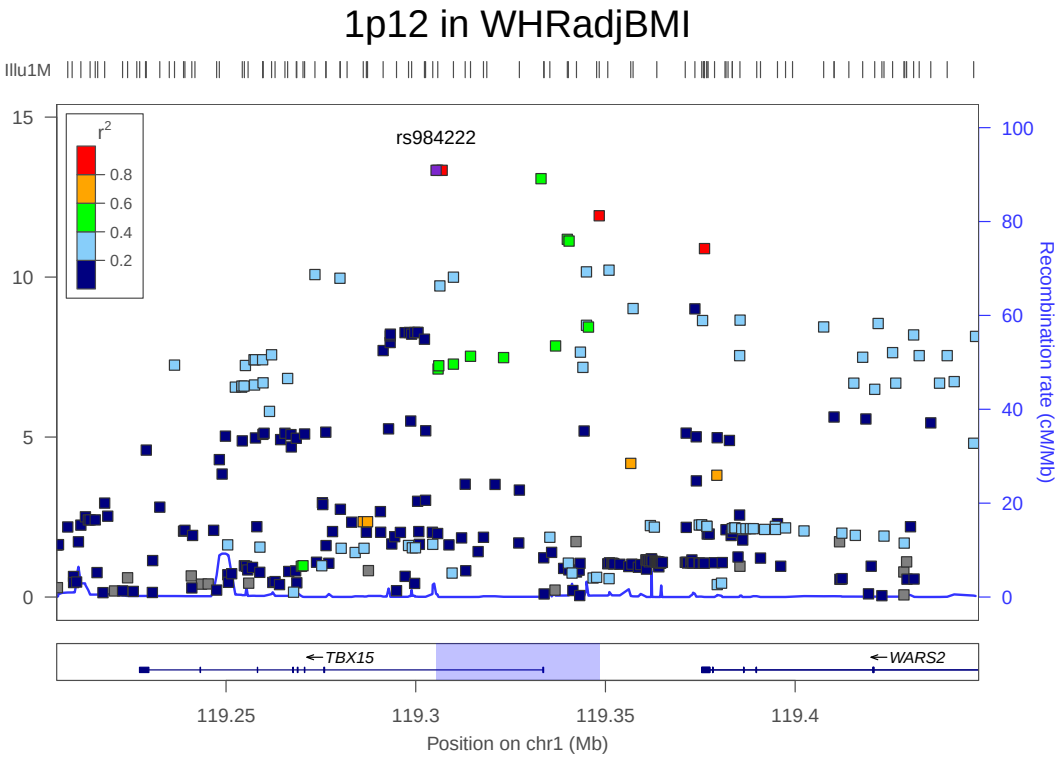
2.2.3.10 Description of the fourteen loci

The extents of each association region were determined by first taking all SNPs within a genetic distance of $1.0cM$ of the lead marker (based on HapMap phase 2[20] fine-

scale recombination rate), then filtering out all SNPs with a p-value within 2 orders of magnitude of the lead marker ($p_{\text{SNP}} > p_{lead\ marker} \times 100$), and finally taking the positions of the first and last SNPs within that set to be the extents of the associated region. This yields a prediction of the associated region that is empirically based on the association signal we observed in our discovery stage data, rather than being based more crudely on HapMap genetic distance or base position criteria.

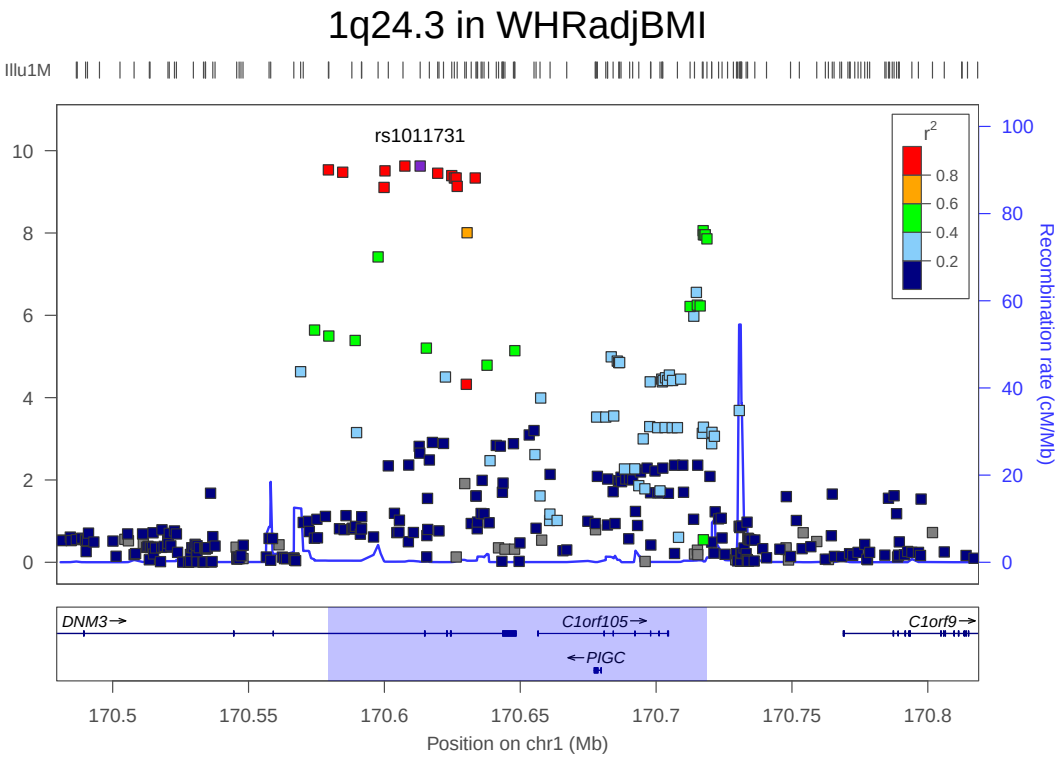
In the locus association plots, each point represents a SNP in the region, with color to indicate LD (r^2) with the lead SNP. Point position along the y-axis indicates $-log_{10}(p)$ for association (left-hand scale). Underneath the points lies the HapMap recombination rate (right-hand scale) traced in blue. Genes from the UCSC genome browser RefSeq genes database[119] are shown as annotation tracks under the plot. The x-axis shows genomic position (in NCBI36 coordinates), and the region highlighted in blue is the associated region for the locus. The locus plots were produced using the standalone version of the LocusZoom software[120].

Figure 2.13: Discovery stage association at *1p12*.



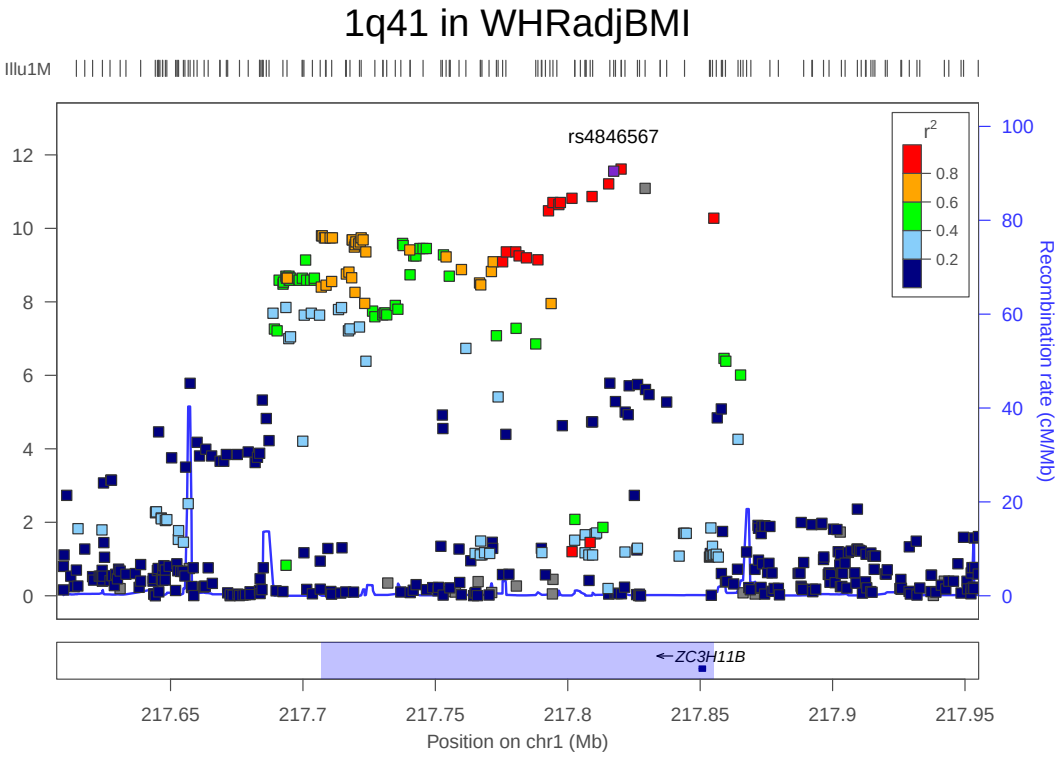
1p12 The *1p12* locus is represented by lead marker rs984222, with association signal for WHRadjBMI extending across $\approx 43kb$ of chromosome 1, ranging from $119305kb - 119348kb$ (see Figure 2.13). One gene (*TBX15*) overlaps this signal region. *TBX15* is a transcription factor reported to be up-regulated in subcutaneous fat compared to intra-adbominal adipocytes, with expression levels closely correlated with patterns of fat distribution (represented by WHR) as well as overall obesity (represented by BMI)[156]. *TBX15* may be involved in adipocyte development, embryonic development, pattern specification, and development of specific adipose depots[156]. In addition, *TBX15* has been implicated in Cousin syndrome, an autosomal recessive disorder characterized by congenital dwarfism, facial dysmorphism, and skeletal anomalies[157], and mutations in Mus musculus T-box 15 (*Tbx15)* have been found to cause significant skeletal anomalies in mouse models[158, 159].

Figure 2.14: Discovery stage association at *1q24.3*.



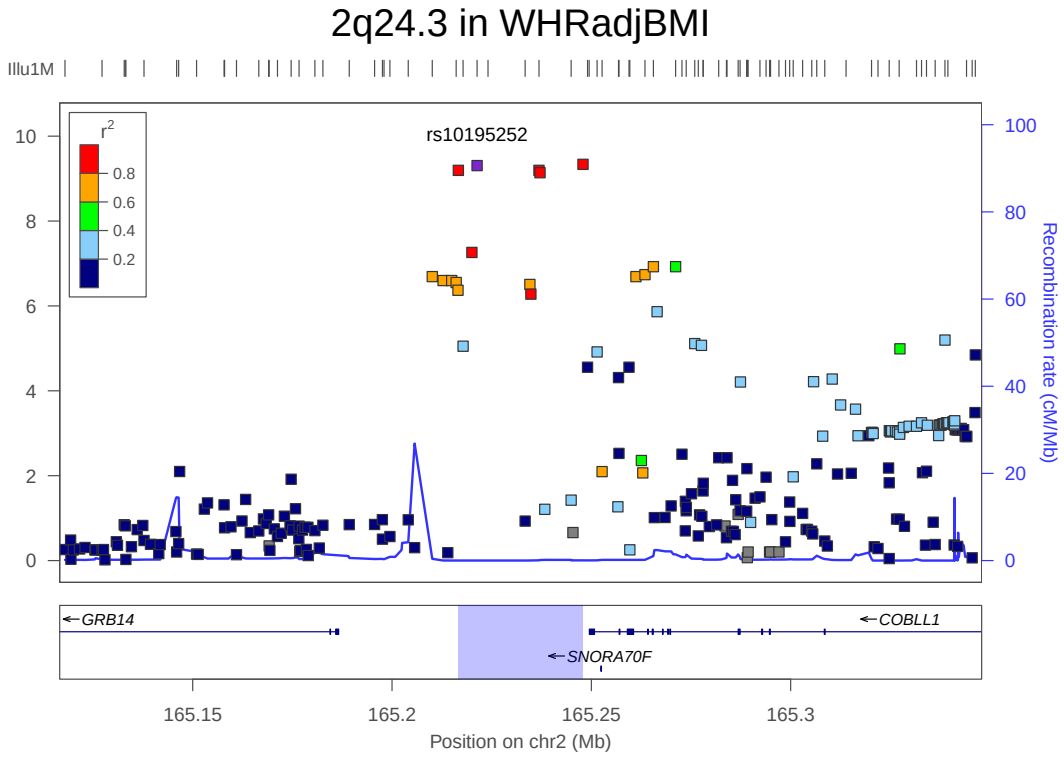
1q24.3 The *1q24.3* locus is represented by lead marker rs1011731, with association signal for WHRadjBMI extending across $\approx 139kb$ of chromosome 1, ranging from $170579kb - 170719kb$ (see Figure 2.14). Three genes (*C1orf105*, *DNM3*, and *PIGC*) overlap this signal region. The lead marker is located with intron 15 of dynamin 3 (*DNM3).* *DNM3* is a member of the dynamin family of enzymes that are important for interactions between the cell membrane and actin cytoskeleton[160]. Dominant negative mutations in transfected dynamin enzymes promote GLUT6 and GLUT8 glucose transporters to the cell surface in cultured rat adipocytes in vitro[161]. *PIGC* encodes a subunit of the enzyme that transfers N-acetylglucosamine to phosphatidylinositol, the first step of glycosylphosphatidylinositol (GPI) lipid anchor biosynthesis. GPI anchors various eukaryotic proteins to the cell membrane[162, 163]. Rare deletions of 1q24.3-q25.1 which includes the associated region are associated with severe growth deficiency, microcephaly, small hands and feet, dysmorphic face, and cognitive defects[164].

Figure 2.15: Discovery stage association at *1q41*.



1q41 The *1q41* locus is represented by lead marker rs4846567, with association signal for WHRadjBMI extending across $\approx 148kb$ of chromosome 1, ranging from $217707kb - 217855kb$ (see Figure 2.15). One gene (*ZC3H11B*) overlaps this signal region. *ZC3H11B* is a pseudogene of unknown function. The nearest genes outside the region are *LYPLAL1* (region is $\approx 293kb$ downstream) and solute carrier family 30, member 10 (*SLC30A10)* (region is $\approx 299kb$ downstream). *LYPLAL1* encodes the lysophospholipase-like 1 protein, which is thought to act as a triglyceride lipase and is reported to be up-regulated in subcutaneous adipose tissue of obese subjects[121]. *SLC30A10* belongs to a family of membrane transporters zinc efflux family (*SLC30*) involved in intracellular zinc homeostasis and is expressed in brain and liver[165]. Another member of the *SLC30* family, solute carrier family 30 (zinc transporter), member 8 (*SLC30A8)*, is associated with T2D risk[166].

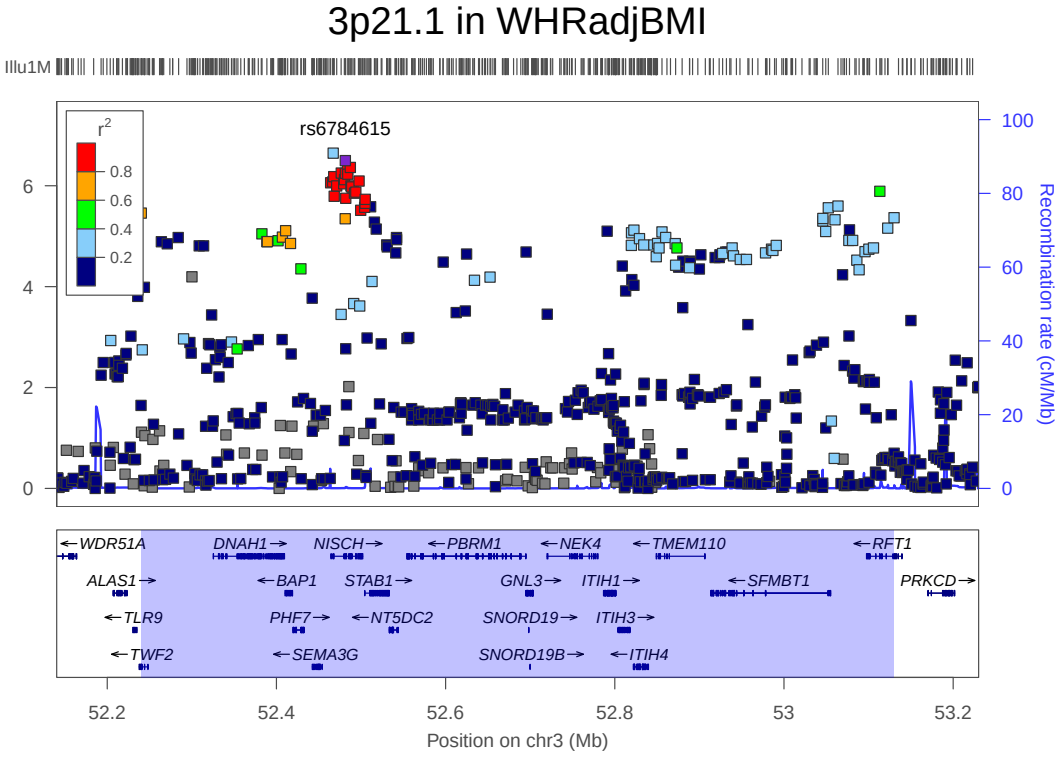
Figure 2.16: Discovery stage association at *2q24.3*.



2q24.3 The *2q24.3* locus is represented by lead marker rs10195252, with association signal for WHRadjBMI extending across $\approx 31kb$ of chromosome 2, ranging from $165217kb - 165248kb$ (see Figure 2.16). One gene (*COBLL1*) overlaps this signal region. The lead marker is located within the first intron of a non-coding splice variant of COBL-like 1 (*COBLL1)* as identified by The GENCODE Project: Encyclopædia of genes and gene variants (GENCODE)[167, 168]. The associated region is located $\approx 1.6kb$ downstream of the coding splice variant of *COBLL1*, which is thought to be involved in neural tube formation[169], and $\approx 30kb$ upstream of *GRB14*. A SNP located just outside our region but within the *COBLL1* transcript is associated with high-density lipoprotein cholesterol (HDL-C)(rs10490694: $\approx 35.5kb$ & $\approx 0.002cM$ from lead marker with $r^2 = 0.1950$ & $D' = 0.9060$)[170]. *GRB14* is a member of a family of SH2-containing adaptors and binds directly to insulin receptors[171, 172]. Interestingly, Mus mus-

culus growth factor receptor bound protein 14 (*Grb14*) deficient mice exhibit increased body weight, mainly explained by increased lean mass on normal diet[144], improved glucose homeostasis despite lower circulating insulin levels, and enhanced insulin signaling in liver and skeletal muscle[173]. *Grb14* expression is increased in adipose tissue of insulin-resistant animal models and in humans with T2D[174], indicating that *Grb14* may modulate insulin sensitivity. The WHRadjBMI signal we observe appears to be distinct from a nearby locus previously associated with smoking initiation and current smoking (rs4423615: $\approx 74.9kb$ & $\approx 0.188cM$ from lead marker with $r^2 = 0.000$ & $D' = 0.009$)[175].

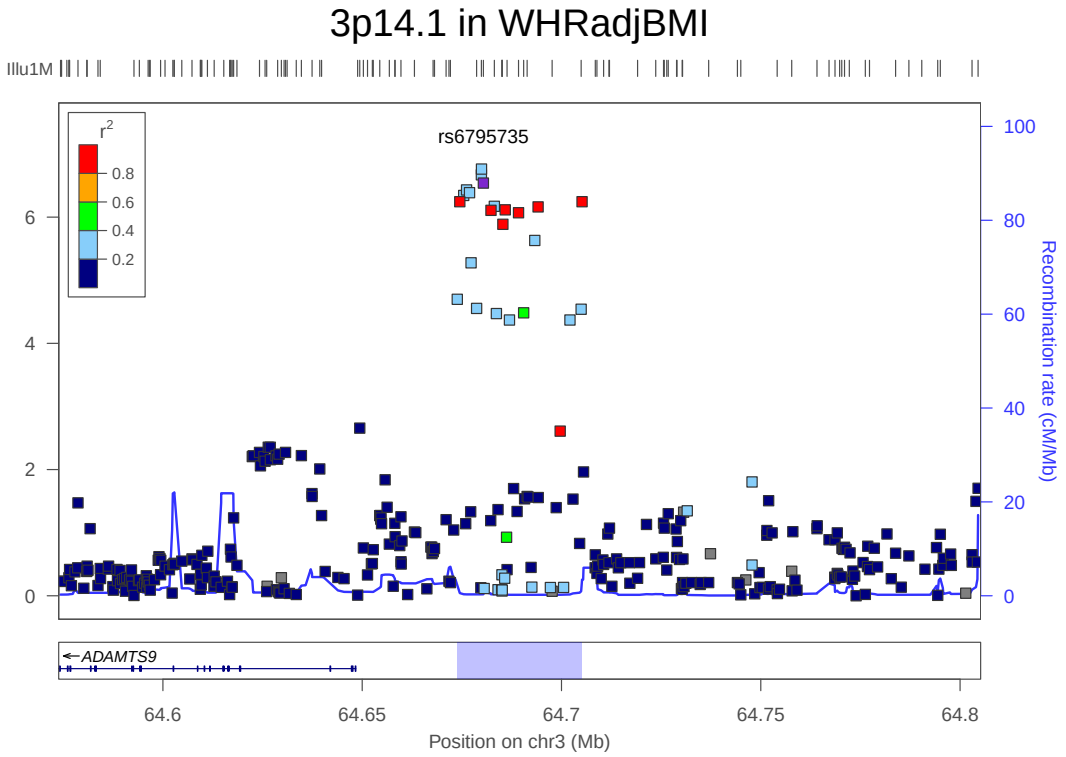
Figure 2.17: Discovery stage association at *3p21.1*.



3p21.1 The *3p21.1* locus is represented by lead marker rs6784615, with association signal for WHRadjBMI extending across $\approx 890kb$ of chromosome 3, ranging from $52240kb - 53130kb$ (see Figure 2.17). Thirty-one genes (*BAP1*, *C3orf78*, *DNAH1*, *GLT8D1*, *GLYCTK*, *GNL3*, *ITIH1*, *ITIH3*, *ITIH4*, *MIR135A1*, *MIRLET7G*, *MUSTN1*,

NEK4, *NISCH*, *NT5DC2*, *PBRM1*, *PHF7*, *PPM1M*, *RFT1*, *SEMA3G*, *SFMBT1*, *SNORD19*, *SNORD19B*, *SNORD69*, *SPCS1*, *STAB1*, *TMEM110*, *TMEM110-MUSTN1*, *TNNC1*, *TWF2*, and *WDR82*) overlap this signal region, as do SNPs associated with major mood disorders (rs2251219: $\approx 78.4kb$ & $\approx 0.028cM$ from lead marker with $r^2 = 0.006$ & $D' = 0.487$)[176], Bipolar disorder (rs1042779: $\approx 315kb$ & $\approx 0.039cM$ from lead marker with $r^2 = 0.004$ & $D' = 0.429$)[177], and height (rs2336725: $\approx 612kb$ from lead marker)[70]. The lead marker of this region is located within the 3' untranslated region (UTR) of nischarin (*NISCH)*, the gene that encodes nischarin, which interacts with insulin receptor substrate 4[178], and has also been reported to play a role in insulin receptor signaling[179]. Sema domain, immunoglobulin domain (Ig), short basic domain, secreted, (semaphorin) 3G (*SEMA3G)* is highly expressed in adipocytes[180]. Also within a cluster of highly associated SNPs in high LD ($r^2 > 0.6$) with the lead marker are SNPs within *STAB1*, a scavenger receptor that plays a role in intracellular trafficking[181] and within BRCA1 associated protein-1 (ubiquitin carboxy-terminal hydrolase) (*BAP1)*, which is thought to help regulate cell growth and proliferation[182].

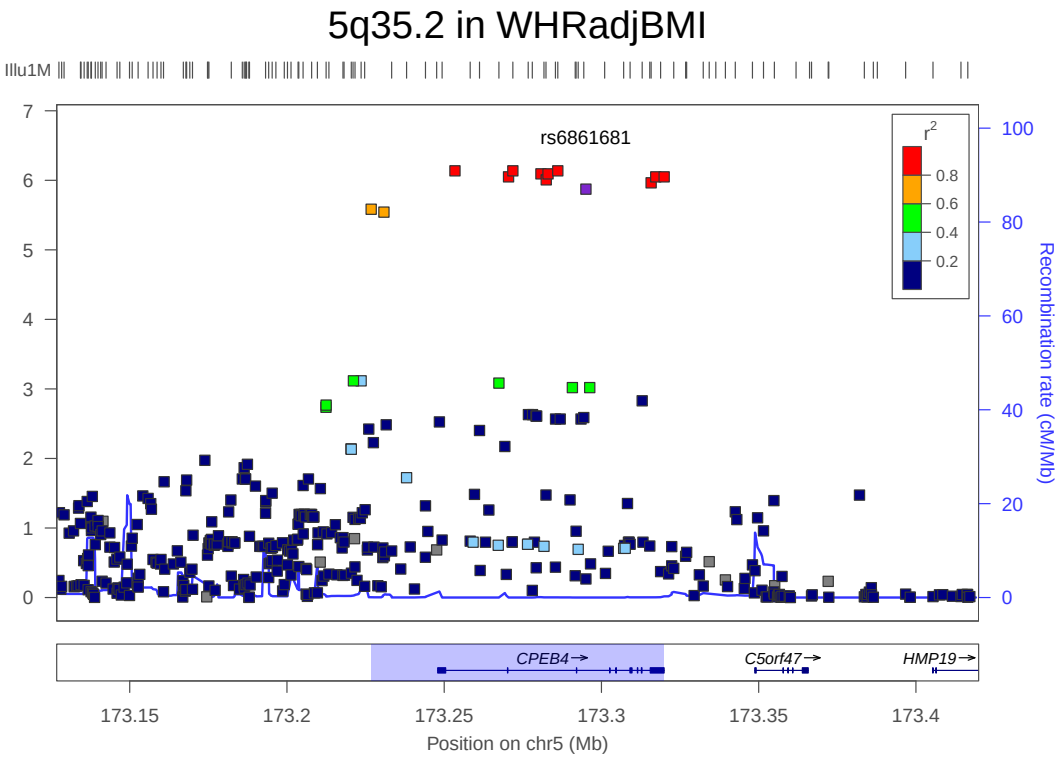
Figure 2.18: Discovery stage association at *3p14.1*.



3p14.1 The *3p14.1* locus is represented by lead marker rs6795735, with association signal for WHRadjBMI extending across $\approx 31kb$ of chromosome 3, ranging from $64674kb - 64705kb$ (see Figure 2.18). Two genes (*ADAMTS9-AS2* and *MIR548AN*) overlap this signal region, as does a previously reported SNP association with T2D (rs4607103: $\approx 6.5kb$ & $\approx 0.003cM$ from lead marker with $r^2 = 0.28$ & $D' = 1.0$)[68]. microRNA 548an (*MIR548AN)* is a microRNA (μRNA) which primarily maps to the X chromosome, but the full length μRNA precursor sequence also maps with 96.4% identity to five fragments across a $\approx 240kb$ window within our signal region with only 3 base mismatches). The function of *MIR548AN* is not known. ADAMTS9 antisense RNA 2 (non-protein coding) (*ADAMTS9-AS2)* is a long non-coding RNA transcript which is an antisense for ADAM metallopeptidase with thrombospondin type 1 motif, 9 (*ADAMTS9)* and which also contains the fragments of *MIR548AN*. This region is located $\approx 26 - -57kb$

upstream of *ADAMTS9*. *ADAMTS9* is a member of the a disintegrin and metalloproteinase with thrombospondin motif (*ADAMTS*) family, a group of genes encoding metalloproteases that lack transmembrane domains and are secreted into the extracellular matrix[183]. Members of the *ADAMTS* family have been implicated in control of organ shape during development and inhibition of angiogenesis[123].

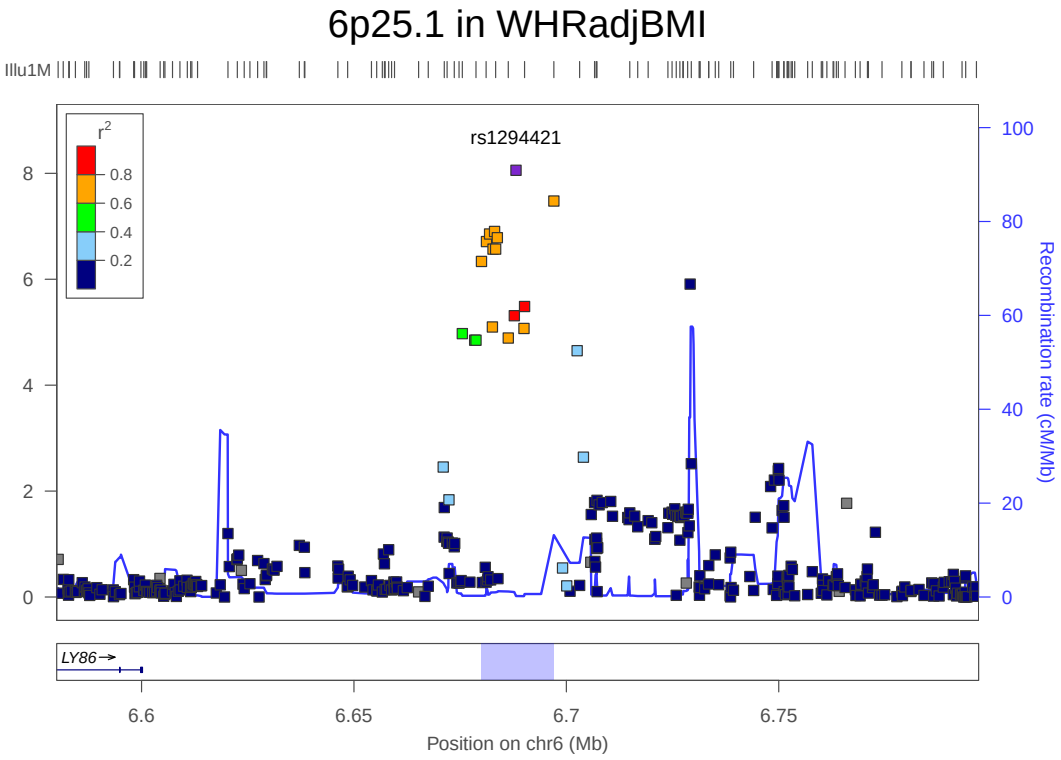
Figure 2.19: Discovery stage association at *5q35.2*.



5q35.2 The *5q35.2* locus is represented by lead marker rs6861681, with association signal for WHRadjBMI extending across $\approx 93kb$ of chromosome 5, ranging from $173227kb - 173320kb$ (see Figure 2.19). One gene (*CPEB4* overlaps this signal region. Cytoplasmic polyadenylation element binding protein 4 (*CPEB4)* is an RNA-binding protein that promotes polyadenylation-induced translation[184]. *CPEB4* nucleates a complex of factors that regulate polyadenylation elongation through a deadenylating enzyme and mediates many processes including

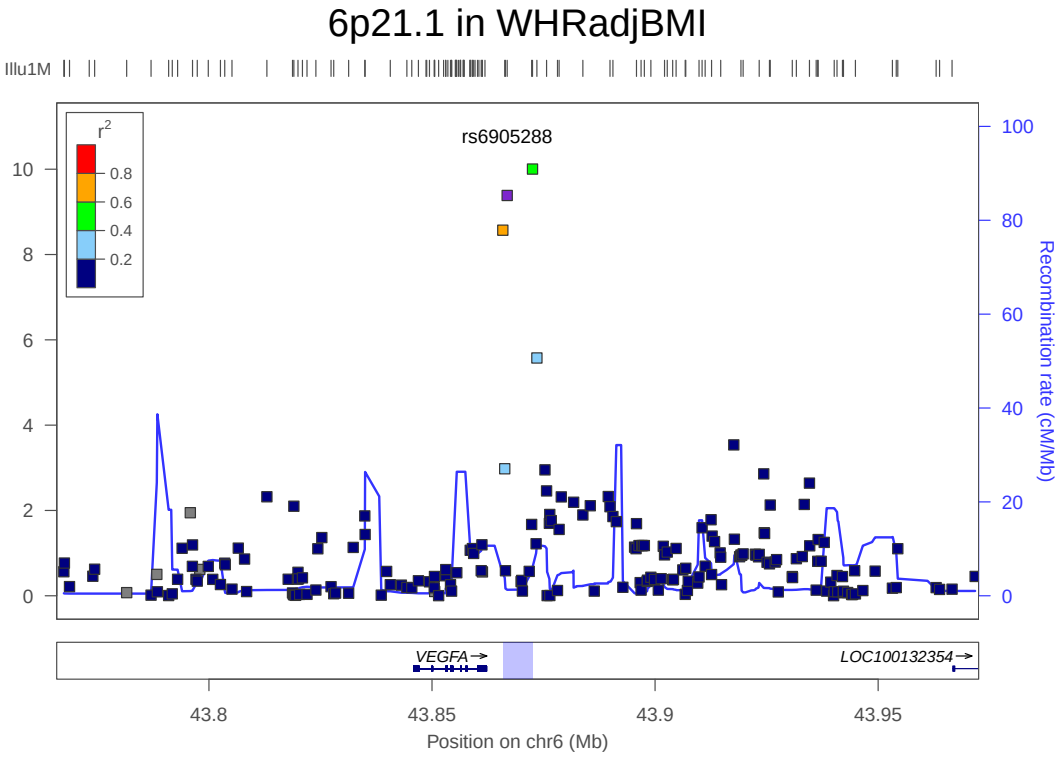
germ-cell development, cell division and cellular senescence and synaptic plasticity[185].

Figure 2.20: Discovery stage association at *6p25.1*.



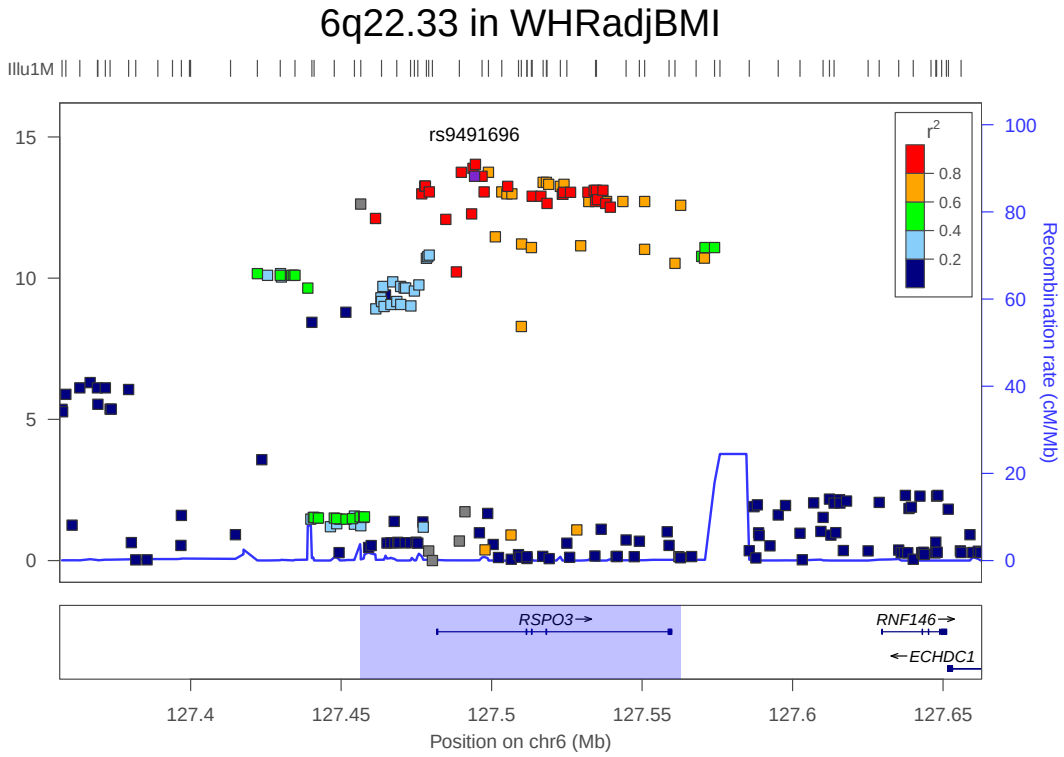
6p25.1 The *6p25.1* locus is represented by lead marker rs1294421, with association signal for WHRadjBMI extending across $\approx 17kb$ of chromosome 6, ranging from $6680kb - 6697kb$ (see Figure 2.20). No genes overlap this signal region. The nearest known protein-coding gene is *LY86* which plays a role in recognition of lipopolysaccharide via the toll-like receptor (*TLR*) pathway when bound as a heterodimer with radioprotective, 105 kDa (RP105). *LY86* is associated with asthma and has been suggested to play a role in autoimmune diseases[186, 187]. This region contains a $2,832bp$ CNV (CNVR2760.1: $6687180--6690011bp$)[188] which is highly correlated with the lead marker ($r^2 = 0.9845$).

Figure 2.21: Discovery stage association at *6p21.1*.



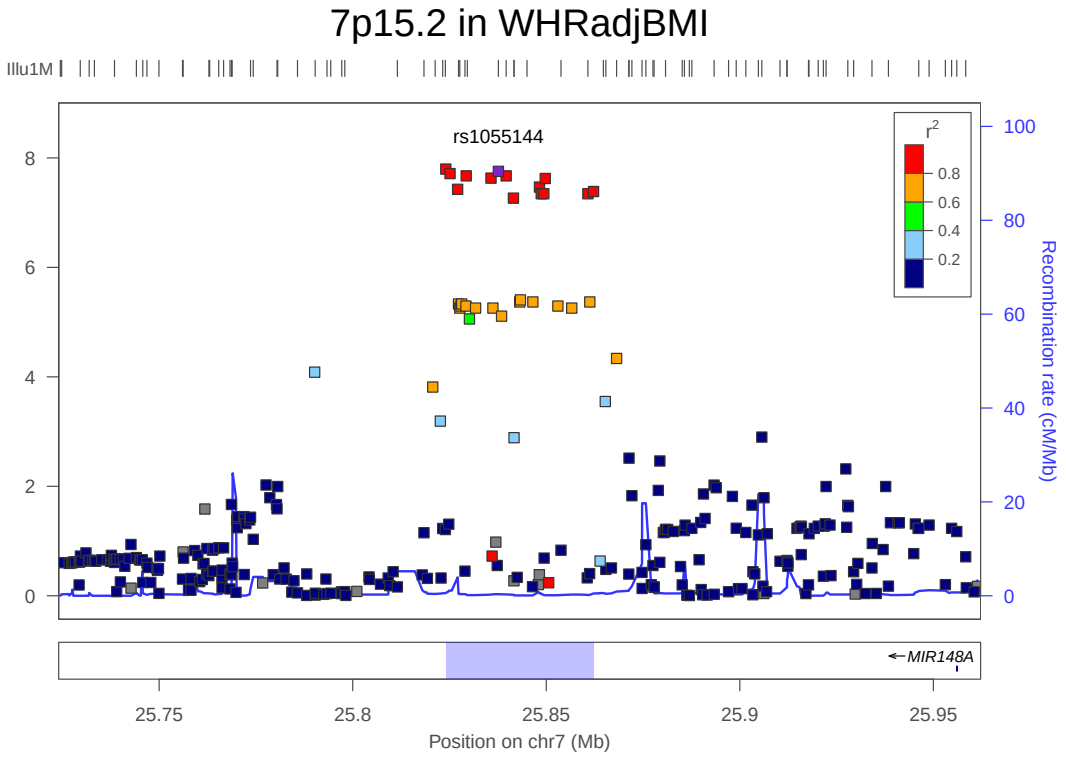
6p21.1 The *6p21.1* locus is represented by lead marker rs6905288, with association signal for WHRadjBMI extending across $\approx 7kb$ of chromosome 6, ranging from $43866kb - 43873kb$ (see Figure 2.21). No genes overlap this signal region. The associated region is located $\approx 3.7 - 10.7kb$ downstream of vascular endothelial growth factor A (*VEGFA)*. Multiple variants and mutations in *VEGFA* are risk factors for diabetic retinopathy[166, 189, 190], and variants in *VEGFA* have been nominally associated with T2D[68]. *VEGFA* is proposed as a key mediator of adipogenesis and angiogenesis[191], is highly expressed in adipose tissue, and has increased expression during adipocyte differentiation[192–195]. *VEGFA* serum concentrations are elevated in overweight and obese patients compared with lean subjects[196] and decrease after weight loss following bariatric surgery, behaving similarly to other hormones related to adipose mass, such as leptin and insulin[197].

Figure 2.22: Discovery stage association at *6q22.33*.



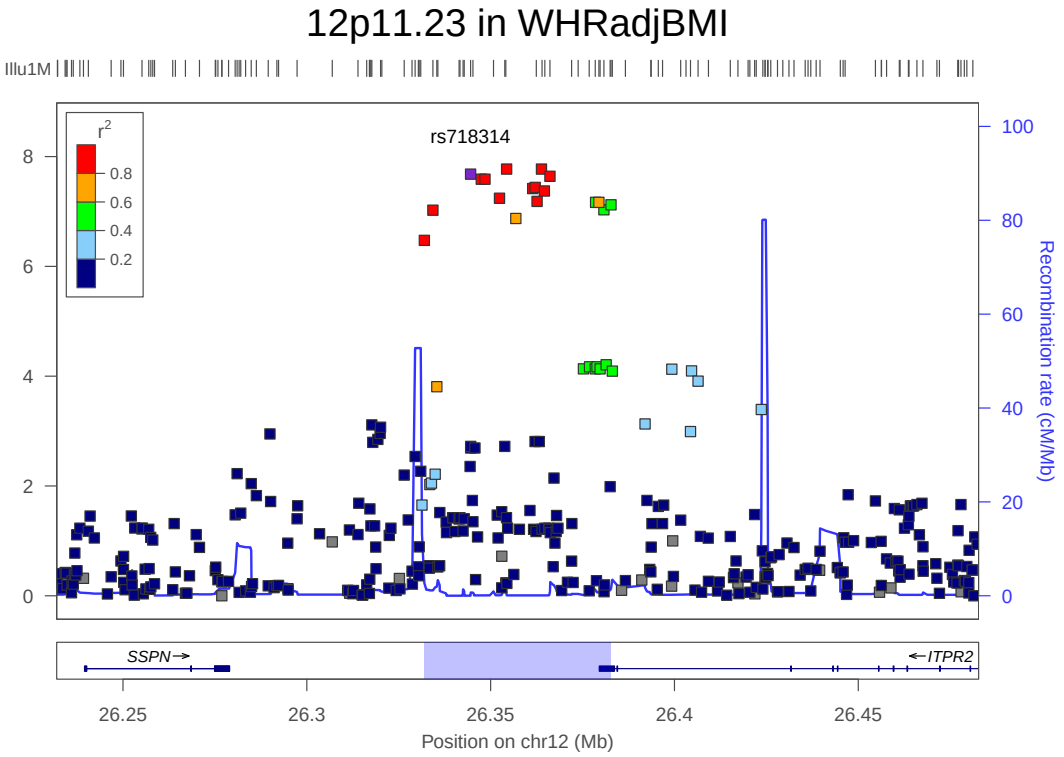
6q22.33 The *6q22.33* locus is represented by lead marker rs9491696, with association signal for WHRadjBMI extending across $\approx 106kb$ of chromosome 6, ranging from $127456kb - 127563kb$ (see Figure 2.22). One gene (*RSPO3*) is entirely contained within this signal region, and no other known genes overlap it. *RSPO3* encodes a secreted protein that regulates beta-catenin signaling[198], and is thought to promote angiogenesis and vascular development[199]. Mus musculus R-spondin 3 homolog (*Rspo3)* knockout mice die due to defects in placental development[200], and *Rspo3* is required for Mus musculus vascular endothelial growth factor A (*Vegfa)* expression and endothelial cell proliferation[199]. *Rspo3* has also been shown to be an oncogene in mouse mammary epithelial cells[201].

Figure 2.23: Discovery stage association at *7p15.2*.



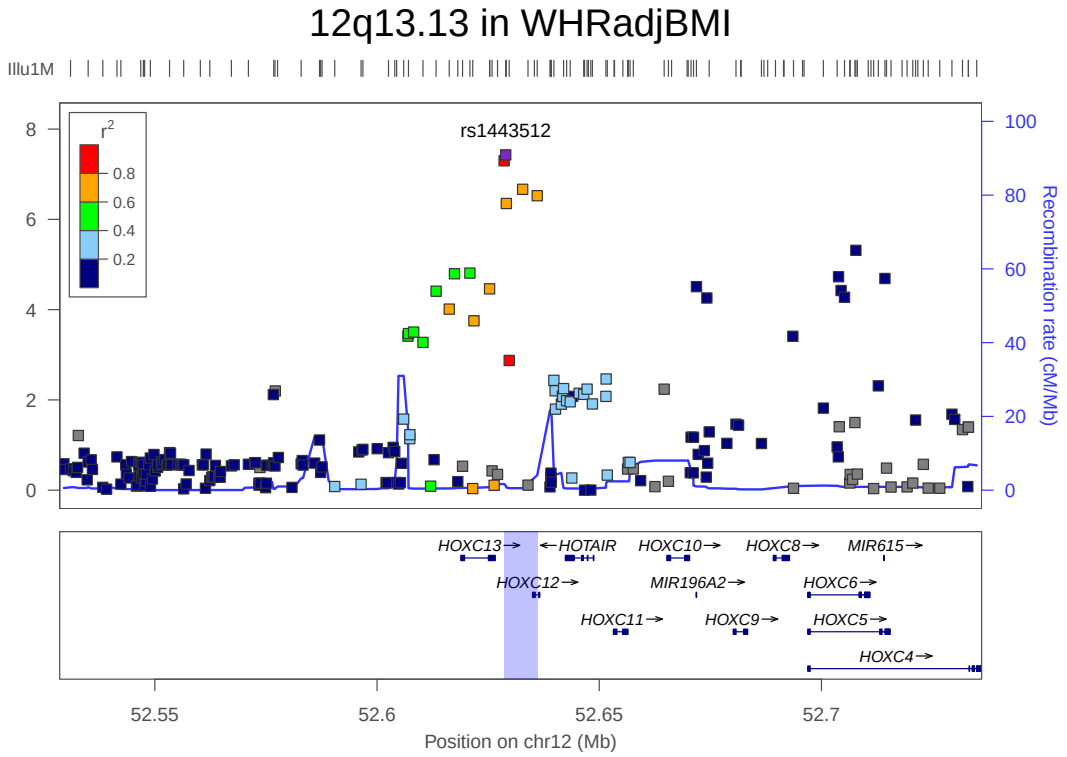
7p15.2 The *7p15.2* locus is represented by lead marker rs1055144, with association signal for WHRadjBMI extending across $\approx 38kb$ of chromosome 7, ranging from $25824kb - 25862kb$ (see Figure 2.23). No genes overlap the signal region. The nearest known gene to this signal region is microRNA 148a (*MIR148A)* ($\approx 93.9 - -132kb$ away). *MIR148A* is known to repress the DNA methyltransferase DNA (cytosine-5-)-methyltransferase 3 beta (DNMT3B)[202]. The nearest protein-coding gene is nuclear factor (erythroid-derived 2)-like 3 (*NFE2L3)*, which this region is $\approx 296 - -334kb$ upstream of. *NFE2L3* encodes a transcription factor that binds antioxidant response elements in target genes and Mus musculus nuclear factor, erythroid derived 2, like 3 (*Nfe2l3)* deficiency in mice has been linked to lymphoma development in vivo[203].

Figure 2.24: Discovery stage association at *12p11.23*.



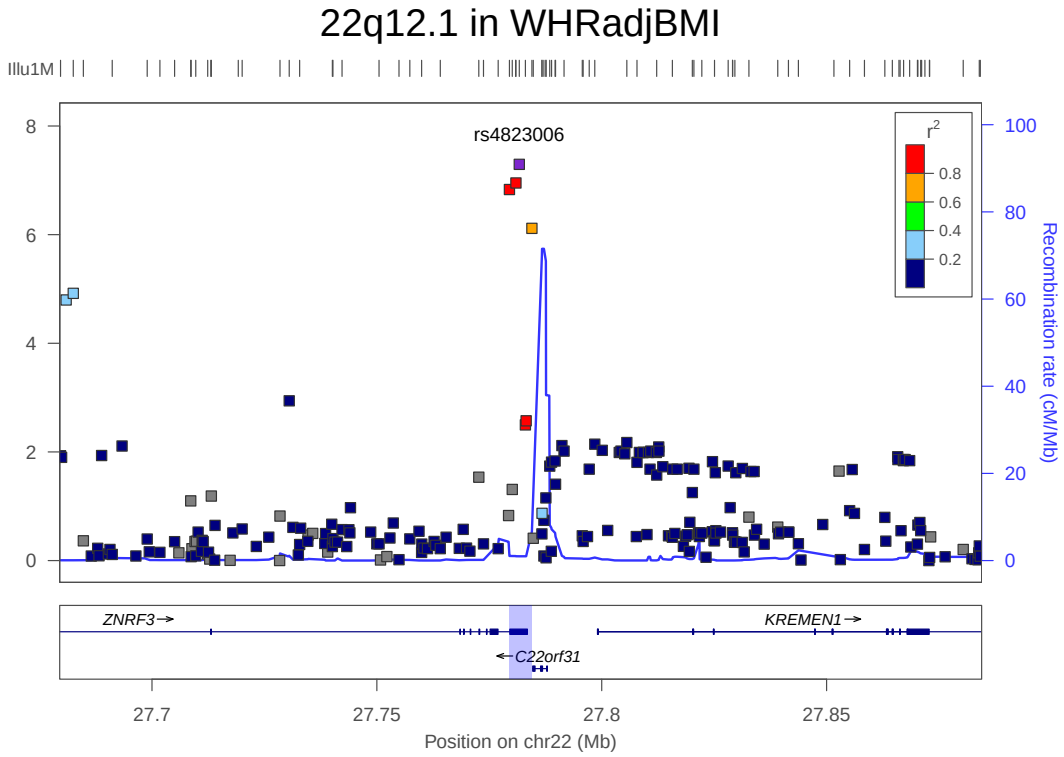
12p11.23 The *12p11.23* locus is represented by lead marker rs718314, with association signal for WHRadjBMI extending across $\approx 51kb$ of chromosome 12, ranging from $26332kb - 26383kb$ (see Figure 2.24). One gene (*ITPR2*) overlaps this signal region. Inositol 1,4,5-trisphosphate receptor, type 2 (*ITPR2)* is an intracellular calcium release channel. Mice lacking both Mus musculus inositol 1,4,5-triphosphate receptor 2 (*Itpr2)* and Mus musculus inositol 1,4,5-triphosphate receptor 3 (*Itpr3)* had impaired calcium signaling, secretion defects, hypoglycemia and lean body type[204].

Figure 2.25: Discovery stage association at *12q13.13*.



12q13.13 The *12q13.13* locus is represented by lead marker rs1443512, with association signal for WHRadjBMI extending across $\approx 7kb$ of chromosome 12, ranging from $52629kb - 52636kb$ (see Figure 2.25). One gene (*HOXC12*) overlaps this signal region. And the lead marker is located between homeobox C13 (*HOXC13)* and homeobox C12 (*HOXC12)*, both members of a family of genes that encode homeobox transcription factors that are important for the spatial distribution of cells during embryonic development[123].

Figure 2.26: Discovery stage association at *22q12.1*.



22q12.1 The *22q12.1* locus is represented by lead marker rs4823006, with association signal for WHRadjBMI extending across $\approx 5kb$ of chromosome 22, ranging from $27779kb - 27784kb$ (see Figure 2.26). One gene (*ZNRF3*) overlaps this signal region. The lead marker is located within the 3' untranslated region of *ZNRF3*, which encodes zinc and ring finger 3.

2.2.4 Discussion

These findings support the hypothesis that the genetic regulation of body fat distribution involves common genetic loci and processes that are largely distinct from those that influence BMI and risk of obesity. This is consistent with the evidence that WHR displays substantial heritability even after adjustment for BMI. The fourteen loci that were GWS in this study do not appear to overlap with those shown to be sig-

nificantly associated with BMI either in previous reports[4, 5, 149] or in the expanded meta-analysis of BMI in a similar set of samples as this study (see Section 2.3).

Another major difference between findings for WHRadjBMI and the findings of GWAMA for BMI (Section 2.3) relates to the evidence for sexual dimorphism that we have observed at several of the WHR-associated loci. Sex differences in the regulation of body fat distribution have long been acknowledged without a clear understanding of the underlying molecular mechanisms. These differences become apparent during puberty and are generally attributed to the influence of sex hormones[205]. Consistent with our findings, variance decomposition studies have shown that the genetic contribution to the overall variance in WHR, WC, and HC is greater in women[118]. Although there is some evidence for loci with differential sex effects influencing lipids[130], uric acid levels[206], and risk of schizophrenia[207], we are unaware of prior reports indicating such strong enrichment of sex-specific associations for any other phenotype, including BMI.

A main objective of GWA efforts is to identify associated regions in order to facilitate characterisation of the mechanisms involved in regulating the trait under study. Although considerable challenges are involved with progressing from the association of SNPs at a genomic locus to the establishment of causal alleles and pathways, we have been able to identify strong candidates for pathways involved at several of the loci. The cis eQTL data implicate *GRB14* as a candidate for the WHRadjBMI association at *2q24.3*, and we were able to show that the same *GRB14* variants are also associated with TG and insulin levels. These inferences about the role of *GRB14* are supported by evidence that *Grb14* deficient mice exhibit improved glucose homeostasis despite lower circulating insulin levels, as well as enhanced insulin signaling in liver and skeletal muscle[173]. The signal near *ADAMTS9* overlaps a previously-reported T2D locus[68], and the lead SNP for WHRadjBMI in our study is identical to the SNP displaying the strongest T2D association in a previous expanded T2D meta-analysis[208]. Evidence that *ADAMTS9* T2D risk alleles are associated with insulin

resistance in peripheral tissues[209] would be consistent with the effect of *ADAMTS9* variants being primarily on body fat distribution and potentially acting on T2D risk through that effect on body fat distribution. At the *6p21.1* locus, *VEGFA* is the most apparent biological candidate given the presumed role of *VEGFA* as a mediator of adipogenesis[191] and evidence that serum levels of *VEGFA* are correlated with obesity[196, 210]. Finally, at the *1p12* locus, *TBX15* emerges as the strongest candidate based on the cis eQTL data in omental fat, marked depot-specific differences in adipose tissue expression in mice and humans and associations between *TBX15* expression in visceral fat and WHR[156, 211].

Pathway analyses resulted in suggestive evidence of an over-representation of developmental processes, which is supported by developmental genes being implicated in fat accumulation[156, 211], as well as body fat distribution[156, 212, 213]. It has also been suggested that developmental genes may determine part of the adipocyte-specific expression patterns that have been observed in different fat depots[156].

Taken together, our findings appear to reveal a set of genes influencing body fat distribution that most likely have their principle effects in adipose or other peripheral tissue, iin contrast to the predominantly hypothalamic processes that have been reported to be involved in the regulation of BMI and overall adiposity[214].

2.3 Meta-analyses of body mass index

This analysis of BMI was undertaken by the GIANT consortium in 2009-2010, and was originally published inSpeliotes et al. [55].

2.3.1 Introduction

Obesity is a major and increasingly prevalent risk factor for multiple disorders, including type 2 diabetes and cardiovascular disease[215, 216]. While lifestyle changes have driven its prevalence to epidemic proportions, heritability studies provide evidence for a substantial genetic contribution ($h^2 \approx 40 - 70\%$) to obesity risk[217, 218]. BMI is an inexpensive, non-invasive measure of obesity that predicts the risk of related complications[219]. Identifying genetic determinants of BMI could lead to a better understanding of the biological basis of obesity.

GWAS of BMI have previously identified ten loci with genome-wide significant ($p < 5 \times 10^{-8}$) associations[4, 5, 57, 108, 149] in or near *FTO*, *MC4R*, transmembrane protein 18 (*TMEM18)*, glucosamine-6-phosphate deaminase 2 (*GNPDA2)*, brain-derived neurotrophic factor (*BDNF)*, neuronal growth regulator 1 (*NEGR1)*, SH2B adaptor protein 1 (*SH2B1)*, ets variant 5 (*ETV5)*, mitochondrial carrier 2 (*MTCH2)*, and potassium channel tetramerisation domain containing 15 (*KCTD15).* Many of these genes are expressed or known to act in the central nervous system, highlighting a likely neuronal component to the predisposition to obesity[5]. This pattern is consistent with results in animal models and studies of monogenic human obesity, where neuronal genes, particularly those expressed in the hypothalamus and involved in regulation of appetite or energy balance, are known to play a major role in susceptibility to obesity[220–222].

The ten previously identified loci account for only a small fraction of the variation in BMI. Furthermore, power calculations based on the effect sizes of established variants have suggested that increasing the sample size would likely lead to the discovery of

additional variants[5]. To identify more loci associated with BMI, we expanded the GIANT consortium GWAMA to include a total of 249,769 individuals of European ancestry.

2.3.2 Results

2.3.2.1 Stage 1 GWA studies identify novel loci associated with BMI

We first conducted a meta-analysis of GWA studies of BMI and up to 2.8 million imputed or genotyped SNPs using data from 46 studies including up to 123,865 individuals. This stage 1 analysis revealed 19 loci associated with BMI at GWS ($p < 5 \times 10^{-8}$) (Figure 2.27). These 19 loci included all ten loci from previous GWA studies of BMI[4, 5, 57, 108, 149], two loci previously associated with body weight[149] and one locus previously associated with WC[3]. The remaining six loci (*1p31.1*, *5q13.3*, *9p21.1*, *14q12*, *15q23*, and *16p12.3*) have not previously been associated with BMI or other obesity-related traits.

Figure 2.27: Genome-wide association plot in stage 1 for BMI. The x-axis represents genomic position and the y-axis is $-log_{10}(p)$. The red line at $p = 5 \times 10^{-6}$ indicates the threshold used for bringing loci forward into the follow-up stage.

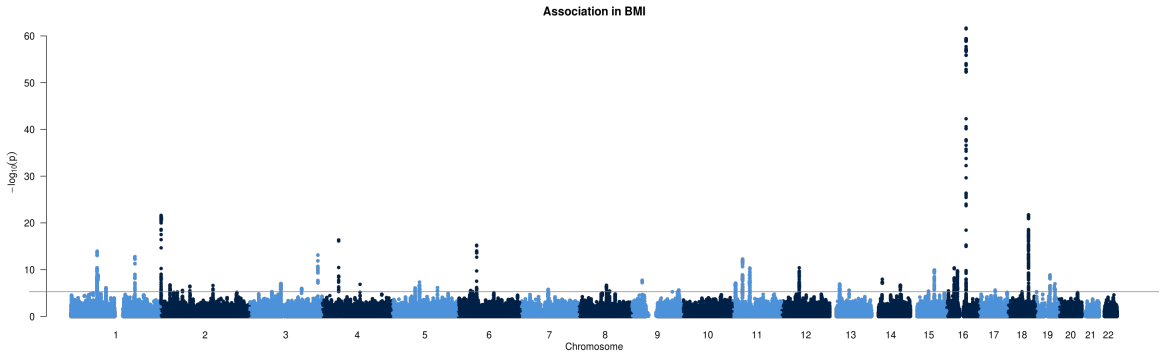


Figure 2.28: QQ plot of genome-wide association in stage 1 for BMI. The x-axis represents expected $-log_{10}(p)$ under H_0 while the y-axis represents the observed $-log_{10}(p)$.

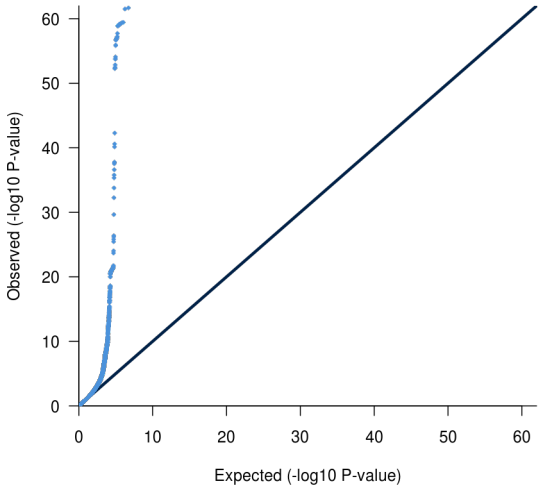


Table 2.3: Results for the 32 loci with genome-wide significance in stage 1+2.

Locus	SNP	EA	EAF	β	SE	s^2 explained	Stage 1 p	Stage 2 p	Stage 1+2 N	Stage 1+2 p	Previous Reports
16q12.2	rs1558902	A	42.0%	0.390	0.020	0.003400	2.05×10^{-62}	1.01×10^{-60}	192344	4.80×10^{-120}	[4, 5, 57, 108, 149]
2p25.3	rs2867125	C	83.0%	0.310	0.030	0.001500	2.42×10^{-22}	4.42×10^{-30}	197806	2.77×10^{-49}	[5, 149]
18q21.32	rs571312	A	24.0%	0.230	0.030	0.001000	1.82×10^{-22}	3.19×10^{-21}	203600	6.43×10^{-42}	[4, 5, 149]
4p13	rs10938397	G	43.0%	0.180	0.020	0.000800	4.35×10^{-17}	1.45×10^{-15}	197008	3.78×10^{-31}	[5]
11p14.1	rs10767664	A	78.0%	0.190	0.030	0.000700	5.53×10^{-13}	1.17×10^{-14}	204158	4.69×10^{-26}	[149]
1p31.1	rs2815752	A	61.0%	0.130	0.020	0.000400	1.17×10^{-14}	2.29×10^{-09}	198380	1.61×10^{-22}	[5, 149]
16p11.2	rs7359397	T	40.0%	0.150	0.020	0.000500	1.75×10^{-10}	7.89×10^{-12}	204309	1.88×10^{-20}	[5, 149]
3q27.2	rs9816226	T	82.0%	0.140	0.030	0.000300	7.61×10^{-14}	1.15×10^{-06}	196221	1.69×10^{-18}	[149]
11p11.2	rs3817334	T	41.0%	0.060	0.020	0.000100	4.79×10^{-11}	1.10×10^{-03}	191943	1.59×10^{-12}	[5]
19q13.11	rs29941	G	67.0%	0.060	0.020	0.000000	1.31×10^{-09}	2.40×10^{-02}	192872	3.01×10^{-09}	[5, 149]
1q25.2	rs543874	G	19.0%	0.220	0.030	0.000700	1.66×10^{-13}	2.41×10^{-11}	179414	3.56×10^{-23}	[149]
6p12.3	rs987237	G	18.0%	0.130	0.030	0.000300	5.97×10^{-16}	2.40×10^{-06}	195776	2.90×10^{-20}	[3]
12q13.13	rs7138803	A	38.0%	0.120	0.020	0.000400	3.96×10^{-11}	7.82×10^{-08}	200064	1.82×10^{-17}	[149]
14q31.1	rs10150332	C	21.0%	0.130	0.030	0.000200	2.03×10^{-07}	2.86×10^{-05}	183022	2.75×10^{-11}	[223]
2p23.3	rs713586	C	47.0%	0.140	0.020	0.000600	1.80×10^{-07}	1.44×10^{-16}	230748	6.17×10^{-22}	
16p12.3	rs12444979	C	87.0%	0.170	0.030	0.000400	4.20×10^{-11}	8.13×10^{-12}	239715	2.91×10^{-21}	
15q23	rs2241423	G	78.0%	0.130	0.020	0.000300	1.15×10^{-10}	1.59×10^{-09}	227950	1.19×10^{-18}	
19q13.32	rs2287019	C	80.0%	0.150	0.030	0.000400	3.18×10^{-07}	1.40×10^{-10}	194564	1.88×10^{-16}	
1p31.1	rs1514175	A	43.0%	0.070	0.020	0.000200	1.36×10^{-09}	7.04×10^{-06}	227900	8.16×10^{-14}	
4q24	rs13107325	T	7.0%	0.190	0.040	0.000300	1.37×10^{-07}	1.93×10^{-07}	245378	1.50×10^{-13}	
5q13.3	rs2112347	T	63.0%	0.100	0.020	0.000200	4.76×10^{-08}	8.29×10^{-07}	231729	2.17×10^{-13}	
9p21.1	rs10968576	G	31.0%	0.110	0.020	0.000200	1.88×10^{-08}	3.19×10^{-06}	216916	2.65×10^{-13}	
19q13.32	rs3810291	A	67.0%	0.090	0.020	0.000200	1.04×10^{-07}	1.59×10^{-06}	233512	1.64×10^{-12}	
2p16.1	rs887912	T	29.0%	0.100	0.020	0.000300	2.69×10^{-06}	1.72×10^{-07}	242807	1.79×10^{-12}	
3p12.1	rs13078807	G	20.0%	0.100	0.020	0.000200	9.81×10^{-08}	5.32×10^{-05}	237404	3.94×10^{-11}	
14q12	rs11847697	T	4.0%	0.170	0.050	0.000100	1.11×10^{-08}	2.25×10^{-04}	241667	5.76×10^{-11}	
2q22.2	rs2890652	C	18.0%	0.090	0.030	0.000200	2.38×10^{-07}	9.47×10^{-05}	209068	1.35×10^{-10}	
1p21.3	rs1555543	C	59.0%	0.060	0.020	0.000100	7.65×10^{-07}	4.48×10^{-05}	243013	3.68×10^{-10}	

Table 2.3: Results for the 32 loci with genome-wide significance in stage 1+2 (continued).

Locus	SNP	EA	EAF	β	SE	s^2 explained	Stage 1 p	Stage 2 p	Stage 1+2 N	Stage 1+2 p	Previous Reports
13q12.2	rs4771122	G	24.0%	0.090	0.030	0.000200	1.20×10^{-07}	8.24×10^{-04}	198577	9.48×10^{-10}	
5q23.2	rs4836133	A	48.0%	0.070	0.020	0.000100	7.04×10^{-07}	1.88×10^{-04}	241999	1.97×10^{-09}	
11p15.4	rs4929949	C	52.0%	0.060	0.020	0.000100	7.57×10^{-08}	1.00×10^{-03}	249791	2.80×10^{-09}	
6p21.31	rs206936	G	21.0%	0.060	0.020	0.000100	2.81×10^{-06}	7.39×10^{-04}	249777	3.02×10^{-08}	

2.3.2.2 Stage 2 follow-up leads to additional novel loci for BMI

To identify additional BMI-associated loci and to validate the loci that reached genome-wide significance in stage 1 analyses, we examined SNPs representing 42 independent loci (including the 19 GWS loci) with stage 1 $p < 5 \times 10^{-6}$. Variants were considered to be independent if the pair-wise LD was low ($r^2 < 0.1$) and if they were separated by at least $1Mb$ in genomic position (see Section 1.3.2.1 for the procedure used to prune the list of associated SNPs into independent loci). In stage 2, we examined these 42 SNPs in up to 125,931 additional individuals (79,561 newly genotyped individuals from 16 different studies and 46,370 individuals from 18 additional studies for which GWA data were available. In a joint analysis of stage 1 and stage 2 results, 32 of the 42 SNPs reached $p < 5 \times 10^{-8}$. Even after excluding SNPs within these 32 confirmed BMI loci, we still observed an excess of small p-values compared to the distribution expected under H_0 (see Figure 2.28), suggesting that more BMI loci remain to be uncovered.

The 32 confirmed associations included all 19 loci with $p < 5 \times 10^{-8}$ at stage 1; 12 additional novel loci at *1p21.3*, *2p23.3*, *2p16.1*, *2q22.2*, *3p12.1*, *4q24*, *5q23.2*, *6p21.31*, *11p15.4*, *13q12.2*, *19q13.32-50.9Mb*, and *19q13.32-52.3Mb*; and one locus (*14q31.1*) previously associated with WC[223] (Table 2.3). In all, our study increased the number of loci robustly associated with BMI from 10 to 32. Four of the 22 novel associations were previously associated with body weight[149] or WC[3, 223], whereas 18 loci had not previously been associated with any obesity-related trait in the general population. While we confirmed all loci previously established by large-scale GWA studies for BMI[4, 5, 57, 108, 149] and WC[3, 223], four loci identified by GWA studies for early-onset or adult morbid obesity: rs1805081 in Niemann-Pick disease, type C1 (*NPC1)* at *18q11.2* ($p = 0.0025$), rs1424233 at *16q23.2* ($p = 0.25$), rs10508503 at *10p13* ($p = 0.64$), and rs473034 at *8p23.1* ($p = 0.23$)[224, 225] showed limited or no evidence of association with BMI in this study.

As expected, the effect sizes of the 18 newly discovered loci are slightly smaller, for a given minor allele frequency, than those of the previously identified variants. The increased sample size also revealed signals with lower minor allele frequency. The BMI-increasing allele frequencies for the 18 newly identified variants ranged from 4%–87%, covering more of the allele frequency spectrum than previous, smaller GWA studies of BMI (24%–83%)[5, 149].

2.3.2.3 Impact of 32 confirmed loci on BMI, obesity, body size, and other metabolic traits

Together, the 32 confirmed BMI loci explained 1.45% of the inter-individual variation in BMI of the stage 2 samples, with the *FTO* SNP accounting for the largest proportion of the variance (0.34%) (Table 2.3). To estimate the cumulative effect of the 32 variants on BMI, we constructed a genetic-susceptibility score that sums the number of BMI-increasing alleles weighted by the overall stage 2 effect sizes in the ARIC study ($N = 8,120$), one of our largest population-based studies. For each unit increase in the genetic-susceptibility score, approximately equivalent to one additional risk allele, BMI increased by $0.17\frac{kg}{m^2}$, equivalent to a 0.435–0.551 kg gain in body weight in adults of 160–180 cm in height. The difference in average BMI between individuals with a high genetic-susceptibility score ($\geq$ 38 BMI-increasing alleles, 1.5% ($N = 124$) of the ARIC sample) and those with a low genetic-susceptibility score ($\leq$ 21 BMI-increasing alleles, 2.2% ($N = 175$) of the ARIC sample) was $2.73\frac{kg}{m^2}$, equivalent to a 6.99–8.85 kg body weight difference in adults of 160–180 cm in height.

All 32 confirmed BMI-increasing alleles showed directionally consistent effects on risk of being overweight (BMI $\geq 25\frac{kg}{m^2}$) or obese (BMI $\geq 30\frac{kg}{m^2}$) in stage 2 samples, with 30 of 32 variants achieving at least nominally significant associations. The BMI-increasing alleles increased the odds of being overweight by 1.013–1.138 fold, and the odds of being obese by 1.016–1.203 fold.

2.3.2.4 Potential functional roles and pathways analyses

Although associated variants typically implicate genomic regions rather than individual genes, we note that some of the 32 loci include candidate genes with established connections to obesity. Several of the 10 previously identified loci are located in or near genes that encode neuronal regulators of appetite or energy balance, including *MC4R*[221, 226], *BDNF*[227], and *SH2B1*[220, 228]. Each of these genes has been tied to obesity, not only in animal models, but also by rare human variants that disrupt each of these genes and lead to severe obesity[153, 229, 230].

To systematically identify biological connections among the genes located near the 32 confirmed SNPs, and to potentially identify new pathways associated with BMI, we performed pathway-based analyses using Meta-Analysis Gene-set Enrichment of variaNT Associations (MAGENTA)[231]. Specifically, we tested for enrichment of BMI genetic associations in biological processes or molecular functions that contain at least one gene from the 32 confirmed BMI loci. Using annotations from the Kyoto encyclopedia of genes and genomes (KEGG), Ingenuity, PANTHER, and gene ontology (GO) databases, we found evidence of enrichment for pathways involved in the platelet-derived growth factor (PDGF) signaling (PANTHER,$p = 8 \times 10^{-4}$, $q_{FDR} = 6.1 \times 10^{-3}$), translation elongation (PANTHER, $p = 8 \times 10^{-4}$, $q_{FDR} = 6.6 \times 10^{-3}$), hormone or nuclear hormone receptor binding (GO, $p < 5 \times 10^{-4}$, $q_{FDR} < 8.5 \times 10^{-3}$), homeobox transcription (PANTHER, $p = 1 \times 10^{-4}$, $q_{FDR} = 1.1 \times 10^{-2}$), regulation of cellular metabolism (GO, $p = 2 \times 10^{-4}$, $q_{FDR} = 3.1 \times 10^{-2}$), neurogenesis and neuron differentiation (GO, $p < 2 \times 10^{-4}$, $q_{FDR} < 3.4 \times 10^{-2}$), protein phosphorylation (PANTHER, $p = 1 \times 10^{-4}$, $q_{FDR} = 4.5 \times 10^{-2}$) and numerous other pathways related to growth, metabolism, immune and neuronal processes (GO, $p < 2 \times 10^{-3}$, $q_{FDR} < 4.6 \times 10^{-2}$).

2.3.2.5 Identifying possible functional variants

We used data from the 1000G and HapMap to explore whether the 32 confirmed BMI SNPs were in LD ($r^2 \geq 0.75$) with common missense SNPs or CNVs. Non-synonymous variants in LD with our signals were present in the *BDNF*, solute carrier family 39 (zinc transporter), member 8 (*SLC39A8)*, POC5 centriolar protein homolog (Chlamydomonas) (*POC5)*, 3-hydroxy-3-methylglutaryl-CoA reductase (*HMGCR)*, glutaminyl-peptide cyclotransferase-like (*QPCTL)*, gastric inhibitory polypeptide receptor (*GIPR)*, *MTCH2*, adenylate cyclase 3 (*ADCY3)*, and SKI family transcriptional corepressor 1 (*SKOR1)* genes. In addition, the rs7359397 signal at 16p11.2 was in LD with coding variants in several genes including *SH2B1*, apolipoprotein B receptor (*APOBR)*, and sulfotransferase family, cytosolic, 1A, phenol-preferring, member 2 (*SULT1A2)*. Furthermore, two SNPs tagged common CNVs. The first CNV was previously identified and is a 45kb deletion near *NEGR1*[5]. The second CNV is a 21kb deletion that lies 50kb upstream of G protein-coupled receptor, family C, group 5, member B (*GPRC5B)*; the deletion allele is tagged by the T-allele of rs12444979 ($r^2 = 1$).

Although simply being correlated with potentially functional variants does not prove that those variants are indeed causal, these do provide clues as to which genes and variants at these loci might be prioritized for future fine-mapping and functional follow-up efforts.

Many of the 32 BMI loci harbor multiple genes, so we examined whether gene eQTL analyses could also direct us to positional candidates. Gene expression data were available for human brain, lymphocytes, blood, SAT, visceral adipose tissue (VAT), and liver[146, 148, 232]. Significant cis-associations, defined at the tissue-specific level, were observed between 14 BMI-associated alleles and expression levels. In several cases, the BMI-associated SNP was the most significant SNP or explained a substantial proportion of the association with the most significant SNP for the gene

transcript in conditional analyses ($p_{adj} < 0.05$). These genes significantly associated included *NEGR1*, zinc finger CCCH-type containing 4 (*ZC3H4)*, transmembrane protein 160 (*TMEM160)*, *MTCH2*, NADH dehydrogenase (ubiquinone) Fe-S protein 3, 30kDa (NADH-coenzyme Q reductase) (*NDUFS3)*, general transcription factor IIIA (*GTF3A)*, *ADCY3*, *APOBR*, *SH2B1*, Tu translation elongation factor, mitochondrial (*TUFM)*, *GPRC5B*, IQ motif containing K (*IQCK)*, *SLC39A8*, sulfotransferase family, cytosolic, 1A, phenol-preferring, member 1 (*SULT1A1)*, and *SULT1A2*, making these genes higher priority candidates for follow-up. However, we note that some BMI-associated variants were correlated with the expression of multiple nearby genes, making it difficult to determine the most relevant gene in those regions.

2.3.3 Discussion

Using a two-stage genome-wide association meta-analysis of up to 249,796 individuals of European descent, we have identified 18 additional loci that are associated with BMI at GWS, bringing the total number of such loci to 32.

The combined effect on BMI of the associated variants at the 32 loci is modest, and account for only 1.45% of the genetic variation in BMI. There is an expectation that additional variance and biology will be explained using complementary approaches that capture variants not examined in the current study, such as lower frequency variants and short insertion-deletion polymorphisms (INDELs).

A primary goal of human genetic discovery is to improve understanding of the biology of conditions such as obesity[233]. The loci identified by this study appear to harbor few, if any, annotated genes with clear connections to the biology of weight regulation. This is likely a reflection of our limited understanding of the biology of BMI and obesity-related traits, and stands in contrast to results from GWAS of some other traits (such as autoimmune diseases or lipid levels). These results suggest that much novel biology underpinning obesity regulation remains to be uncovered, and GWAS

studies such as this one may provide a starting point to the discovery of that biological underpinning. A further examination of the associated loci through a combination of resequencing and fine-mapping to find causal variants, and genomic and experimental studies designed to assign function, could uncover novel insights into the biology of obesity.

In conclusion, we have performed GWA studies in large samples to identify numerous genetic loci associated with variation in BMI, a common measure of obesity. Because current lifestyle interventions are largely ineffective in addressing the challenges of growing obesity[234, 235], new insights into biology are needed to guide the development and application of future therapies and interventions.

2.4 Meta-analyses of weight

This analysis of adult weight was performed in 2008-2009 based on data collected by the GIANT consortium. It has not been published elsewhere.

2.4.1 Background

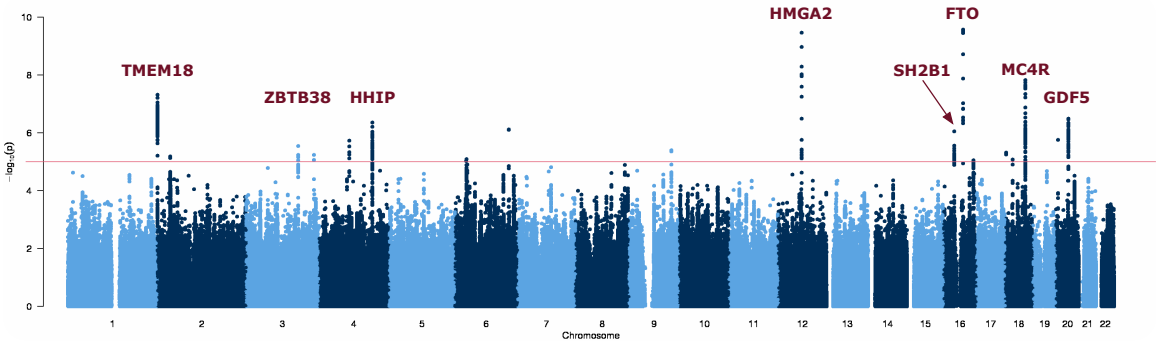
BMI is the traditional epidemiological measure of obesity, originally intended to classify inactive individuals with average body composition. BMI is a measure of weight in kilograms divided by height in meters squared. Square meters are used as an approximation of the natural relationship between height and weight, but analysis of the power law between weight and height has yielded estimates in the range of 2.3–2.7. While it is interesting to analyse genetic associations with BMI directly, it should yield greater power to perform the association using weight, which represents both obesity and overall body size, and then correct for height as a covariate in order to hone in on the obesity phenotype.

We performed a meta-analysis of adult weight in 13 studies, including both population-based and case-control studies (including case cohorts for T2D, CAD, and hypertension). All populations were of European descent, including cohorts from Italy, Switzerland, Germany, Sweden, Finland, the United Kingdom, and the United States. Case-control based studies were analyzed as separate case and control populations, making a total of 15 populations from the 13 GIANT consortium studies. Each of the 15 populations had been phenotyped and genotyped previously using study-specific methods, and QC was also performed separately by each study. QC typically included exclusion of markers significantly out of HWE or with high missing data rates. Genotypes were then imputed to obtain probably genotypes for $\approx 2.5M$ markers, based on the HapMap CEU reference panel. In parallel, phenotype data was prepared by stratifying by gender and then performing an inverse-normal transform on the weight

data. All $\approx 2.5M$ imputed genotypes were then tested for association with the transformed weight phenotype, using an additive model with age as a covariate. Data from all studies was then collected into the meta-analysis phase where QC was performed on the association results, excluding markers with low imputation quality scores or MAF less than 1%, and then doing a GC correction of the SE estimates to correct for possible population stratification. Finally, we performed a meta-analysis of the association results using the inverse-variance method to generate summary statistics (including effect estimates, SE, and a corresponding p-value) before doing a final GC correction to correct the SE in the meta-analysis results for any remaining population stratification. The resulting list of markers was then independentised into a list of markers with a HapMap r^2 of less then 0.2 between any two. An overview of our analysis procedure is shown in Figure 2.2.

The results exceeded the genome-wide signifiance threshold of 5×10^{-8} for *FTO*, *HMGA2*, *MC4R*, and *TMEM18*. *FTO*, *MC4R*, and *TMEM18* had already been established as being associated with BMI [4, 5, 57], while *HMGA2* had been established as being associated with height [236], so this analysis confirmed those findings (as expected since weight reflects both obesity and overall body size). All results are presented in a GWA plot in Figure 2.29.

Figure 2.29: Genome-wide association plot of overall (men and women meta-analysed together) results for GIANT weight, with known obesity genes labelled.

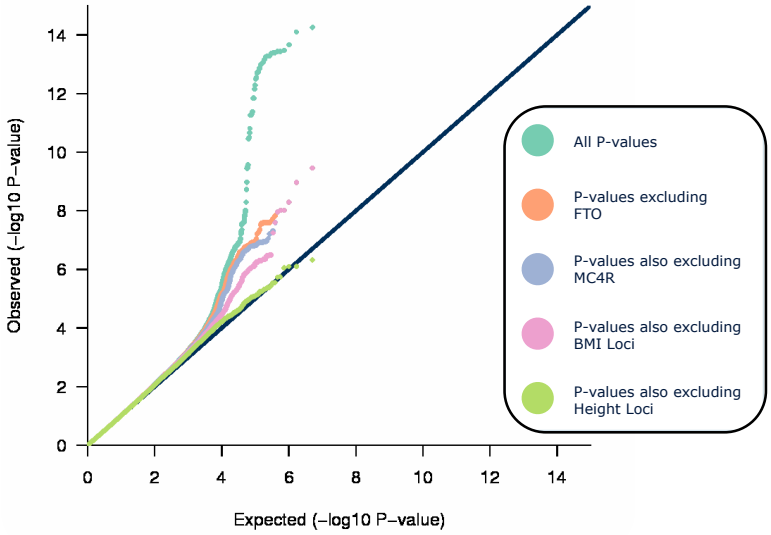


After excluding previously known loci, an excess of low p-values over what is expected by chance remained (see Figure 2.30), so it is likely that some of the signals below

the genome-wide significance threshold of 5×10^{-8} may be real associations with weight.

The top associations were in growth differentiation factor 5 (*GDF5)* ($p = 3.2 \times 10^{-7}$), in hedgehog interacting protein (*HHIP)* ($p = 4.4 \times 10^{-7}$), downstream of RAB32, member RAS oncogene family (*RAB32)* ($p = 7.8 \times 10^{-7}$), in *SULT1A1* ($p = 8.9 \times 10^{-7}$), upstream of solute carrier family 23 (nucleobase transporters), member 2 (*SLC23A2)* ($p = 1.8 \times 10^{-6}$), upstream of RasGEF domain family, member 1B (*RASGEF1B)* ($p = 1.9 \times 10^{-6}$), and in zinc finger and BTB domain containing 38 (*ZBTB38)* ($p = 2.9 \times 10^{-6}$). Of these, *GDF5*, *HHIP*, and *ZBTB38* had already been established as being associated with height [71, 237, 238], and *SULT1A1* has later been established as being associated with BMI [55], but no other associations appear to have been published for *RAB32*, *SLC23A2*, or *RASGEF1B*; so these may be novel associations with weight, though as we did not reach genome-wide significance, follow-up is required to establish this.

Figure 2.30: Exclusive QQ plots of overall (men and women meta-analysed together) results for GIANT weight, shown first with no exclusions, and then excluding known obesity and height genes.



As noted above, the weight phenotype appears to pick up signals of obesity as well

as overall body size (represented by height). As a result, it is somewhat difficult to interpret an association with weight. It would be more interesting to perform analyses of weight, adjusted for height, as an alternative to analysing BMI, since the weight-adjusted-for-height analysis may yield greater power than BMI does to detect obesity variants.

Results of the weight meta-analysis have not been published elsewhere, but the results for 43 candidate SNPs were provided to deCODE for inclusion in their paper on obesity as confirmation of their findings[149] and the upcoming expanded GIANT analyses should provide a follow-up to this work.

2.5 Sex-specific meta-analysis of nine anthropometric traits

These sex-specific analyses of nine anthropometric traits were undertaken in 2009-2011 by the GIANT consortium gender working group. A separate manuscript describing the work is being prepared and will be submitted for publication elsewhere.

Many anthropometric traits such as height, adiposity, and fat distribution are markedly different between men and women and may explain a portion of sex-specific susceptibilities to certain diseases. Some sexual dimorphism in body composition is already apparent during childhood, but it becomes much more pronounced as boys reach adolescence and tend to become taller and more muscular, while girls of the same age begin to experience an increase in average fat mass[205, 239–241].

These substantial differences in anthropometry may reflect sex-specific differences in biological processes such as adipogenesis, lipid storage, or lipolysis, which would suggest that genetic factors that modify or regulate genes involved in those pathways may also have sexually dimorphic effects.

GWAS have successfully identified genetic loci reproducibly associated with several anthropometric traits, including height[70, 71, 242–244], BMI[4, 5, 55, 57], and WHR[3, 54]. While most these studies did not perform genome-wide sex-stratified analyses – all with the exception ofLindgren et al. [3], whose sex-specific analysis identified the first known common genetic variant associated with WHR, follow-up analyses for loci found to be associated in overall analyses were performed in some of the studies. For example, inHeid et al. [54] a sex-specific analyses of the 14 SNPs examined in follow-up studies showed that effects were significantly more pronounced in women than in men for seven of the loci, and a follow-up analyses of overall GWAS findings for height found some evidence of sex-specific association in men[243].

The obvious differences in physical appearance between men and women along with

with the strong evidence of sex-specific effects for the recently identified WHR loci raises the question as to whether a more systematic approach would be able to discover additional sexually dimorphic genetic variants influencing anthropometric traits. GWAS stratified by sex not only improves power to identify sex-specific associations, but also allows for formal tests for sex differences, so we set out to investigate whether we could detect additional sexually dimorphic associations with anthropometric measures using sex-stratified GWAS, and if so, whether these association signals were of concordant effect direction (CED) or opposite effect direction (OED).

We defined OED signals as having an effect in both men and women, but of opposite direction, while CED signals include both sex-specific effects (in which the effect in either men or women is not significantly different from 0) and also those with an effect in both men and women in the same direction but of different magnitudes. Within the GIANT consortium, we performed meta-analyses of sex-specific GWAS on six anthropometric measures from 117 studies comprising up to 270,775 individuals (up to 122,981 men and 147,794 women) in order to investigate the extent and nature of sex-specific genetic effects on anthropometry.

2.5.1 Results of sex-specific analyses

2.5.1.1 Discovery meta-analysis of sex-specific genome-wide association study for anthropometric traits

We performed sex-specific analyses in 46 studies (see Table 2.4) comprising a total of up to 60,586 men and 73,137 women, testing $\approx$ 2.8 million SNPs for association with six anthropometric traits: height, weight, BMI, WC, HC, and WHR. The latter three traits were analyzed both unadjusted and adjusted for BMI: WC-adjusted-for-BMI (WC$_{ADJ}$BMI), HC-adjusted-for-BMI (HC$_{ADJ}$BMI), and WHR$_{ADJ}$BMI. This yielded nine phenotypes in total on which we performed inverse-variance weighted fixed-effects meta-analysis (see Section 1.3.1.3) stratified by sex, which resulted in a total of 18 meta-analyses (9 phenotypes, each analysed in both men and women). Study-specific

analysis methods have been described previously[54, 55, 70] and in general followed standard procedures for GWAS (see Section 1.2.4).

Table 2.4: Sample size of studies in the discovery stage.

	Hip		WC		WHR		Height		Weight		BMI	
Study	**Men**	**Women**	**Men**	**Women**	**Men**	**Women**	**Men**	**Women**	**Men**	**Women**	**Men**	**Women**
ADVANCE Cases			114	160			114	161	114	161	114	161
ADVANCE Controls			128	179			128	181	128	182	128	181
AGES Reykjavik			1351	1865			1352	1867	1351	1856	1351	1856
Amish	465	435	465	435	465	435	470	437	468	437	468	437
ARIC	3821	4283	3821	4283	3821	4283	3823	4287	3822	4286	3822	4286
B58C T1DGC	1198	1292	1198	1292	1198	1292	1261	1330	1259	1328	1259	1328
B58C WTCCC	700	722	700	722	700	722	741	738	741	738	741	738
BRIGHT	667	982	668	984	667	982	719	1087	719	1087	719	1087
CAD WTCCC							1491	388	1489	387	1489	387
CAPS1 Cases							489		485		484	
CAPS1 Controls							491		485		483	
CAPS2 Cases							1483		1424		1423	
CAPS2 Controls							519		504		500	
CHS	1274	1943	1274	1944	1274	1943	1277	1955	1276	1952	1276	1952
CoLaus	2544	2861	2546	2862	2544	2861	2547	2862	2547	2861	2547	2861
DECODE	2610	3273	2610	3273	2610	3273	9213	17586	9213	17586	9213	17586
DGI Cases	509	449	509	449	509	449	687	630	688	629	688	629
DGI Controls	241	228	241	228	241	228	553	537	553	535	553	535
EGCUT	195	228	195	228	195	228	697	720	697	720	697	720
EPIC Obesity Study	1130	1284	1130	1284	1130	1284	1621	1931	1131	1284	1131	1284
ERF EUROSPAN	890	1170	890	1170	890	1170	890	1170	890	1170	890	1170
Fenland	615	787	615	787	615	787	615	787	615	787	615	787
FHS Cases	208	233	208	233	208	233	208	233	208	233	208	233
FHS Controls	207	207	207	207	207	207	208	207	208	207	208	207
FRAM	1562	1715	1562	1715	1562	1715	3700	4389	3706	4388	3706	4388
FTC		120		120		120		125		125		125

Table 2.4: Sample size of studies in the discovery stage (continued).

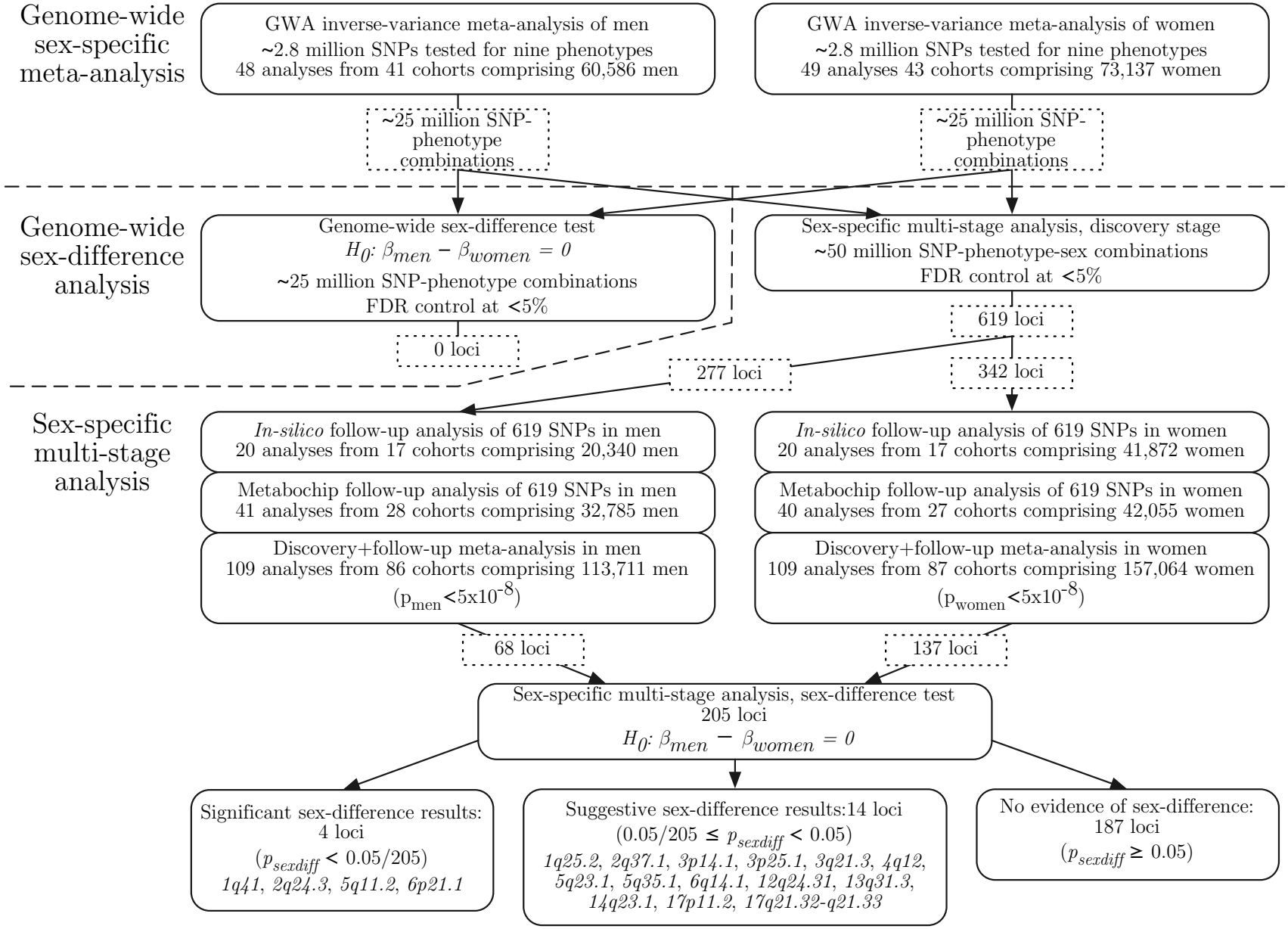
	Hip		WC		WHR		Height		Weight		BMI	
Study	**Men**	**Women**	**Men**	**Women**	**Men**	**Women**	**Men**	**Women**	**Men**	**Women**	**Men**	**Women**
FUSION Cases	620	464	620	465	620	464	617	465	623	469	623	469
FUSION Controls	571	598	571	598	571	598	569	598	572	599	572	599
GENMETS Cases	425	423	425	423	425	423	410	414	410	414	425	432
GENMETS Controls	401	423	401	423	401	423	401	422	401	422	401	423
GerMiFSI							394	206	394	206	394	206
GerMiFSII							901	223	901	223	901	223
KORAS3	812	826	812	827	812	826	813	830	813	829	813	829
KORAS4	883	929	883	929	883	929	883	928	883	929	883	928
MICROS	118	175	118	175	112	169	467	612	468	612	475	622
MIGEN							1622	1030	1623	1028	1619	1028
NBS WTCCC							696	745	694	743	694	743
NFBC1966	2246	2245	2246	2245	2246	2245	2250	2249	2250	2247	2250	2247
NHS		1632		1632		1632		2265		2265		2265
NSPHS	307	339					308	344	307	342	307	340
NTRNESDA	1206	2303	1206	2303	1206	2303	1211	2311	1210	2306	1210	2306
ORCADES	324	371	324	371	324	371	324	371	324	371	332	384
PLCO							2244		2238		2238	
PROCARDIS	612	346	612	346	612	346	1700	612	1700	612	1700	612
RS I	2266	3202	2266	3205	2266	3202	2372	3375	2375	3383	2372	3372
RUNMC							1777	1096	1777	1096	1777	1096
SARDINIA	1886	2415	1886	2416	1886	2415	1883	2415	1883	2415	1885	2416
SASBAC Cases								794		794		793
SASBAC Controls								758		760		755
SEARCH UKOPS								1592		1581		1556
SHIP	2019	2073	2019	2073	2019	2073	2019	2073	2019	2073	2019	2073
T2D WTCCC	1085	786	1085	786	1085	786	1105	798	1105	798	1105	798
TwinsUK		1096		1096		1094		1479		1477		1477

Table 2.4: Sample size of studies in the discovery stage (continued).

	Hip		WC		WHR		Height		Weight		BMI	
Study	**Men**	**Women**	**Men**	**Women**	**Men**	**Women**	**Men**	**Women**	**Men**	**Women**	**Men**	**Women**
VIS	325	458	325	459	325	458	325	459	325	445	328	467

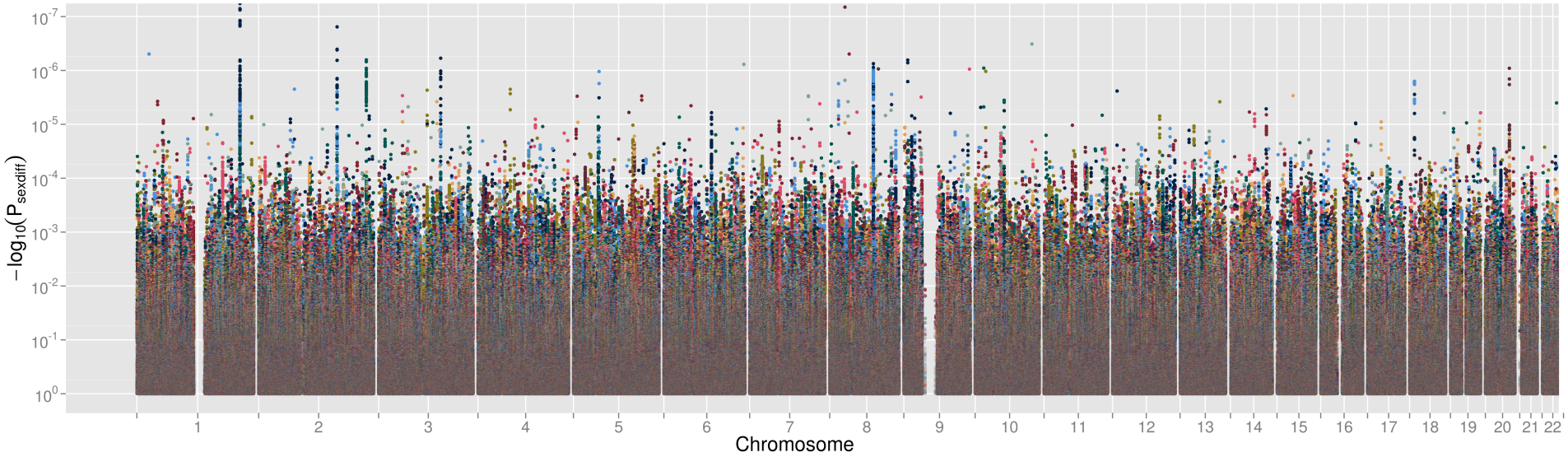
We were interested in identifying loci with OED and CED effects for the nine phenotypes. To optimize power, we performed two types of analyses for each phenotype: a genome-wide sex-difference analysis and a sex-specific multi-stage analysis (see Figure 2.31). For the genome-wide sex-difference analysis, we conducted a test of the significance of the difference in effect estimate between men and women (sex-difference test) for all markers genome-wide, which is best powered to detect OED signals for which there is a significant association in both sexes. For the sex-specific multi-stage analysis, we evaluated the significance of sex-specific association p-values in the discovery stage and subsequently tested for sex-difference in a follow-up stage. The sex-specific multi-stage analysis is best powered to detect CED signals, but could also pick up OED signals, though with less power than a genome-wide test of sex difference.

Figure 2.31: Overall diagram of sex-specific analyses.



The genome-wide sex-difference analysis did not identify any SNP that showed significant sex-difference controlling for FDR at 5% (Figure 2.32), so no SNP was taken forward from this analysis into follow-up. When examining sex-difference p-values separately by phenotype, we did observe a deviation from the expected distribution of small sex-difference p-values (under the null hypothesis of no sex-difference), particularly for the waist phenotypes (WHRadjBMI, WHR, WCadjBMI), which is illustrated in the phenotype-specific QQ-plot (Figure 2.33).

Figure 2.32: Discovery stage genome-wide sex-difference p-values for all nine phenotypes, with the x-axis representing genomic position chromosome-by-chromosome and with p-value for sex difference on the y-axis on a $-log_{10}$ scale. Colour indicates which of the nine phenotypes each point represents, and for p-values above a threshold of $p > 1 \times 10^{-4}$, increasing alpha blending (transparency) is used to facilitate display of overlapping points. WHRadjBMI is represented by dark blue, WHR by light blue, WCadjBMI by dark green, WC by light green, HCadjBMI by red, HC by pink, BMI by brown, weight by light tan, and height by olive green.



The sex-specific multi-stage analysis showed an excess of small p-values (Figure 2.34) in the discovery stage, and controlling for FDR at 5% (see Section 1.3.3.3) resulted in the selection of 619 independent SNPs (254 loci for height, 50 for weight, 51 for BMI, 38 for WC, 48 for WCadjBMI, 49 for HC, 45 for HCadjBMI, 31 for WHR, and 53 for WHRadjBMI) which showed the most significant sex-specific association with any of the nine phenotypes to be taken forward into the follow-up stage. The equivalent p-value threshold of the FDR control was $p < 2 \times 10^{-5}$. Of these 619 loci, 310 would have been selected in an overall analysis of men and women together using the same threshold of $p < 2 \times 10^{-5}$, while the other 309 would not have been selected by such an overall analysis.

All of the SNPs selected for follow-up came from the sex-specific multi-stage analysis, while no SNPs were selected from the genome-wide sex-difference analysis.

Figure 2.33: QQ plot of discovery stage genome-wide sex-difference p-values observed plotted vs their expected distribution. Colour indicates which of the nine phenotypes the point represents, and for p-values above a threshold of $p > 1 \times 10^{-4}$, increasing alpha blending (transparency) is used to allow display of overlapping points. The $p_{expected} = p_{observed}$ line is drawn in black. WHRadjBMI is represented by dark blue, WHR by light blue, WCadjBMI by dark green, WC by light green, HCadjBMI by red, HC by pink, BMI by brown, weight by light tan, and height by olive green.

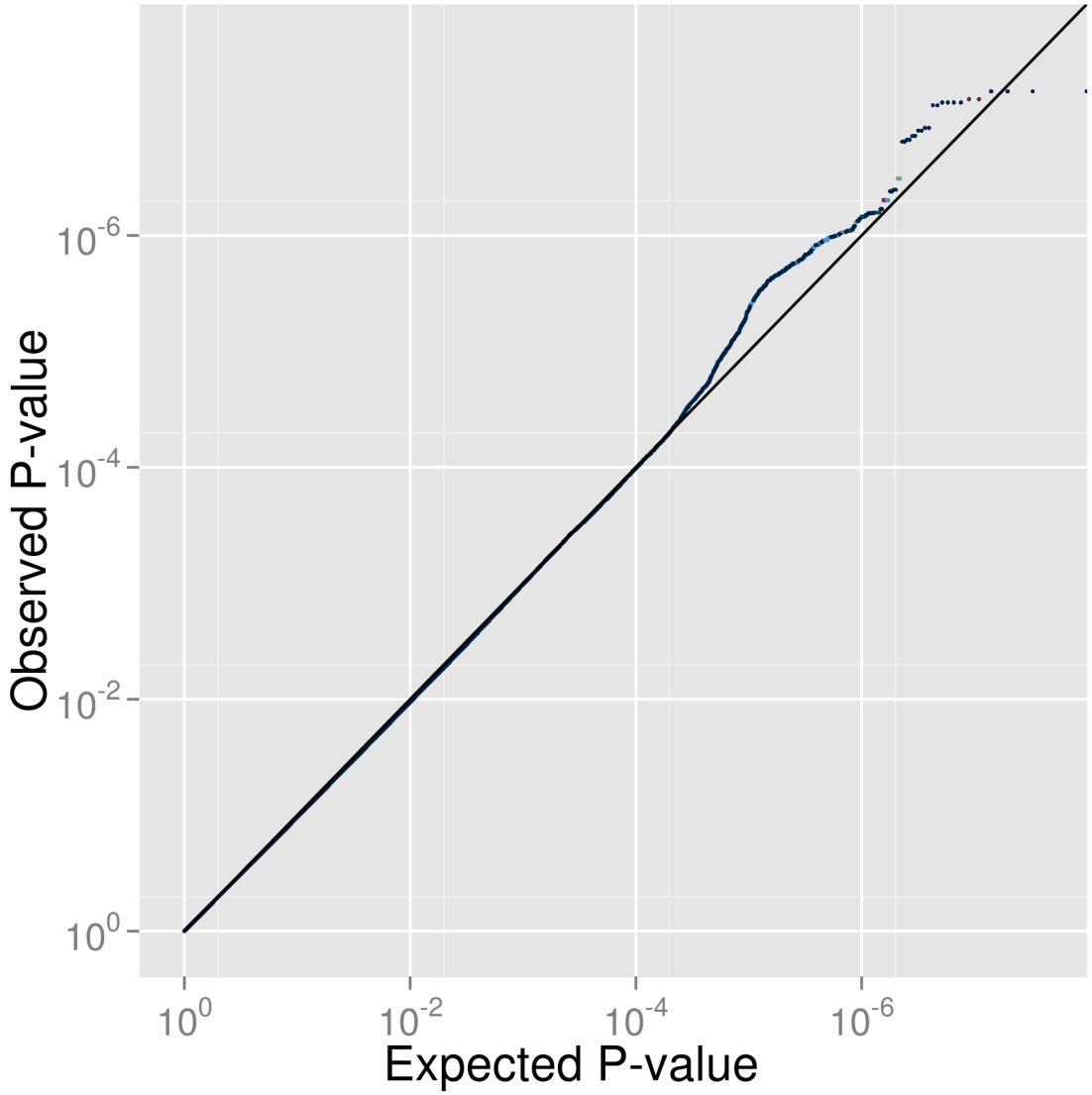
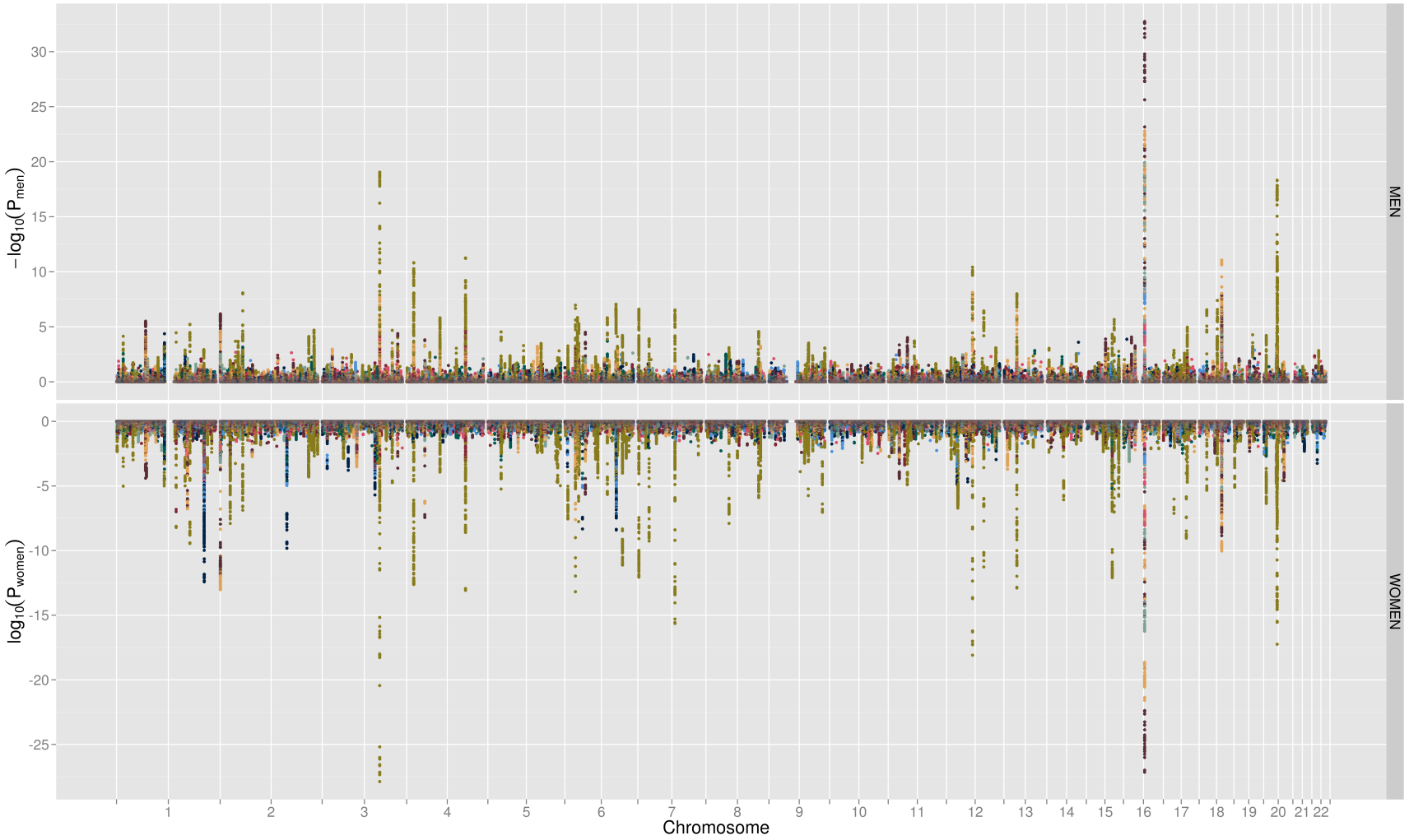


Figure 2.34: Discovery stage genome-wide sex-specific association p-values for all 9 phenotypes in men and women, with the x-axis representing genomic position chromosome-by-chromosome and sex-specific p-value for association on the y-axis on a $-log_{10}$ scale. To aid interpretation of the relative association between men and women, the scale for women is reversed to log_{10}. Colour indicates which of the nine phenotypes each point represents, and for p-values above a threshold of $p > 1 \times 10^{-4}$, increasing alpha blending (transparency) is used to facilitate display of overlapping points. WHRadjBMI is represented by dark blue, WHR by light blue, WCadjBMI by dark green, WC by light green, HCadjBMI by red, HC by pink, BMI by brown, weight by light tan, and height by olive green.



In the follow up stage, we examined the sex-specific associations of the 619 SNPs for the nine phenotypes in 18 studies with in-silico genotype information (up to 20,340 men and 41,872 women; see Table 2.5) and in 28 studies typed on Metabochip (up to 32,785 men and 42,055 women; see Table 2.5).

Figure 2.35: QQ plot of discovery stage sex-specific association p-values observed plotted vs their expected distribution. Colour indicates which of the nine phenotypes the point represents, and for p-values above a threshold of $p > 1 \times 10^{-4}$, increasing alpha blending (transparency) is used to allow display of overlapping points. The $p_{expected} = p_{observed}$ line is drawn in black. WHRadjBMI is represented by dark blue, WHR by light blue, WCadjBMI by dark green, WC by light green, HCadjBMI by red, HC by pink, BMI by brown, weight by light tan, and height by olive green.

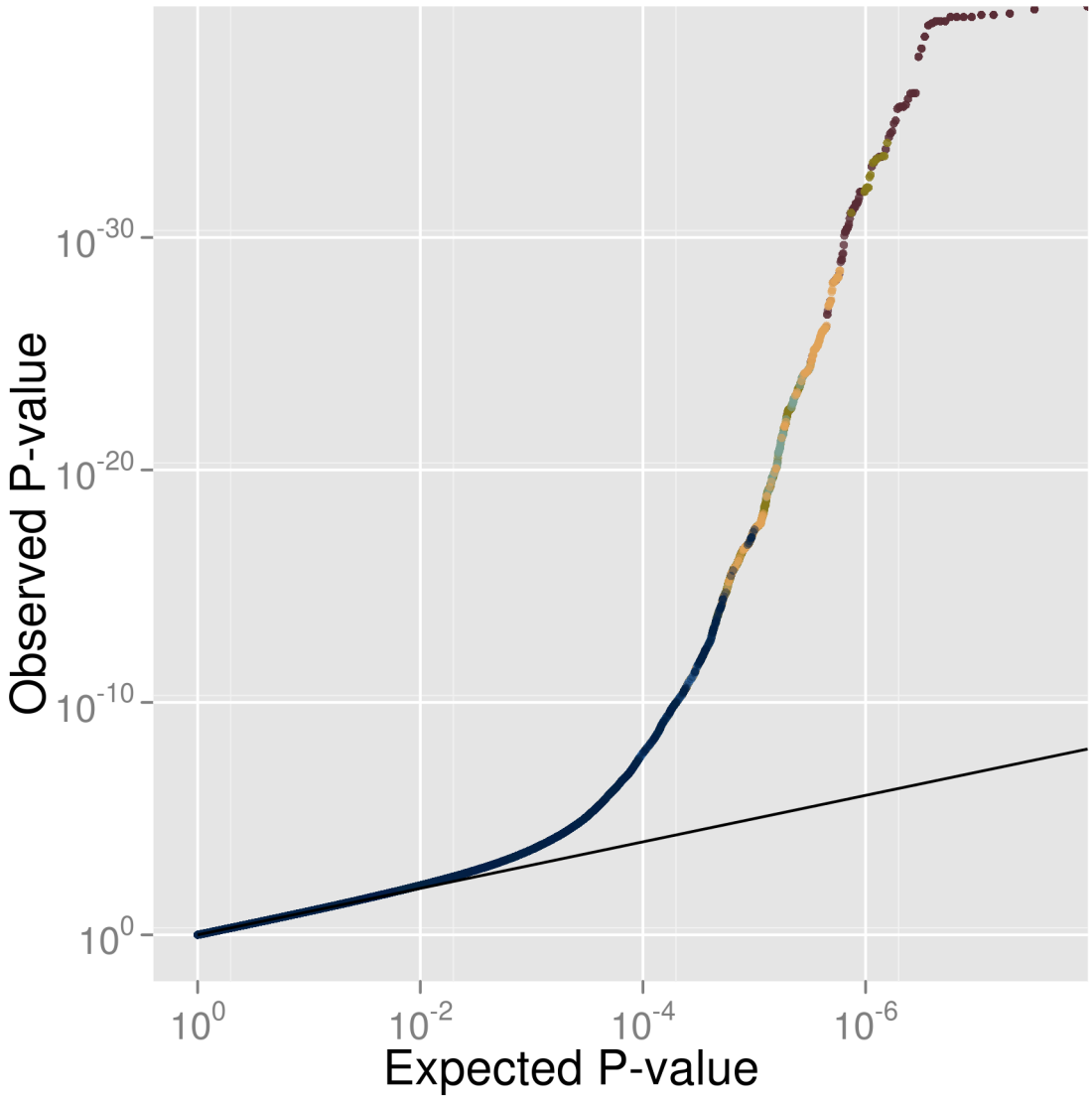


Table 2.5: Sample size of studies in the follow-up stage. Type indicates whether the study was an *in silico* (GWAS) or MetaboChip (MC) study.

		Hip		WC		WHR		Height		Weight		BMI	
Study	**Type**	**Men**	**Women**	**Men**	**Women**	**Men**	**Women**	**Men**	**Women**	**Men**	**Women**	**Men**	**Women**
FamHS	GWAS	1027	1194	1027	1194	1027	1194	1028	1195			1028	1195
BSN	GWAS	537	748	537	749	536	748	558	770			558	769
Corogene Cases	GWAS	1015	881	1015	881	1015	881	1005	880			1005	880
Corogene Controls	GWAS							1128	549			1127	549
EGCUT	GWAS	358	357	358	357	358	357	351	349			358	357
FINGEST	GWAS							756	206			750	204
GOOD	GWAS	938		938		938		938				938	
HBCS	GWAS	735	988	735	989	735	988	736	990			736	990
Hypergenes Cases	GWAS	170	152	170	152	170	152	1072	538			1072	538
Hypergenes Controls	GWAS	120	146	120	146	120	146	998	684			998	684
Lifelines	GWAS	1373	1993	1373	1993	1373	1993	1373	1994			1373	1994
MGS	GWAS							1247	1350			1247	1350
PLCO2 Cases	GWAS							2043	934			2041	933
PLCO2 Controls	GWAS							649	544			645	544
PREVEND	GWAS	1964	1873	1964	1873	1964	1873	1966	1872			1966	1871
QIMR	GWAS							1470	2157			1470	2157
RS2	GWAS	877	1035	877	1035	877	1035	876	1035			876	1035
RS3	GWAS	842	1085	842	1085	842	1085	877	1129			877	1129
Sorbs	GWAS	361	515	363	516	361	515	361	515			361	515
WGHS	GWAS		20538		20566		20529		23099				
YFS	GWAS	907	1075	907	1075	907	1075	911	1084			908	1081
ADVANCE CAD Cases	MC			678	222			678	222	678	222	678	222

Table 2.5: Sample size of studies in the follow-up stage (continued).

		Hip		WC		WHR		Height		Weight		BMI	
Study	**Type**	**Men**	**Women**	**Men**	**Women**	**Men**	**Women**	**Men**	**Women**	**Men**	**Women**	**Men**	**Women**
ADVANCE CAD Controls	MC			421	247			421	247	421	247	421	247
AMCPAS Cases	MC	368	122					368	122	368	122	368	122
BC58	MC	1231	897	1232	897	1231	897	1234	901	1232	902	1232	902
BHS Cases	MC	79	40	80	40	79	40	83	42	82	42	82	42
BHS Controls	MC	85	51	86	51	85	51	87	51	87	51	87	51
CARDIOGENICS Cases	MC	328	51					328	51	328	51	328	51
CARDIOGENICS Controls	MC	148	227					148	227	148	227	148	227
D2D2007 DPS DRSEXTRA FUSIONS2 METSIM Cases	MC	2151	757	2150	757	2150	757	2152	759	2152	759	2152	759
D2D2007 DPS DRSEXTRA FUSIONS2 METSIM Controls	MC	3669	2957	3669	2955	3669	2954	3672	2963	3672	2963	3672	2963
DILGOM	MC	1769	2098	1765	2098	1763	2092	1775	2114	1769	2115	1769	2115
DUNDEE Cases	MC	1926	1347	1750	1215			1926	1347	1923	1346	1922	1345
DUNDEE Controls	MC	1918	1789	1917	1787			1918	1789	1918	1788	1918	1788
EAS	MC	353	378					353	378	352	377	353	378
EGCUT CAD Cases	MC	346	354	346	354	346	354	346	354	346	354	346	354
EGCUT Controls	MC	340	600	340	600	340	600	341	601	341	601	341	601
EGCUT DB Cases	MC	272	417	272	417	272	417	355	613	355	613	355	613

Table 2.5: Sample size of studies in the follow-up stage (continued).

		Hip		WC		WHR		Height		Weight		BMI	
Study	**Type**	**Men**	**Women**	**Men**	**Women**	**Men**	**Women**	**Men**	**Women**	**Men**	**Women**	**Men**	**Women**
Ely	MC	736	852	736	853	735	852	744	855	744	856	744	855
EPIC Cases	MC	432	295	432	295	432	295	432	294	432	292	432	291
EPIC Controls	MC	410	549	411	550	410	549	410	552	411	551	410	551
Fenland	MC	1485	1698	1485	1698	1484	1698	1486	1698	1486	1698	1486	1698
GLACIER	MC	2381	3666	874	1445			2381	3666	2381	3666	2381	3666
HNR	MC	2262	2256	2262	2256	2262	2256	2262	2256	2262	2256	2262	2256
HUNT TROMSO Cases	MC	492	470	492	470	492	470	651	637	651	636	651	636
HUNT TROMSO Controls	MC	566	546	566	547	566	546	749	718	748	718	748	717
IMPROVE	MC	1663	1775	1663	1775	1663	1775	1650	1776			1666	1783
KORAS3	MC	603	660	603	660	603	660	603	671	599	659	599	659
KORAS4	MC	585	624	585	624	585	624	585	632	582	624	582	624
LURIC Cases	MC	1652	567	1652	567	1652	567	1678	573	1678	574	1678	574
LURIC Controls	MC	327	310	327	310	327	310	330	315	330	315	330	315
MORGAM Cases	MC	984	222	983	222	983	222	1707	246	1708	246	1707	246
MORGAM Controls	MC	1084	287	1084	287	1084	287	2094	319	2095	320	2095	319
NSHD	MC	463	519	463	519	463	518	464	520	464	516	464	515
PIVUS	MC	485	483	485	483	485	483	490	488	490	488	490	488
SCARFSHEEP Cases	MC	886	309	886	310	893	311	893	311	716	259	882	310
SCARFSHEEP Controls	MC	1171	510	1173	512	1181	514	1181	514	944	433	1177	509
STR	MC	563	778	563	780	563	778	808	1377			838	1378
THISEAS Cases	MC	423	78	423	78	423	78	423	78	423	78	423	78
THISEAS Controls	MC	416	530	416	530	416	530	416	530	416	530	416	530

Table 2.5: Sample size of studies in the follow-up stage (continued).

		Hip		WC		WHR		Height		Weight		BMI	
Study	**Type**	**Men**	**Women**	**Men**	**Women**	**Men**	**Women**	**Men**	**Women**	**Men**	**Women**	**Men**	**Women**
ULSAM	MC	1095		1095		1095		1112		1116		1112	
Whitehall	MC	1700	535	1700	536	1700	535	1699	535	1700	536	1699	535

Using the joint association p-value combining discovery and follow-up results, we identified 205 SNPs out of the 619 that reached genome-wide significance ($p \leq 5 \times 10^{-8}$) (158 for height, 12 for weight, 15 for BMI, 2 for WC, 2 for WCadjBMI, 0 for HC, 0 for HCadjBMI, 0 for WHR, and 16 for WHRadjBMI). Of these 205 SNPs, 194 belonged to the set of 310 loci that would have been selected in a discovery stage overall analysis using the same p-value threshold as our discovery stage.

To evaluate sex-differences, we performed a sex-difference test on these 205 SNPs based only on follow-up data (to avoid selection bias similar to the winner's curse). This resulted in:

- Four SNPs with significant ($p \leq \frac{0.05}{205}$) sex difference (3 for WHRadjBMI, and 1 for WCadjBMI). Three of these four were previously reported as trait associations, all three of which would have been selected in an overall analysis, while the fourth is a novel sexually dimorphic trait association that would not have been selected in an overall discovery stage analysis ($p_{overall} = 1.3 \times 10^{-3}$).
- 14 loci with nominally significant ($p \leq 0.05$) sex-difference (4 for WHRadjBMI, 8 for height, and 2 for weight), 9 of which are near previously published trait association loci.
- 187 SNPs with no evidence ($p > 0.05$) of sex-difference, 157 of which were previously reported overall trait associations (potential evidence for overall trait association with the remaining 30 will be evaluated in future overall GIANT analyses).

All 18 SNPs identified with genome-wide significant association (see Table 2.6) and either significant or suggestive sex difference had consistent effect direction across the two stages (Figure 2.36).

				Discovery Stage		Follow-up Stage			Discovery + Follow-up	
SNP	Trait	Sex	Refs.	q_{men}	q_{women}	p_{men}	p_{women}	$p_{sexdiff}$	p_{men}	p_{women}
rs6717858	WHRadjBMI	F	[54]	1.00	1.52×10^{-10}	7.05×10^{-02}	7.27×10^{-16}	2.15×10^{-11}	0.61	1.99×10^{-29}
rs2820443	WHRadjBMI	F	[3, 54]	0.99	3.81×10^{-13}	0.94	1.83×10^{-20}	5.20×10^{-10}	0.37	4.62×10^{-37}
rs1358980	WHRadjBMI	F	[54]	0.97	4.75×10^{-09}	0.22	2.75×10^{-19}	9.06×10^{-08}	4.85×10^{-02}	2.41×10^{-31}
rs11743303	WCadjBMI	F		1.00	9.74×10^{-03}	0.34	1.43×10^{-06}	1.07×10^{-04}	0.57	2.69×10^{-11}
rs2371767	WHRadjBMI	F	[54]	0.99	1.65×10^{-04}	1.22×10^{-02}	1.71×10^{-16}	4.28×10^{-04}	8.34×10^{-03}	7.07×10^{-23}
rs2093210	HEIGHT	F	[70]	6.89×10^{-02}	8.58×10^{-07}	2.60×10^{-03}	1.03×10^{-15}	1.48×10^{-03}	4.07×10^{-07}	5.24×10^{-25}
rs10478424	WHRadjBMI	F		1.00	3.10×10^{-02}	0.27	7.61×10^{-05}	3.34×10^{-03}	0.76	3.45×10^{-09}
rs4684854	WHRadjBMI	F		1.00	2.27×10^{-04}	0.26	2.96×10^{-07}	8.44×10^{-03}	0.41	4.17×10^{-14}
rs6439167	HEIGHT	F	[70]	3.96×10^{-02}	4.09×10^{-03}	6.04×10^{-03}	9.68×10^{-12}	1.15×10^{-02}	4.59×10^{-07}	1.77×10^{-16}
rs7598759	HEIGHT	F		0.83	9.22×10^{-03}	2.91×10^{-02}	2.30×10^{-09}	1.36×10^{-02}	1.02×10^{-03}	4.48×10^{-14}
rs13192994	HEIGHT	F	[70]	0.11	1.20×10^{-03}	4.35×10^{-10}	8.95×10^{-04}	1.55×10^{-02}	7.76×10^{-13}	1.28×10^{-09}
rs2227901	HEIGHT	M	[70]	2.20×10^{-03}	0.31	2.79×10^{-07}	3.20×10^{-02}	1.84×10^{-02}	4.74×10^{-13}	6.10×10^{-05}
rs4646404	WHRadjBMI	F		0.86	1.31×10^{-02}	0.18	1.37×10^{-06}	2.26×10^{-02}	9.27×10^{-03}	2.18×10^{-11}
rs12812519	WEIGHT	M	[70]	2.52×10^{-02}	0.21	3.27×10^{-05}	0.41	3.21×10^{-02}	9.26×10^{-10}	5.03×10^{-04}
rs543874	WEIGHT	F	[55]	4.33×10^{-02}	1.71×10^{-07}	4.65×10^{-03}	7.44×10^{-08}	3.40×10^{-02}	3.41×10^{-07}	3.15×10^{-18}
rs2072153	HEIGHT	M	[70]	2.96×10^{-02}	0.14	4.73×10^{-05}	2.75×10^{-02}	4.11×10^{-02}	4.85×10^{-09}	9.27×10^{-06}
rs7320982	HEIGHT	M	[70]	4.60×10^{-03}	0.30	2.92×10^{-08}	1.75×10^{-03}	4.28×10^{-02}	1.82×10^{-13}	2.06×10^{-06}
rs13183458	HEIGHT	F		0.93	2.63×10^{-02}	0.68	7.30×10^{-04}	4.99×10^{-02}	8.87×10^{-02}	2.35×10^{-08}

Table 2.6: SNPs representing loci that showed evidence for sex difference in the follow-up stage. Eighteen SNPs with signfiicant ($p_{sexdiff} < 2 \times 10^{-4}$ in follow-up data) or suggestive ($p_{sexdiff} < 0.05$ in follow-up data) evidence of sex difference among the 205 SNPs showing significant association ($p_{men} < 5 \times 10^{-8}$ or $p_{women} < 5 \times 10^{-8}$) in combined analysis. Loci are sorted by $p_{sexdiff}$, and previously reported loci have a reference to the publication(s) in the "Refs." column.

Three of the four associations with significant sex difference were near (within $1Mb$ and LD $r^2 \geq 0.2$) previously described loci for WHRadjBMI and with previously established sex-difference (*2q24.3*, *1q41*, *6p21.1*)[3, 54]. The fourth locus (*5q11.2*: $p_{sexdiff} = 1.1 \times 10^{-4}$) is a novel association with WCadjBMI in women ($p_{women} = 2.7 \times 10^{-11}$), but not in men ($p_{men} = 0.6$).

Nine of the fourteen SNPs with suggestive sex difference were near (within 1 Mb and LD $r^2 \geq 0.2$) published hits associated with height, BMI, or WHRadjBMI[3, 54, 55, 70], but none were previously known to have sex-specific effects (*3p14.1*, *14q23.1*, *3q21.3*, *6q14.1*, *4q12*, *12q24.31*, *SEC16B*, *17q21.32-q21.33*, *13q31.3*). Five of the fourteen loci are novel associations with WHRadjBMI, WCadjBMI, or height (see Table 2.6).

Age-stratified analyses and association with other related phenotypes: In order to investigate whether any of the 18 identified sexually dimorphic associations would possibly be due to age-related effects (e.g. menopausal status in women), we re-analyzed the discovery studies stratified by age with a cut-off at 50 years and by sex. Among the 18 identified associations with established sex difference, we found no difference of the associations between the two age groups (Supplementary Table 7). To investigate potential enrichment of association with other related phenotypes, we (add OTHER PHENOTYPE data here).

2.5.2 Descriptions of the 18 sex-specific loci

The extents of each association region were determined by first taking all SNPs within a genetic distance of $1.0cM$ of the lead marker (based on HapMap phase 2[20] fine-scale recombination rate), then filtering out all SNPs with a p-value within 2 orders of magnitude of the lead marker ($p_{\text{SNP}} > p_{lead\ marker} \times 100$), and finally taking the positions of the first and last SNPs within that set to be the extents of the associated region. This yields a prediction of the associated region that is empirically based on the association signal we observed in our discovery stage data, rather than being

Figure 2.36: Effect estimates with 95% confidence intervals from follow-up data.

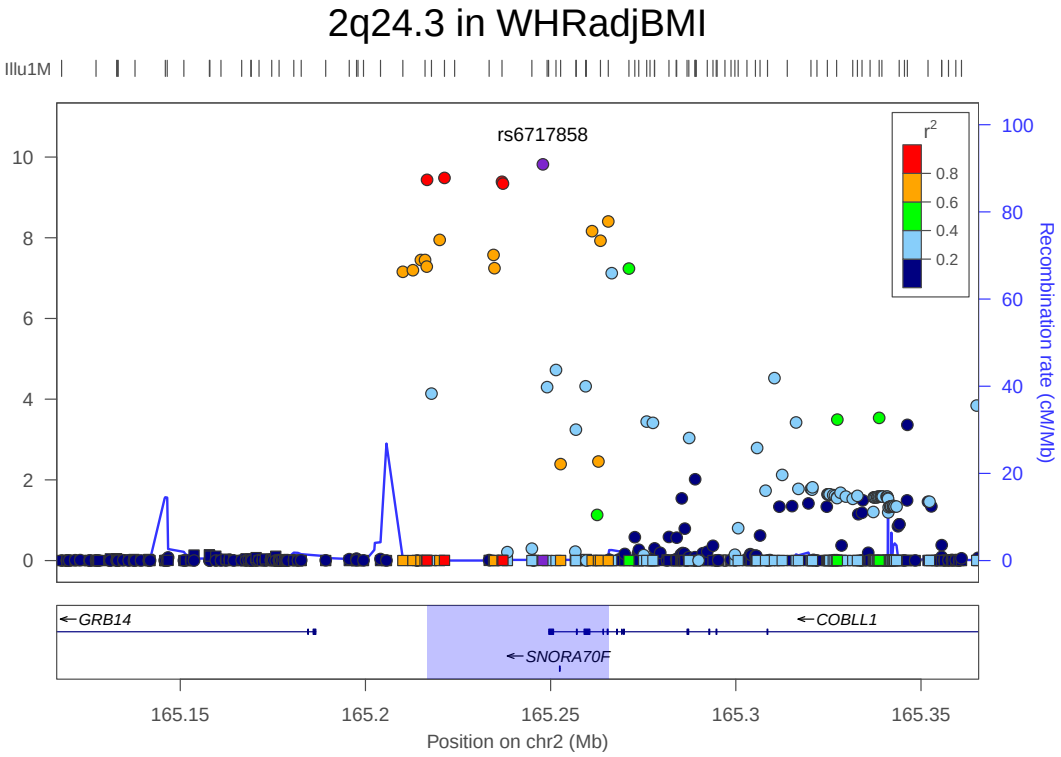
based more crudely on HapMap genetic distance or base position criteria.

In the locus association plots, each point represents a SNP in the region, with color to indicate LD (r^2) with the lead SNP. Point position along the y-axis indicates $-log_{10}(q)$ for association (left-hand scale). Underneath the points lies the HapMap recombination rate (right-hand scale) traced in blue. Genes from the UCSC genome browser RefSeq genes database[119] are shown as annotation tracks under the plot. The x-axis shows genomic position (in NCBI36 coordinates), and the region highlighted in blue is the associated region for the locus. The locus plots were produced using the standalone version of the LocusZoom software[120].

2.5.2.1 WHRadjBMI loci

2q24.3 The *2q24.3* locus is represented by lead marker rs6717858, with association signal for WHRadjBMI extending across $\approx 2kb$ of chromosome 49, ranging from $165216kb - 165265kb$ (see Figure 2.37). Two genes (*COBLL1* and *SNORA70F*) overlap this signal region, as does a previously reported SNP association with WHRadjBMI in tight LD with the lead marker (rs10195252: $\approx 26.5kb$ & $\approx 0.001cM$ from lead marker with $r^2 = 0.94$ & $D' = 1.0$)[55]. Heid et al. also presented eQTL data which suggested that *GRB14* expression was associated with rs10195252 genotype but *COBLL1* was not[55]. Our region lies $\approx 30 - 79kb$ upstream of *GRB14*, which is a member of a family os SH2-containing adaptors.

Figure 2.37: Discovery stage association at *2q24.3*.

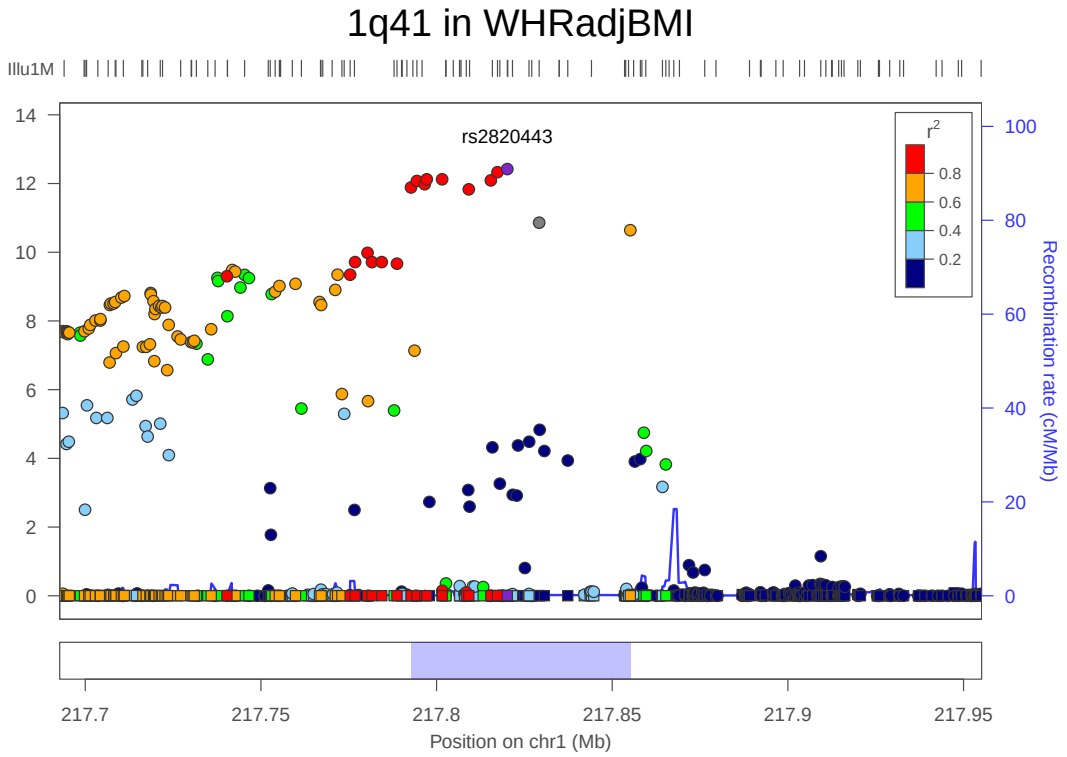


The *GRB14* protein binds directly to the insulin receptor[171, 172], and likely has an inhibitory effect on receptor tyrosine kinase signaling as well as on insulin receptor signaling, thereby regulating growth and metabolism. *Grb14* deficient mice exhibit enhanced body weight, mainly explained by increased lean mass on normal diet[144], improved glucose homeostasis despite lower circulating insulin levels, and enhanced insulin signaling in liver and skeletal muscle[173]. *Grb14* expression is increased in adipose tissue of insulin-resistant animal models and type 2 diabetic human patients[174], suggesting that *Grb14* may modulate insulin sensitivity. The WHR signal appears to be distinct from a *GRB14* locus previously reported as associated with both smoking initiation and current smoking (rs4423615: $\approx 101kb$ & $\approx 0.19cM$ from lead marker with $r^2 < 0.001$ & $D' = 0.01$)[175]. *COBLL1* may be involved in neural tube formation[169], is expressed at higher levels in tumors associated with good prognosis in mesothe-

lioma after surgery[245], and its knockdown led to increased apoptosis in both normal and tumor cells[246].

1q41 The *1q41* locus is represented by lead marker rs2820443, with association signal for WHRadjBMI extending across $\approx 1kb$ of chromosome 62, ranging from $217793kb - 217855kb$ (see Figure 2.38). One gene (*ZC3H11B*) overlaps this signal region, as does a previously reported SNP associated with WHRadjBMI (rs4846567: $\approx 2.8kb$ & $\approx 0.0002cM$ from lead marker with $r^2 = 0.96$ & $D' = 1.0$)[55]. *ZC3H11B* is a pseudogene with no known function. Another previously reported SNP association with WHRadjBMI lies $\approx 82 - 144kb$ from the association region (rs2605100: $\approx 109kb$ & $\approx 0.04cM$ from lead marker with $r^2 = 0.68$ & $D' = 0.84$)[3]. Excluding *ZC3H11B*, the signal region is nearest to *SLC30A10* ($\approx 299 - 361kb$ downstream) and *LYPLAL1* ($\approx 340 - 402kb$ downstream). *SLC30A10* belongs to a family of membrane transporters involved in intracellular zinc homeostasis and is expressed in brain and liver[165]. *LYPLAL1* encodes lysophospholipase-like 1 protein, which is thought to act as a triglyceride lipase and is reported to be up-regulated in subcutaneous adipose tissue of obese subjects[121]. Intergenic variants near *LYPLAL1* have also been associated with fatty liver disease (rs12137855: $\approx 305kb$ & $\approx 0.22cM$ from lead marker with $r^2 = 0.1$ & $D' = 0.40$)[247].

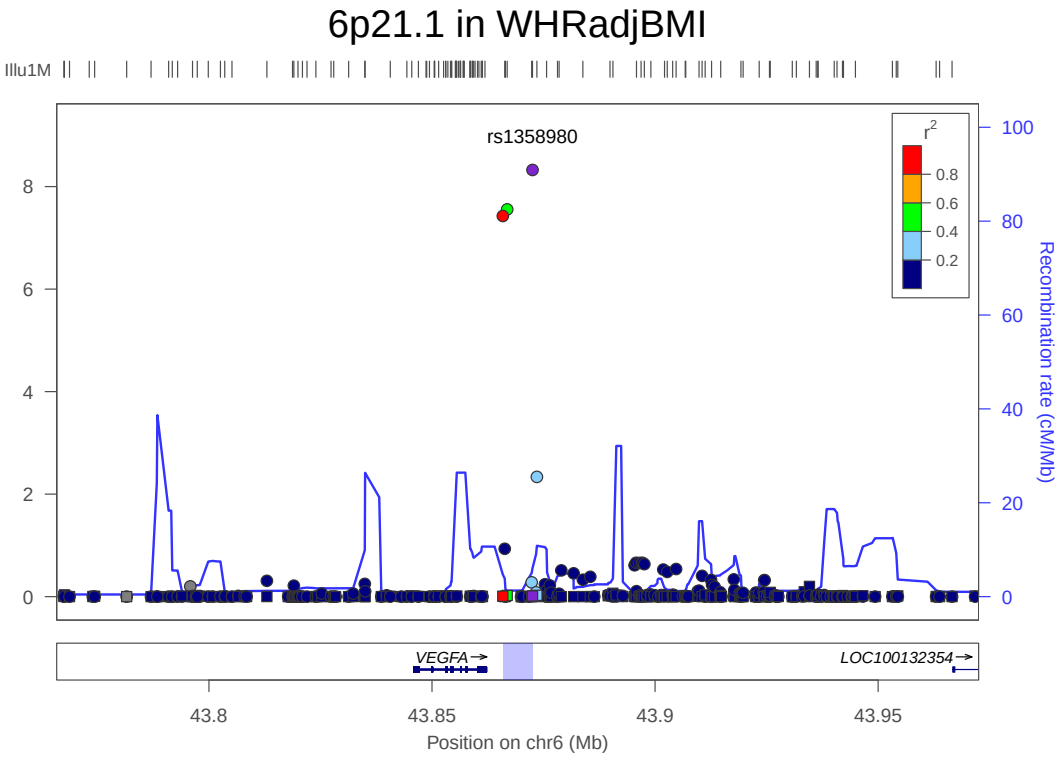
Figure 2.38: Discovery stage association at *1q41*.



6p21.1 The *6p21.1* locus is represented by lead marker rs1358980, with association signal for WHRadjBMI extending across $\approx 6kb$ of chromosome 7, ranging from $43872kb - 43865kb$ (see Figure 2.39). No genes overlap this signal region, but it does include a SNP previously associated with WHRadjBMI (rs6905288: $\approx 5.6kb$ & $\approx 0.01cM$ from lead marker with $r^2 = 0.5$ & $D' = 0.91$)[54]. The associated region is located $\approx 3.7 - 10.7kb$ downstream of *VEGFA*. Multiple variants and mutations in *VEGFA* are risk factors for diabetic retinopathy[166, 189, 190], and variants in *VEGFA* have been nominally associated with T2D[68]. *VEGFA* is proposed as a key mediator of adipogenesis and angiogenesis[191], is highly expressed in adipose tissue, and has increased expression during adipocyte differentiation[192–195]. *VEGFA* serum concentrations are elevated in overweight and obese patients compared with lean subjects[196] and decrease after weight loss following bariatric surgery, behaving similarly to other hormones related to

adipose mass, such as leptin and insulin[197]. Variants near *VEGFA* have also been associated with kidney function (rs881858: $\approx 42kb$ & $\approx 0.2cM$ from lead marker with $r^2 = 0.01$ & $D' = 0.18$)[248] and advanced age related macular degeneration (rs4711751: $\approx 64kb$ & $\approx 0.2cM$ from lead marker with $r^2 = 0.04$ & $D' = 0.21$ in 1000G data)[249] although both appear likely to be distinct from the signal at this locus.

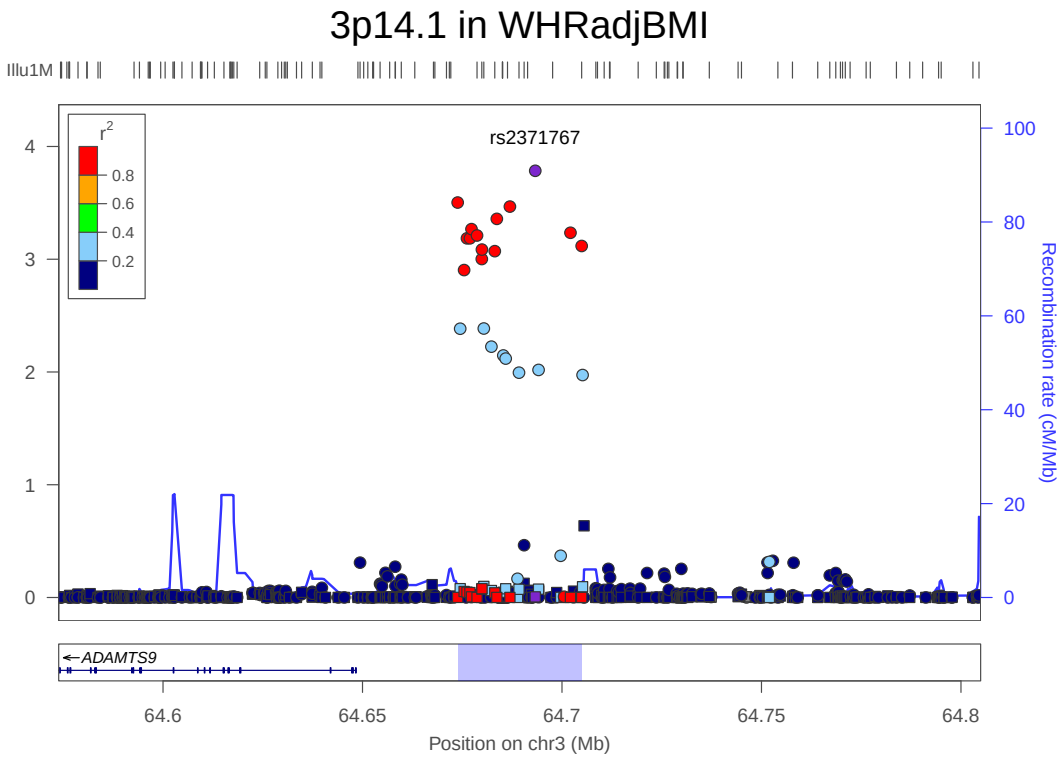
Figure 2.39: Discovery stage association at *6p21.1*.



3p14.1 The *3p14.1* locus is represented by lead marker rs2371767, with association signal for WHRadjBMI extending across $\approx 3kb$ of chromosome 31, ranging from $64704kb - 64673kb$ (see Figure 2.40). Two genes (*ADAMTS9-AS2* and *MIR548AN*) overlap this signal region, as do three previously reported SNP associations, one with WHRadjBMI (rs6795735: $\approx 12.9kb$ & $\approx 0.004cM$ from lead marker with $r^2 = 0.311$ & $D' = 1.0$)[54], and two with T2D (rs4607103: $\approx 6.4kb$ & $\approx 0.001cM$ from lead marker with $r^2 = 0.90$ & $D' = 1.0$ and rs4411878:

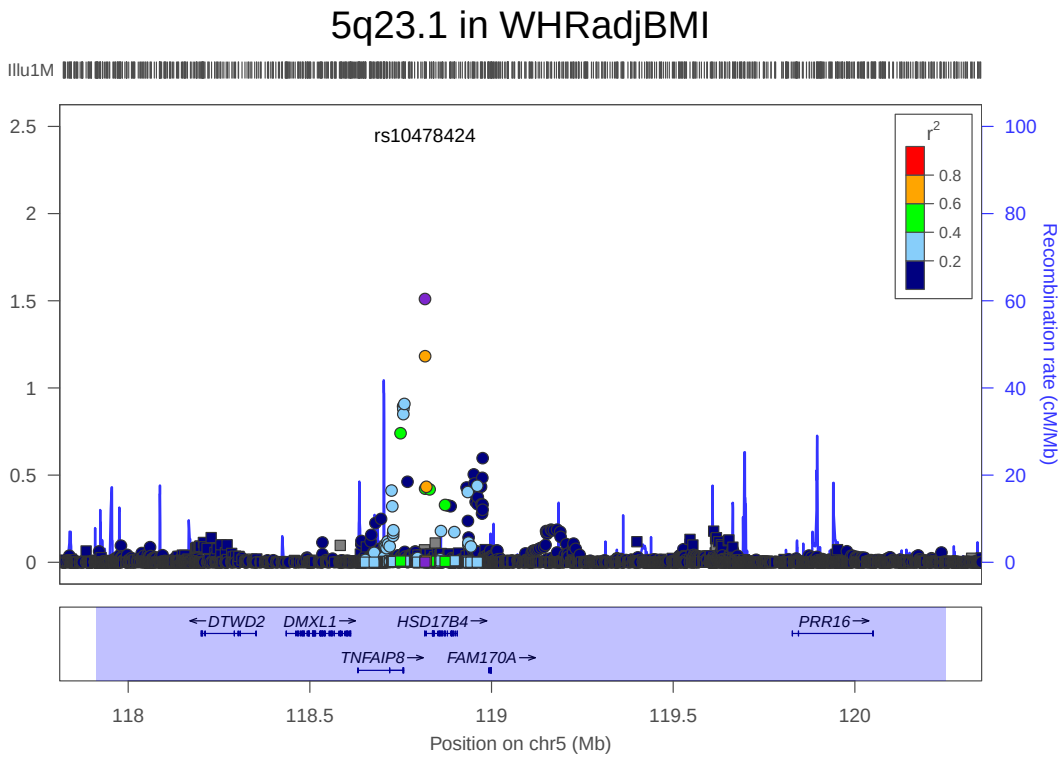
$\approx 14.6kb$ & $\approx 0.005cM$ from lead marker with $r^2 = 0.85$ & $D' = 0.95$)[68, 250]. The T2D association is possibly mediated through decreased insulin sensitivity of peripheral tissues[209]. *MIR548AN* is a μRNA which primarily maps to the X chromosome, but the full length μRNA precursor sequence also maps with 96.4% identity to five fragments across a $\approx 240kb$ window within our signal region with only 3 base mismatches). The function of *MIR548AN* is not known. *ADAMTS9-AS2* is a long non-coding RNA transcript which is an antisense for *ADAMTS9* and which also contains the fragments of *MIR548AN*. This region is located $\approx 25 - 56kb$ upstream of *ADAMTS9*. *ADAMTS9* is a member of the *ADAMTS* family, a group of genes encoding metalloproteases that lack transmembrane domains and are secreted into the extracellular matrix[183]. Members of the *ADAMTS* family have been implicated in control of organ shape during development and inhibition of angiogenesis[123].

Figure 2.40: Discovery stage association at *3p14.1*.



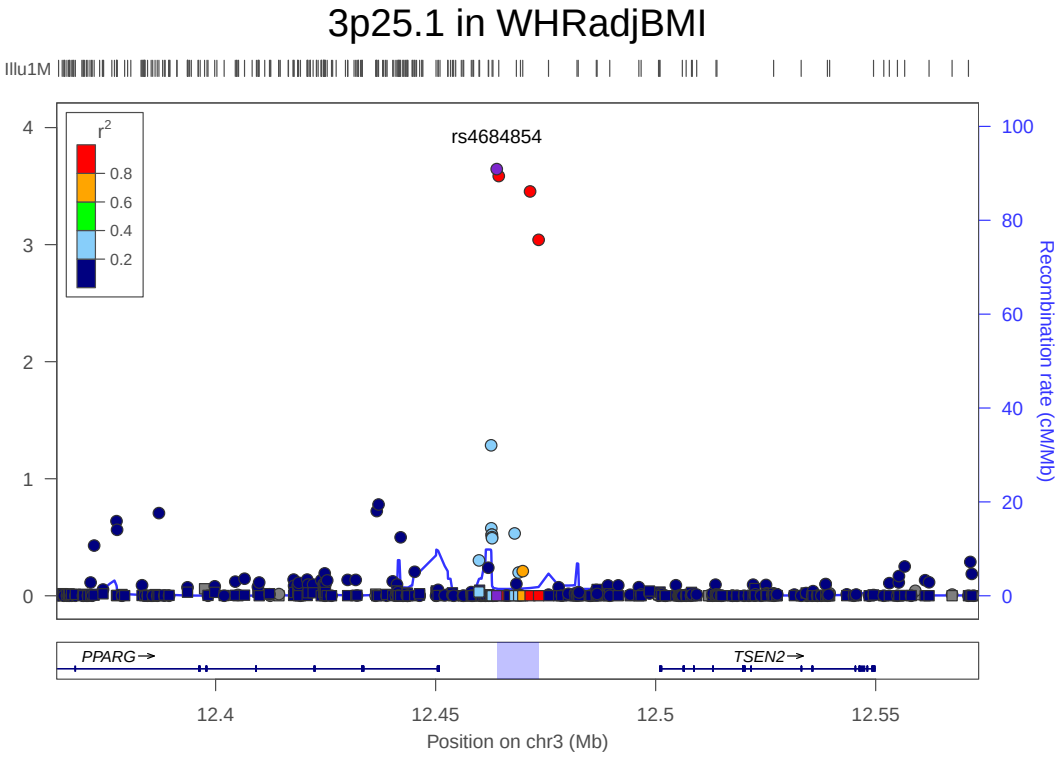
5q23.1 The *5q23.1* locus is represented by lead marker rs10478424, with association signal for WHRadjBMI extending across $\approx 5kb$ of chromosome 2337, ranging from $117911kb - 120249kb$ (see Figure 2.41). Nine genes (*DMXL1*, *DTWD2*, *FAM170A*, *HSD17B4*, *MIR1244-1*, *MIR1244-2*, *MIR1244-3*, *PRR16*, and *TNFAIP8*) overlap this signal region. The lead marker, rs10478424, is located in an intronic region of hydroxysteroid (17-beta) dehydrogenase 4 (*HSD17B4)*. The protein encoded by *HSD17B4* is a bifunctional enzyme that is involved in the peroxisomal beta-oxidation pathway for fatty acids. Mutations in this gene are known to cause D-bifunctional protein (D-BP) deficiency, an autosomal-recessive disorder of peroxisomal fatty acid beta-oxidation that is generally fatal within the first two years of life[251, 252]. Expression levels of *HSD17B4* have been associated with prostate cancer severity[253], and it is also a significant independent predictor of poor patient outcome[254].

Figure 2.41: Discovery stage association at *5q23.1*.



3p25.1 The *3p25.1* locus is represented by lead marker rs4684854, with association signal for WHRadjBMI extending across $\approx 3kb$ of chromosome 10, ranging from $12463kb - 12473kb$ (see Figure 2.42). No known genes overlap this region, although structural variants including a CNV[255] and an inversion[256] have been reported that overlap this region. However, it does lie $\approx 13kb$ downstream of the well known T2D susceptibility gene peroxisome proliferator-activated receptor gamma (*PPARG)*, although the T2D associated SNP appears to be distinct from this locus (rs1801282: $\approx 96kb$ & $\approx 0.11cM$ from lead marker with $r^2 = 0.04$ & $D' = 0.61$)[257–259]. The protein product encoded by *PPARG* is $PPAR - \gamma$ which is a regulator of adipocyte differentiation. Additionally, $PPAR - \gamma$ has been implicated in the pathology of numerous diseases including obesity[260–263], diabetes[264, 265], atherosclerosis[266] and cancer[264, 267–269]. Interestingly, previous studies of the Pro12Ala polymorphism in *PPARG* have demonstrated genotypeâĂŞby-sex interaction with BMI[270], fatty acid concentrations during the first 24h after birth were related to *PPARG* expression in female but not in male lambs[271], and female 12Ala mutation carriers had greater risk of developing abdominal obesity than female non-carriers while male 12Ala mutation carriers had no significant increase in risk[272].

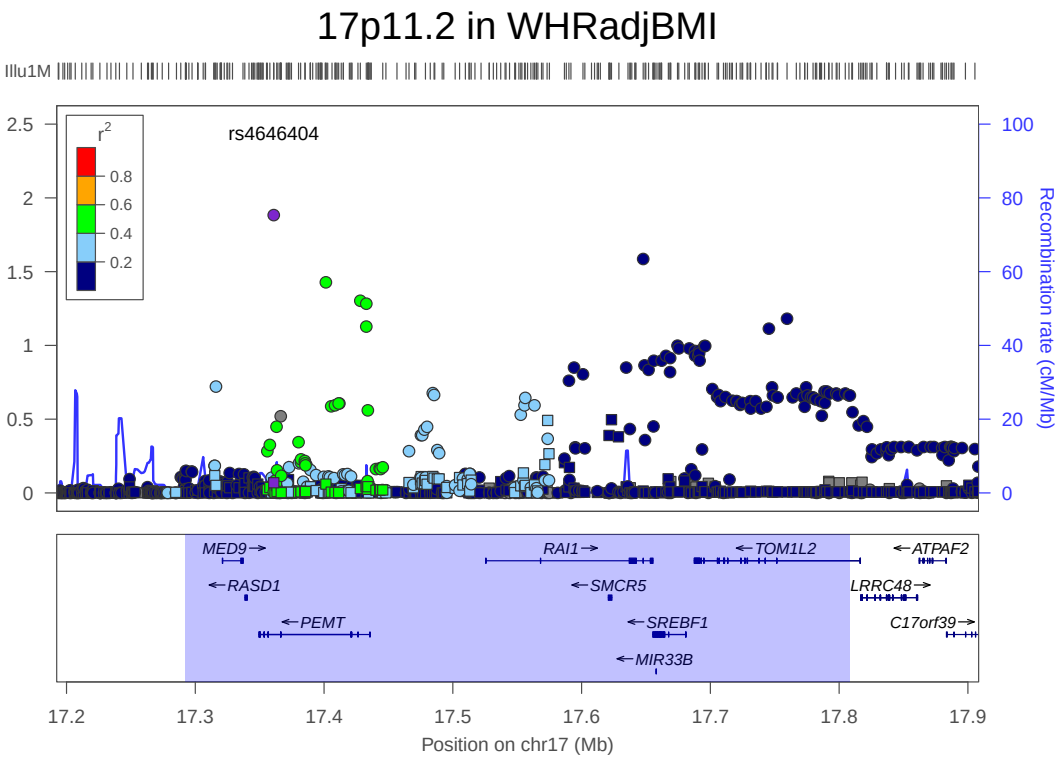
Figure 2.42: Discovery stage association at *3p25.1*.



17p11.2 The *17p11.2* locus is represented by lead marker rs4646404, with association signal for WHRadjBMI extending across $\approx 17kb$ of chromosome 516, ranging from $17292kb - 17808kb$ (see Figure 2.43). Eight genes (*MED9*, *MIR33B*, *PEMT*, *RAI1*, *RASD1*, *SMCR5*, *SREBF1*, and *TOM1L2*) overlap this signal region, as does a SNP previously reported to be associated with CAD (rs12936587: $\approx 124kb$ & $\approx 0.07cM$ from lead marker with $r^2 = 0.34$ & $D' = 0.81$)[273]. Also within the signal region are two rare SNPs clinically linked to Smith-Magenis syndrome (SMS) in patients without the chromosome 17 deletion found in most SMS patients (rs104894633: $\approx 281.5kb$ from lead marker and rs104894634: $\approx 280.7kb$ from lead marker; no LD information available)[274]. SMS is a syndrome consisting of multiple congenital anomalies, including distinctive craniofacial features, skeletal features such as short stature, fatty liver, delayed speech and language skills, sleep disturbances, behavioral problems, cognitive impairment,

and mental retardation.

Figure 2.43: Discovery stage association at *17p11.2*.



SMS is thought to be caused in most cases by the deletion of a $\approx 3.5mb$ region of 17p11.2, encompassing many genes[275]. In patients without a deletion, mutations within the retinoic acid induced 1 (*RAI1)* gene (including rs104894633 and rs104894634) can cause many of the SMS phenotypes, though some phenotypes such as short stature, hearing loss, speech and motor delay, and cardiovascular anomolies appear to not be associated with *RAI1* mutations but only with deletion of other genes in the region[275]. Analysis comparing phenotypes in patients with variously sized deletions covering different genes in the region suggest that the region contributing to short stature is located between partially within the associated region, overlapping the genes sterol regulatory element binding transcription factor 1 (*SREBF1)* and target of myb1-like 2 (chicken) (*TOM1L2)*[275]. Our lead marker is located within phosphatidylethanolamine

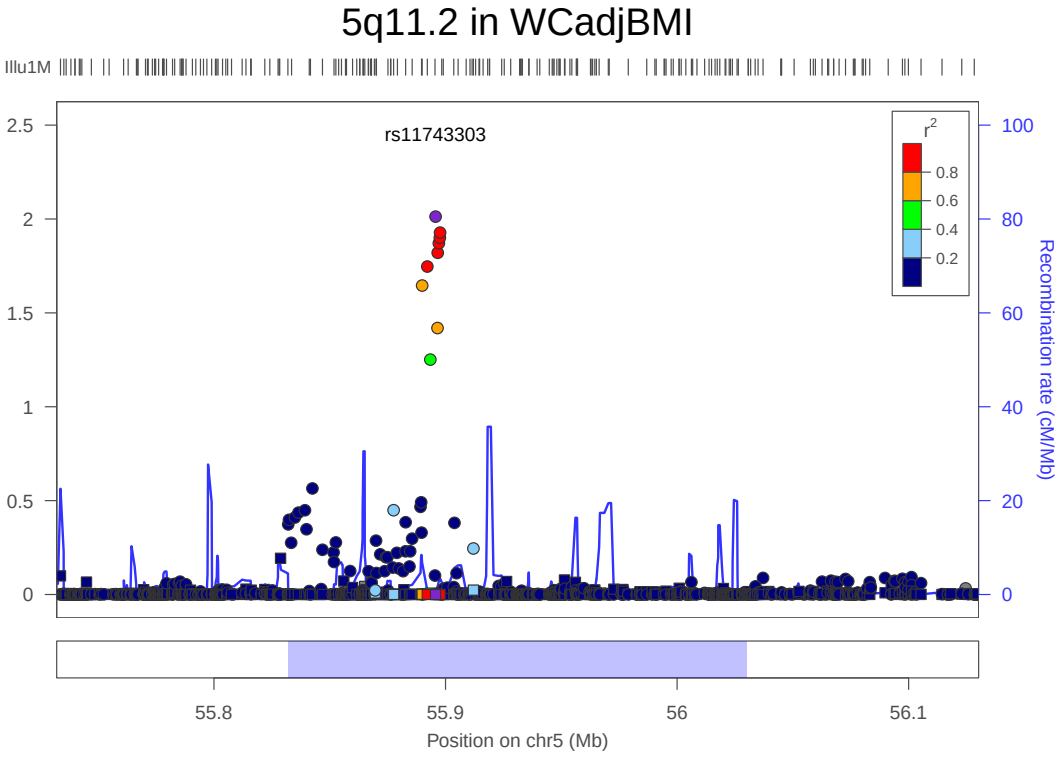
N-methyltransferase (*PEMT)*, which encodes an enzyme which converts phosphatidylethanolamine to phosphatidylcholine by sequential methylation in the liver. The protein localizes to the endoplasmic reticulum and mitochondria associated membranes[123] and is thought to be responsible for the fatty liver phenotype sometimes seen in SMS[275]. microRNA 33b (*MIR33B)* is an μRNA that has been found to be involved in control of cholesterol homeostasis[276], while mediator complex subunit 9 (*MED9)* is a member of the mediator complex of proteins that promote activation of RNA polymerase II through direct interactions with transcription factors[123].

2.5.2.2 WCadjBMI loci

5q11.2 The *5q11.2* locus is represented by lead marker rs11743303, with association signal for WCadjBMI extending across $\approx 5kb$ of chromosome 198, ranging from $55832kb - 56030kb$ (see Figure 2.44). No known genes overlap this region, but it does overlap with another SNP (rs6867983: $\approx 5.8kb$ & $\approx 0.02cM$ from lead marker with $r^2 = 0.71$ & $D' = 1.0$) reported as a suggestive association with TG[277]. The associated region also lies $\approx 116 - 314kb$ upstream of mitogen-activated protein kinase kinase kinase 1 (*MAP3K1)*, a serine/threonine kinase that occupies a pivotal role in a network of phosphorylating enzymes integrating cellular responses to a number of mitogenic and metabolic stimuli, including insulin (MIM 176730) and many growth factors[123]. Mutations in *MAP3K1* are associated with gonadal dysgenesis[278], and a SNP within *MAP3K1* (rs889312: $\approx 172kb$ & $\approx 0.44cM$ from lead marker with $r^2 = 0.005$ & $D' = 0.10$) has been reported to be associated with breast cancer[279, 280], possibly as a gene-gene interaction with breast cancer 2, early onset (*BRCA2)*[281]. It is also $\approx 384 - 582kb$ upstream of ankyrin repeat domain 55 (*ANKRD55)*, which harbors SNPs reported to be associated with longevity (rs415407: $\approx 445kb$ & $\approx 1.1cM$ from lead marker with $r^2 = 0.005$ & $D' = 0.12$)[282], Rheumatoid Arthritis

(RA) (rs6859219: $\approx 421kb$ & $\approx 0.93cM$ from lead marker with $r^2 = 0.008$ & $D' = 0.10$)[283], and Celiac Disease (CD) (rs1020388: $\approx 300kb$ & $\approx 0.71cM$ from lead marker with $r^2 = 0.003$ & $D' = 0.11$)[284].

Figure 2.44: Discovery stage association at *5q11.2*.

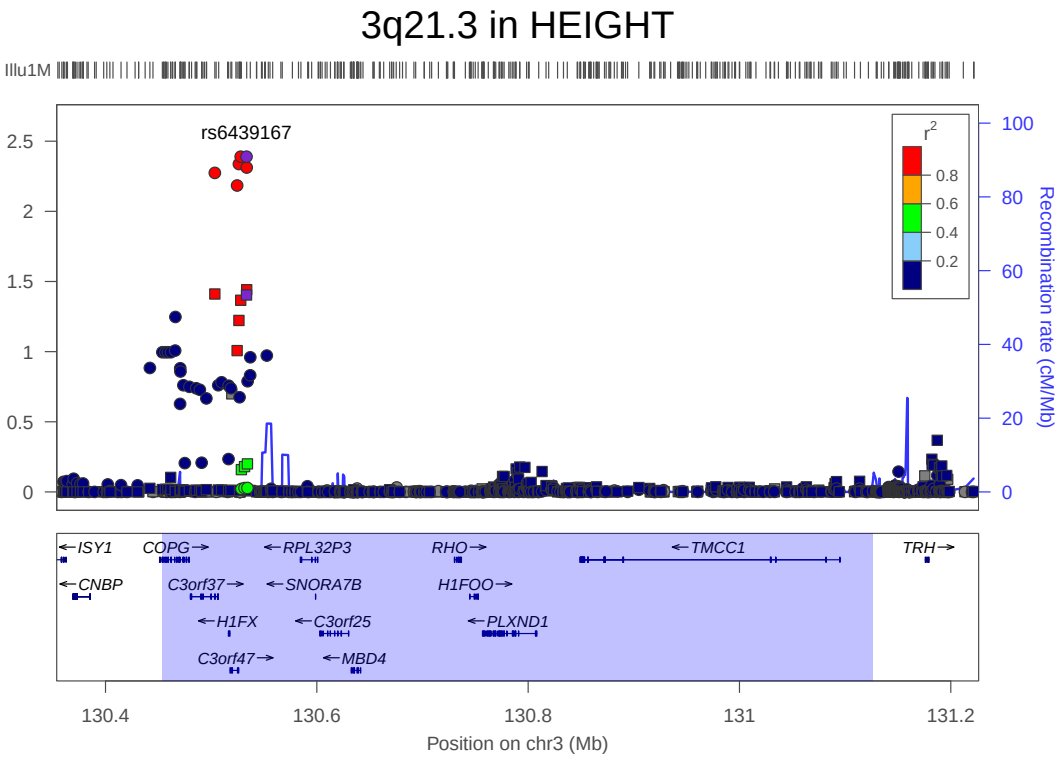


2.5.2.3 Height loci

3q21.3 The *3q21.3* locus is represented by lead marker rs6439167, with association signal for Height extending across $\approx 3kb$ of chromosome 673, ranging from $130454kb - 131126kb$ (see Figure 2.45). Thirteen genes (*C3orf25*, *C3orf37*, *COPG*, *H1FOO*, *H1FX*, *H1FX-AS1*, *IFT122*, *MBD4*, *PLXND1*, *RHO*, *RPL32P3*, *SNORA7B*, and *TMCC1*) overlap this signal region. The lead marker has been previously reported as associated with height[70] and is located $\approx 5.7kb$ downstream of H1FX antisense RNA 1 (non-protein coding) (*H1FX-AS1)* and $\approx 15.6kb$ upstream of H1 histone family, member X (*H1FX)*. *H1FX* is a gene that encodes

a member of the histone H1 family. Histones are basic nuclear proteins that are responsible for the nucleosome structure of the choromosomal fiber in eukaryotes. Another gene, coatomer protein complex, subunit gamma (*COPG*) is $\approx 54kb$ from the lead marker and is associated with anorexia nervosa[285].

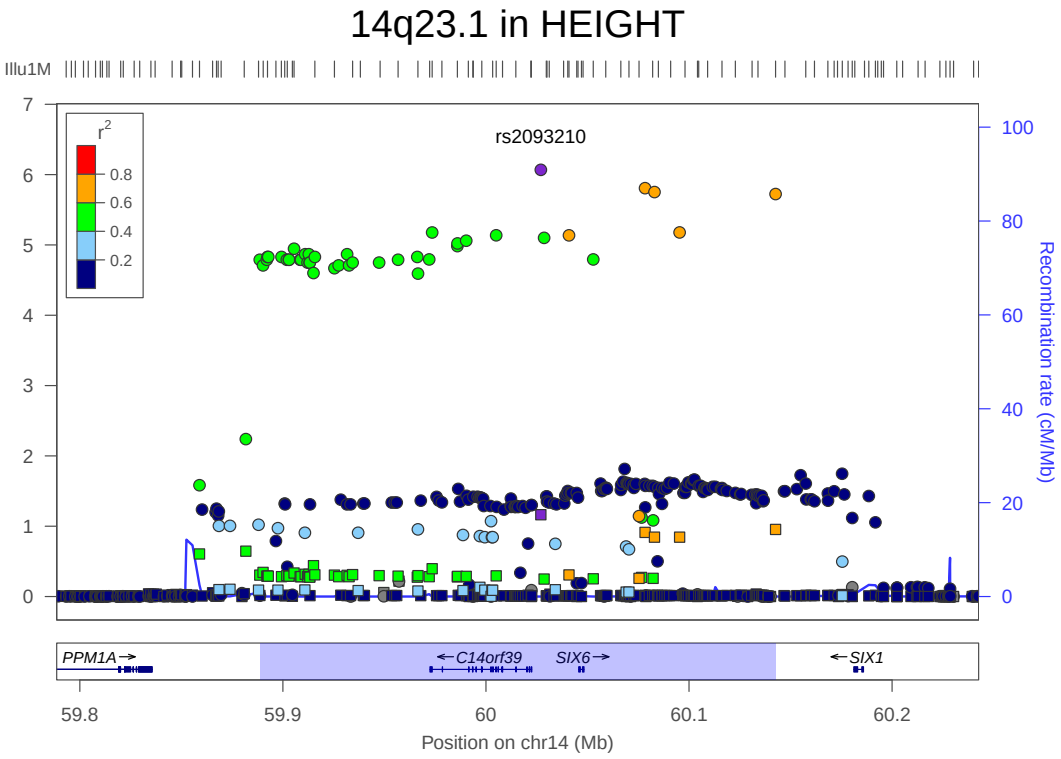
Figure 2.45: Discovery stage association at *3q21.3*.



14q23.1 The *14q23.1* locus is represented by lead marker rs2093210, with association signal for Height extending across $\approx 14kb$ of chromosome 254, ranging from $59888kb - 60142kb$ (see Figure 2.46). Two genes (*C14orf39* and *SIX6*) overlap this region, as does a SNP associated with optic disc size and cup/disc ratio (rs10483727: $\approx 116kb$ & $\approx 0.01cM$ from lead marker with $r^2 = 0.80$ & $D' = 0.96$)[286, 287] and an inversion that does not overlap any known genes[256]. The lead marker was also previously reported to be associated with height[70], and is $\approx 4.5kb$ upstream of chromosome 14 open reading frame 39 (*C14orf39*) and $\approx 18.6kb$ upstream of SIX homeobox 6 (*SIX6*). *C14orf39* is

a protein-coding gene on the opposite strand from *SIX6* and encodes the protein SIX6OS1, which is of unknown function[123]. The *SIX6* gene codes for a homeobox protein and is thought to be involved in eye development. Defects in this gene are a cause of isolated microphthalmia with cataract type 2 (MCOPCT2)[288]. Mutations in this gene have also been associated with anophthalmia[289] and pituitary abnormalities[290]. Our associated region is also located $\approx 53-307kb$ downstream from protein phosphatase, Mg2+/Mn2+ dependent, 1A (*PPM1A)*, a gene involved in TGF beta signaling[291].

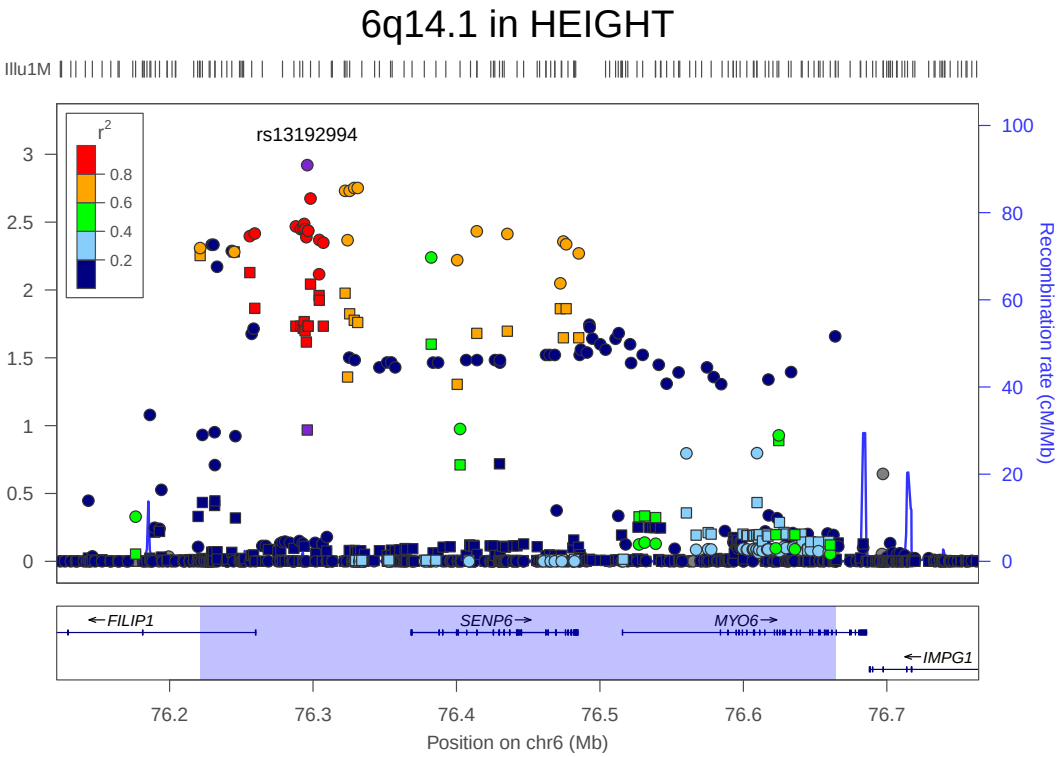
Figure 2.46: Discovery stage association at *14q23.1*.



6q14.1 The *6q14.1* locus is represented by lead marker rs13192994, with association signal for Height extending across $\approx 6kb$ of chromosome 443, ranging from $76221kb - 76664kb$ (see Figure 2.47). Three genes (*FILIP1*, *MYO6*, and *SENP6*) overlap this signal region, as does a previously reported association with height (rs9360921: $\approx$ 26*kb*, 0.003cM, r2=0.628, D'=0.92 with lead marker)[70]. The

lead marker is $\approx 36kb$ upstream of filamin A interacting protein 1 (*FILIP1*), which controls the start of neocortical cell migration from the ventricular zone and is moderately expressed in adult heart and brain. The lead marker is $\approx 72kb$ upstream of SUMO1/sentrin specific peptidase 6 (*SENP6*), which encodes a SUMO-1-specific protease and was found to be highly expressed in human reproductive organs[292]. The lead marker is also $\approx 220kb$ upstream of myosin VI (*MYO6*), mutations in which have been found in patients with non-syndromic autosomal dominant recessive hearing loss[293, 294].

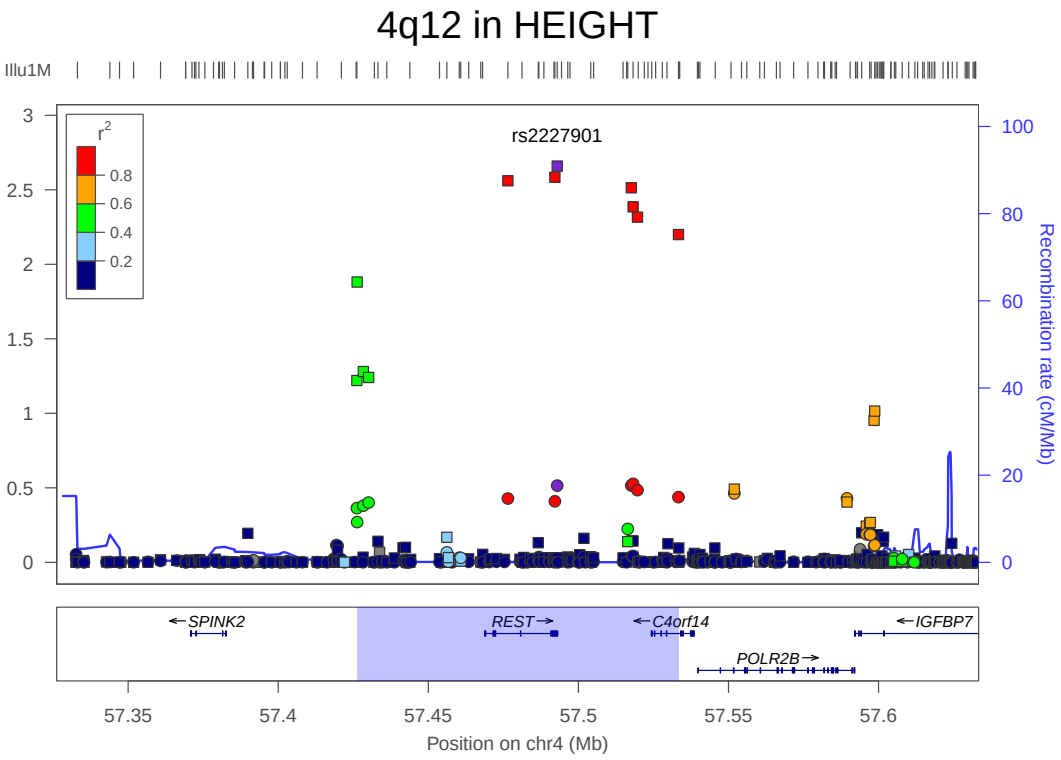
Figure 2.47: Discovery stage association at *6q14.1*.



4q12 The *4q12* locus is represented by lead marker rs2227901, with association signal for Height extending across $\approx 4kb$ of chromosome 107, ranging from $57426kb - 57533kb$ (see Figure 2.48). Two genes (*C4orf14* and *REST*) overlap this signal region, as do SNPs associated with bleomycin (BLM) sensitivity (rs708547: $\approx 23kb$ & $\approx 0.0006cM$ from lead marker with $r^2 = 0.08$ & $D' = 1.0$)[295] and

height (rs17081935: $\approx 25kb$ & $\approx 0.0007cM$ from lead marker with $r^2 = 1.0$ & $D' = 1.0$)[70]. The lead marker is a synonymous polymorphism within the RE1-silencing transcription factor (*REST*) gene. This gene encodes a transcriptional repressor that represses neuronal genes in non-neuronal tissues. *REST* has been associated with mutagen sensitivity[295], with transcriptional and epigenetic dysregulation of Huntington's disease[296], and with breast cancer[297].

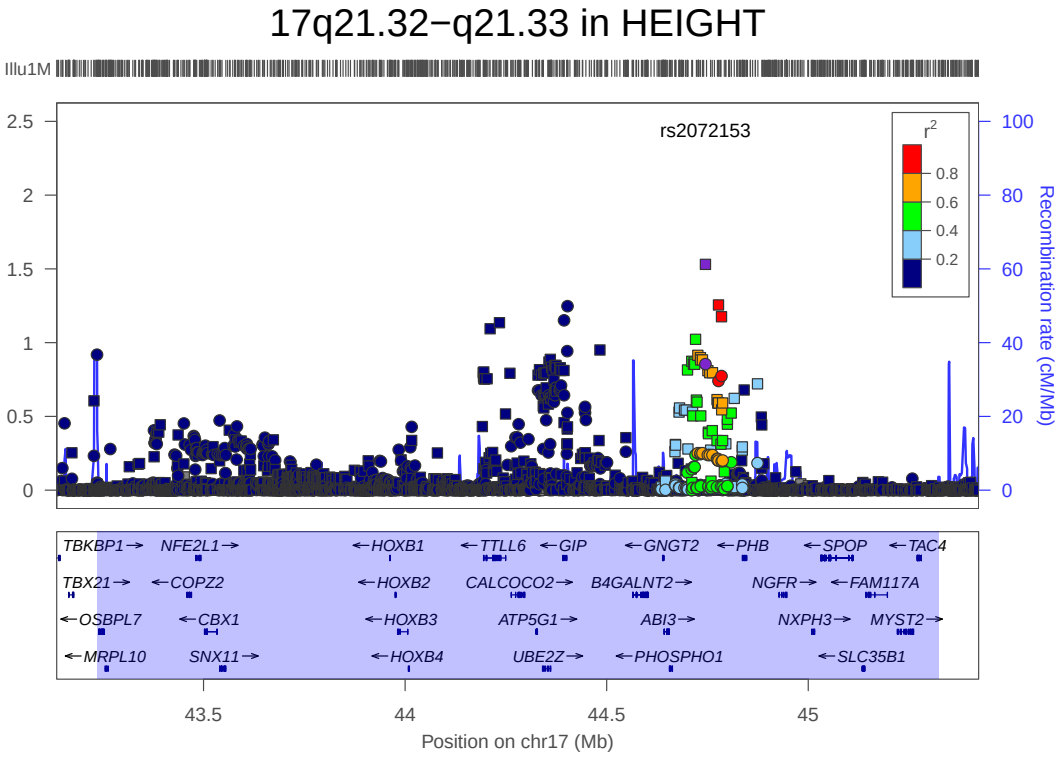
Figure 2.48: Discovery stage association at *4q12*.



17q21.32-q21.33 The *17q21.32-q21.33* locus is represented by lead marker rs2072153, with association signal for Height extending across $\approx 17kb$ of chromosome 2087, ranging from $43235kb - 45322kb$ (see Figure 2.49). Fifty-one genes (*ABI3*, *ATP5G1*, *B4GALNT2*, *CALCOCO2*, *CBX1*, *CDK5RAP3*, *COPZ2*, *FAM117A*, *GIP*, *GNGT2*, *HOXB1*, *HOXB13*, *HOXB13-AS1*, *HOXB2*, *HOXB3*, *HOXB4*, *HOXB5*, *HOXB6*, *HOXB7*, *HOXB8*, *HOXB9*, *IGF2BP1*, *KAT7*, *LRRC46*, *MIR10A*, *MIR1203*, *MIR152*, *MIR196A1*, *MIR3185*, *MRPL10*, *NFE2L1*, *NGFR*, *NXPH3*, *OSBPL7*, *PHB*, *PHOS-*

PHO1, *PNPO*, *PRAC*, *PRR15L*, *SCRN2*, *SKAP1*, *SLC35B1*, *SNF8*, *SNX11*, *SP2*, *SP6*, *SPOP*, *TAC4*, *TTLL6*, *UBE2Z*, and *ZNF652*) overlap this signal region, as do SNPs associated with BMI (rs3764400: $\approx 1266kb$ from lead marker)[55], bipolar disorder (rs1035050: $\approx 174kb$ & $\approx 0.27cM$ from lead marker with $r^2 = 0.05$ & $D' = 0.44$)[177], cognitive performance (rs2326017: $\approx 669kb$ from lead marker)[298], coronary heart disease (rs46522: $\approx 401kb$ & $\approx 0.51cM$ from lead marker with $r^2 = 0.008$ & $D' = 0.14$)[273], diastolic blood pressure (rs16948048: $\approx 50kb$ & $\approx 0.004cM$ from lead marker with $r^2 = 0.24$ & $D' = 0.87$)[299, 300], ovarian cancer (rs2084881: $\approx 1033kb$; rs9303542: $\approx 979kb$ from lead marker)[301], primary tooth development (rs6504340: $\approx 773kb$ from lead marker; rs9674544: $\approx 305kb$ & $\approx 0.42cM$ from lead marker with $r^2 = 0.01$ & $D' = 0.15$)[302], and prostate cancer (rs7210100: $\approx 47kb$ & $\approx 0.003cM$ from lead marker with $r^2 = 0.002$ & $D' = 1.0$)[303]. In addition, the lead marker has been previously reported as an association with height in a GWAMA[70]. The lead marker is located within an intronic region of zinc finger protein 652 (*ZNF652)*, a DNA binding transcription factor[304]. Gene expression studies have shown that a SNP in modest LD with the lead marker (rs8064621: $\approx 25kb$ & $\approx 0.002cM$ from lead marker with $r^2 = 0.58$ & $D' = 1.0$) is strongly associated with *ZNF652* expression in omental and subcutaneous adipose tissue[70].

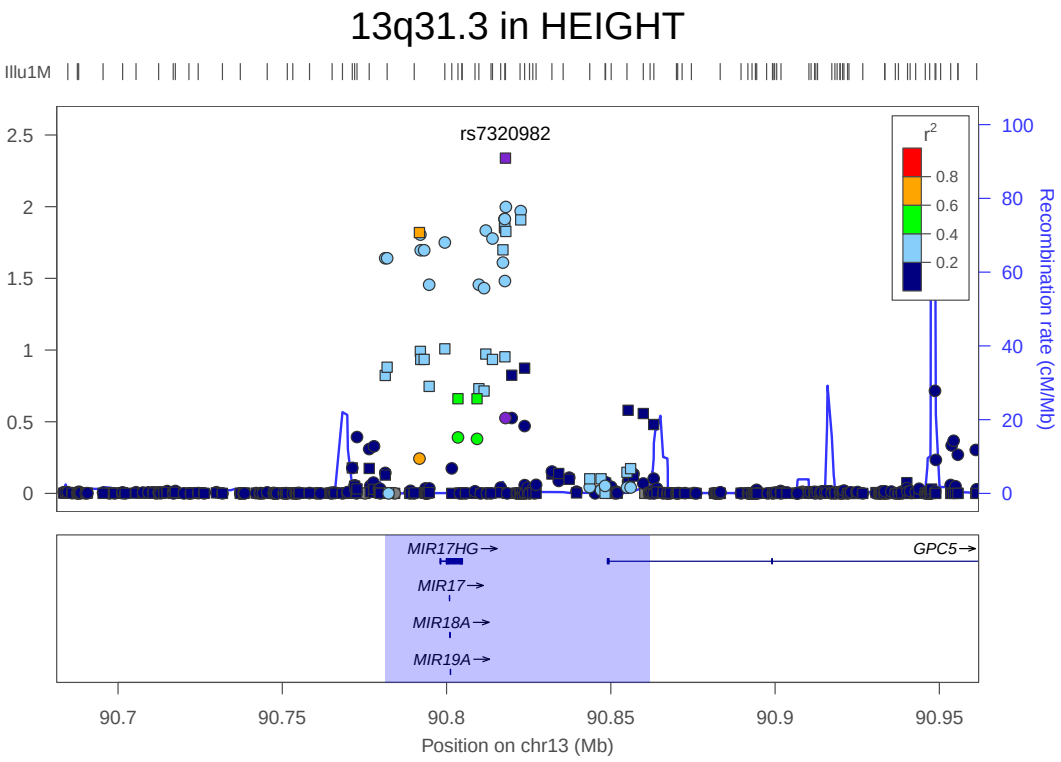
Figure 2.49: Discovery stage association at *17q21.32-q21.33*.



13q31.3 The *13q31.3* locus is represented by lead marker rs7320982, with association signal for Height extending across $\approx 13kb$ of chromosome 81, ranging from $90781kb - 90862kb$ (see Figure 2.50). Eight genes (*GPC5*, *MIR17*, *MIR17HG*, *MIR18A*, *MIR19A*, *MIR20A*, *MIR19B1*, and *MIR92A1*) overlap this signal region, as do SNPs previously associated with height (rs7319045: $\approx 4kb$ & $\approx 0.002cM$ from lead marker with $r^2 = 0.395$ & $D' = 1.0$;)[70, 244, 305]. The nearest gene to the lead marker is miR-17-92 cluster host gene (non-protein coding) (*MIR17HG)*, located $\approx$ $13kb$ downstream ($\approx$ $0.004cM$) from the lead marker. This gene encodes a polycistronic primary transcript containing several mature μRNAs, including the microRNA 17 (*MIR17)*, microRNA 18a (*MIR18A)*, microRNA 19a (*MIR19A)*, microRNA 20a (*MIR20A)*, microRNA 19b-1 (*MIR19B1)*, and microRNA 92a-1 (*MIR92A1)* genes. These μRNAs are associated with clinical outcomes in cancer and as an oncogenic/tumor-suppressive regulator[306–312].

Expression of one of the mature μRNA, *MIR18A*, was found to be significantly increased in female over male hepatocellular carcinoma (HCC) tissues[313]. Only slightly farther from the lead marker ($\approx 31kb$ and $\approx 0.012cM$) and also within the associated region is glypican 5 (*GPC5)*, a cell surface heparan sulfate proteoglycan that may play a role in the control of cell division and growth regulation. In addition to the height variants listed above, variants near *GPC5* have also been previously reported as associated with acquired nephrotic syndrome (rs16946160: $\approx 184kb$ & $\approx 0.36cM$ from lead marker with $r^2 = 0.006$ & $D' = 0.52$)[314], multiple sclerosis (rs9523762: $\approx 1312kb$ from lead marker)[315, 316], and lung cancer risk (rs2352028: $\approx 425kb$ & $\approx 0.72cM$ from lead marker with $r^2 = 0.006$ & $D' = 0.29$)[317].

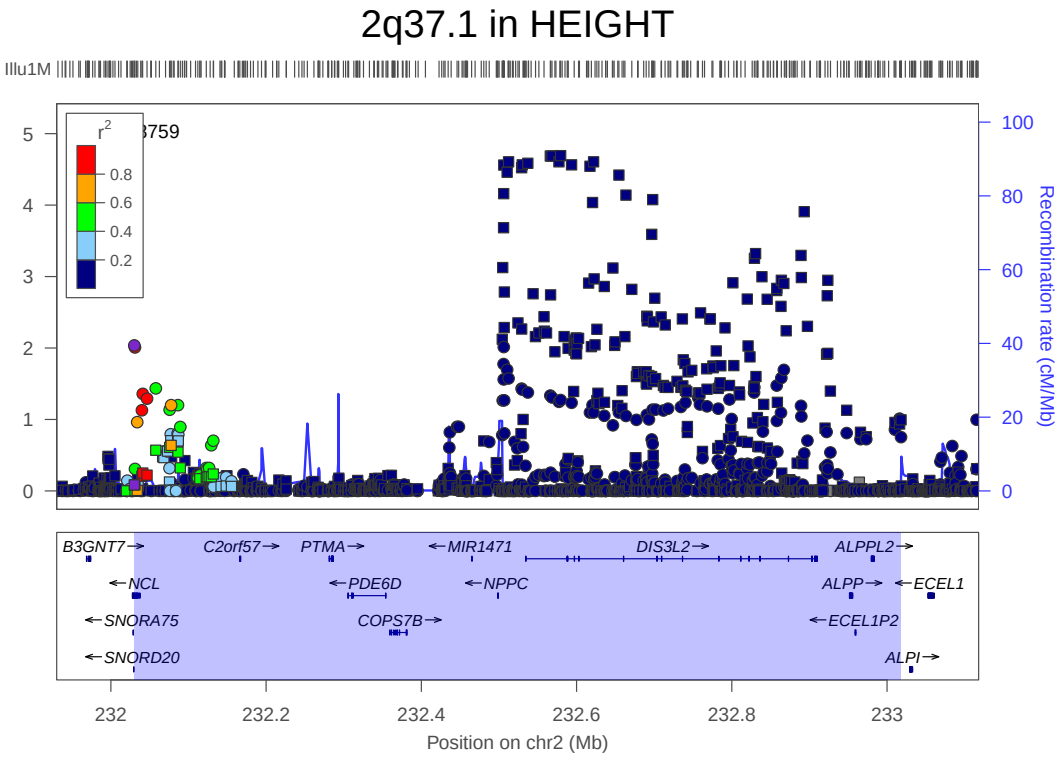
Figure 2.50: Discovery stage association at *13q31.3*.



2q37.1 The *2q37.1* locus is represented by lead marker rs7598759, with association signal for Height extending across $\approx 2kb$ of chromosome 988, ranging

from $232030kb - 233018kb$ (see Figure 2.51). Seventeen genes (*ALPP*, *ALPPL2*, *LINC00471*, *C2orf57*, *COPS7B*, *DIS3L2*, *ECEL1P2*, *MIR1244-1*, *MIR1244-2*, *MIR1244-3*, *MIR1471*, *NCL*, *NMUR1*, *NPPC*, *PDE6D*, *PTMA*, and *SNORD82*) overlap this signal region, and two markers within the region have been previously reported to be associated with body height (rs749052: $\approx 475kb$ & $\approx 0.68cM$ from lead marker with $r^2 = 0.01$ & $D' = 0.41$)[70, 238]. Our lead marker is intronic within nucleolin (*NCL*) which is a eukaryotic nucleolar phosphoprotein involved in synthesis and maturation of ribosomes[123].

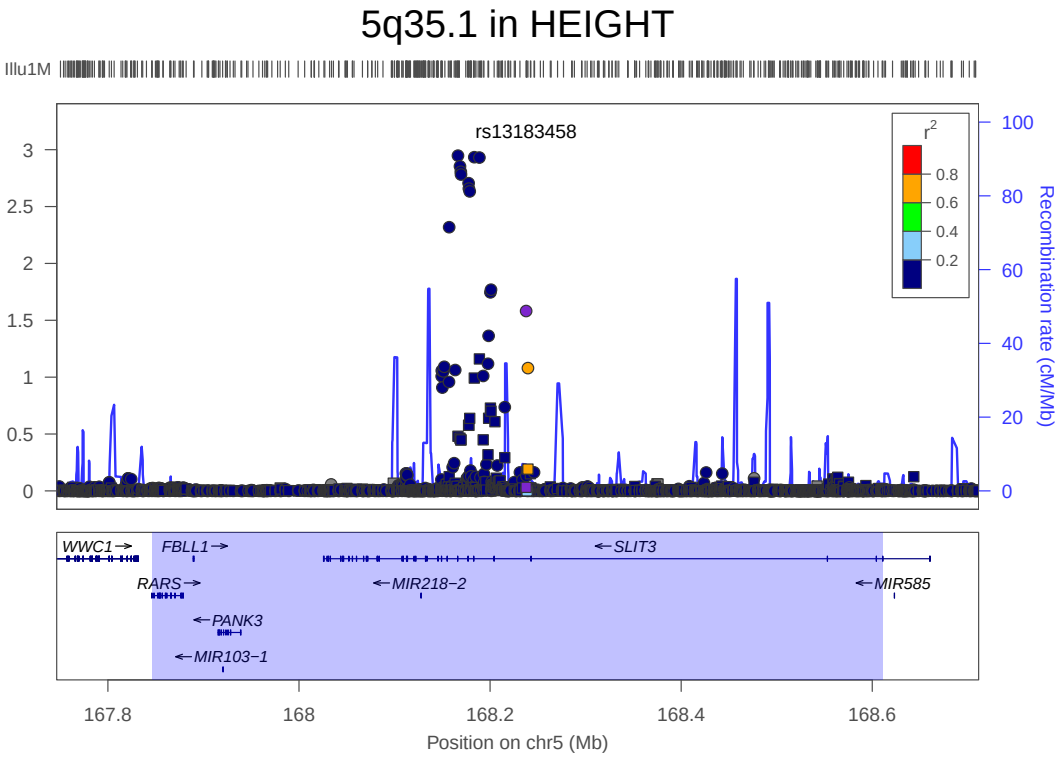
Figure 2.51: Discovery stage association at *2q37.1*.



5q35.1 The *5q35.1* locus is represented by lead marker rs13183458, with association signal for Height extending across $\approx 5kb$ of chromosome 765, ranging from $167846kb - 168611kb$ (see Figure 2.52). Seven genes (*FBLL1*, *MIR103A1*, *MIR103B1*, *MIR218-2*, *PANK3*, *RARS*, *SLIT3*) overlap this signal region, as does a SNP previously reported as associated with height (rs4282339: $\approx 49kb$ &

$\approx 0.19cM$ from lead marker with $r^2 = 0.03$ & $D' = 0.26$)[70]. The lead marker is an intronic variant within the slit homolog 3 (Drosophila) (*SLIT3)* gene. Cloning and expressions of three mammalian homologues of Drosophila slit suggest possible roles for Slit in the formation and maintenance of the nervous system[318]. SLT3 protein may act as a molecular guidance cue in cellular migration, and its function may be mediated by interaction with roundabout homolog receptors. *SLIT3* was also suggested to be associated with schizophrenia[319].

Figure 2.52: Discovery stage association at *5q35.1.*

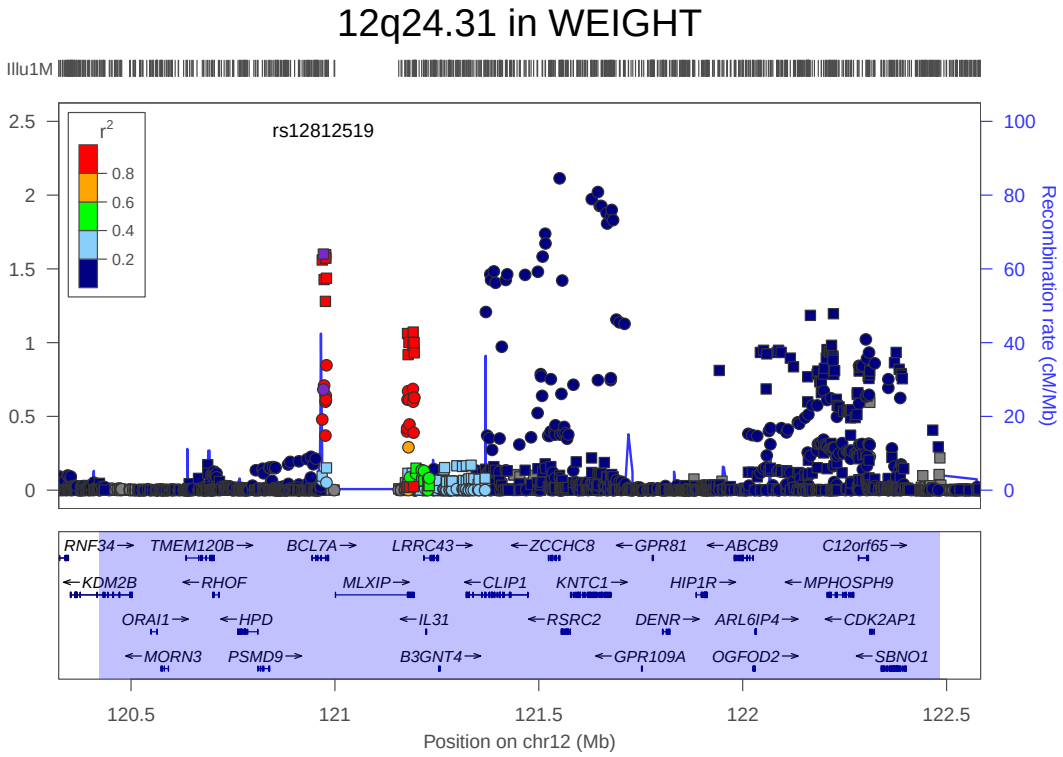


2.5.2.4 Weight loci

12q24.31 The *12q24.31* locus is represented by lead marker rs12812519, with association signal for Weight extending across $\approx 12kb$ of chromosome 2061, ranging from $120422kb - 122484kb$ (see Figure 2.53). Thirty-eight genes (*ABCB9*, *ARL6IP4*, *B3GNT4*, *BCL7A*, *C12orf65*, *CCDC62*, *CDK2AP1*, *CLIP1*, *DENR*, *DIA-*

BLO, *HCAR1*, *HCAR2*, *HCAR3*, *HIP1R*, *HPD*, *IL31*, *KDM2B*, *KNTC1*, *LRRC43*, *MIR4304*, *MLXIP*, *MORN3*, *MPHOSPH9*, *OGFOD2*, *ORAI1*, *PITPNM2*, *PSMD9*, *RHOF*, *RILPL2*, *RSRC2*, *SBNO1*, *SETD1B*, *SETD8*, *TMEM120B*, *VPS33A*, *VPS37B*, *WDR66*, *ZCCHC8*) overlap this signal region, as do SNPs associated with platelet count (rs7961894: $\approx 122kb$ & $\approx 0.21cM$ from lead marker with $r^2 = 0.02$ & $D' = 0.34$)[320], Parkinson Disease (rs12817488: $\approx 891kb$ from lead marker)[321], and body height (rs11830103: $\approx 1418kb$ from lead marker)[70]. Our lead marker is located within B-cell CLL/lymphoma 7A (*BCL7A)* gene, which is directly involved, along with Myc and IgH, in a three-way gene translocation in a Burkitt lymphoma cell line[123]. As a result of the gene translocation, the N-terminal region of the gene product is disrupted, which is thought to be related to the pathogenesis of a subset of high-grade B cell non-Hodgkin lymphoma[322]. The N-terminal segment involved in the translocation includes the region that shares a strong sequence similarity with those of B-cell CLL/lymphoma 7B (*BCL7B)* and B-cell CLL/lymphoma 7C (*BCL7C).* Two transcript variants encoding different isoforms have been found for this gene.

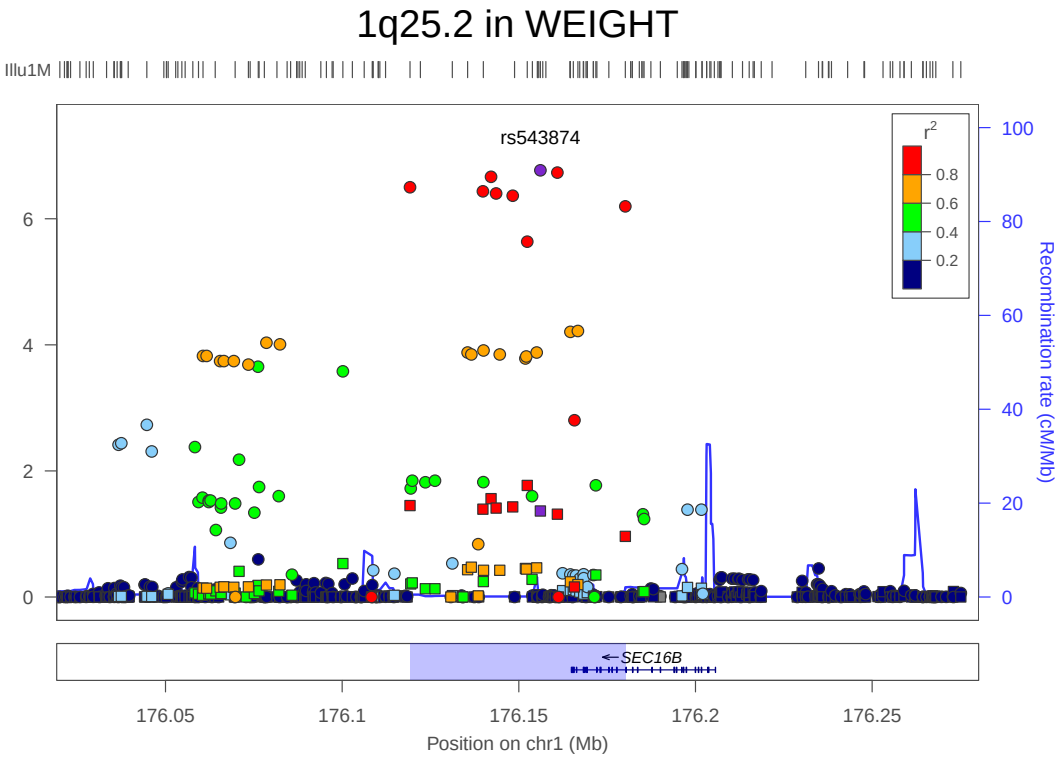
Figure 2.53: Discovery stage association at *12q24.31.*



1q25.2 The *1q25.2* locus is represented by lead marker rs543874, with association signal for Weight extending across $\approx 1kb$ of chromosome 61, ranging from $176119kb - 176180kb$ (see Figure 2.54). One gene (*SEC16B*) overlaps this signal region. The lead marker is in a conserved transcription factor binding site (TFBS) and has been previously reported as an association with BMI[55]. The signal region also contains two other SNPs with previously reported trait associations for age at menarche (rs633715: $\approx 37kb$ & $\approx 0.009cM$ from lead marker with $r^2 = 0.917$ & $D' = 0.957$), and for weight and BMI (rs10913469: $\approx 24kb$ & $\approx 0.019cM$ from lead marker with $r^2 = 0.959$ & $D' = 1.0$)[149]. Our lead marker is located $\approx 8kb$ downstream of SEC16 homolog B (S. cerevisiae) (*SEC16B)*, a mammalian homolog of S. cerevisiae Sec16 that is required for organization of transitional ER sites and protein export[323, 324]. This gene may be involved in the regulation of the effect of Ca2+ on liver cell functions and may play a pivotal

role in maintaining liver cell homeostasis and function. The associated region is also located $\approx 149-210kb$ upstream of RAS protein activator like 2 (*RASAL2)*, a Ras activating protein that regulates Ras activity, which is involved in cell growth and differentiation[325].

Figure 2.54: Discovery stage association at *1q25.2*.



2.5.3 Sex-specific analysis methods

2.5.3.1 Anthropometric phenotypes

The anthropometric traits we examined are height (cm), weight (kg), BMI ($\frac{kg}{m^2}$, computed as weight divided by height squared), WC (cm), HC (cm), and WHR ($unitless$). The latter three were analyzed both with and without adjustment for BMI, yielding nine phenotypes in total (height, weight, BMI, WC, HC, WHR, WCadjBMI, HCadjBMI, and WHRadjBMI). Phenotypes were normalized by individual studies but using a consistent transformations across studies. All traits except for height used an

inverse-normal transform (see Section 1.2.2.2) to ensure normality and consistency across studies and between men and women. All studies also adjusted for age by including *age* and age^2 as covariates in the regression model, along with BMI for the three BMI-adjusted phenotypes. Some studies also included additional covariates such as principal components (see Section 1.2.5.3) or subpopulation identifiers to correct for population stratification within the study.

2.5.3.2 Discovery GWAS

We included 46 discovery studies (60,586 men, 73,137 women) on height, weight and BMI, 34 studies (36231 men, 45192 women) on WC, 33 studies (34,942 men, 43,316 women) on HC and 32 studies (34,629 men, 42,969 women) on WHR. Each of the discovery studies was genotyped using Affymetrix or Illumina genotyping arrays. To enable meta-analyses across different SNP panels, each group performed genotype imputation using HapMap phase 2[20] HapMap population comprising samples from the CEPH collection (Utah residents with ancestry from northern and western Europe) (CEU) using MACH[64], IMPUTE[27] or BimBam[326].

2.5.3.3 Follow-up studies

For the follow-up stage, we included data from *in silico* GWAS as well as studies performing genotyping using the custom MetaboChip array.

in silico GWAS studies: We included 18 studies (20,340 men, 41,872 women) for height, weight, and BMI and 14 studies (11,225 men, 32,610 women) for WC, WCadjBMI, HC, HCadjBMI, WHR and WHRadjBMI genotyped via genome-wide SNP chips with subsequent imputation much like the Discovery GWAS (see Section 2.5.3.2).

MetaboChip studies: We also included 28 studies (42,055 men, 32,785 women) for height, weight and BMI and 26 studies (36,671 men, 28,326 women) for WC,

WCadjBMI, HC, HCadjBMI, WHR and WHRadjBMI genotyped via a custom iSELECT array known as "MetaboChip" which contains $\approx 195K$ SNPs submitted by several consortia performing GWAMA on traits such as T2D, lipid & glycaemic traits, and anthropometric traits. Some SNPs from a preliminary version of our sex-specific discovery stage analysis were submitted to the MetaboChip effort and were included in the design.

2.5.3.4 Analyses and quality control on the study level

Each study conducted sex-stratified association testing between the genotyped or imputed SNPs and each of the nine anthropometric phenotypes under an additive genetic model. Case/control studies additionally stratified by case/control status. In each study, the additive genetic effect for each SNP was estimated using a linear regression model on the transformed phenotypes using software packages such as MACH2QTL, SNPTEST[27], ProbABEL[327], GenABEL[328], Merlin[329] or PLINK[37]. Details on the study-specific analysis software were given previously (see Sections 2.1, 2.2, and 2.3 as well as Speliotes et al. [55][55],Heid et al. [54][54], andLango Allen et al. [70][70].

All study-specific files were processed by standardized QC routines that included checks of allele frequencies, compliance with Hapmap alleles, file completeness, number of markers, and ranges of test-statistics. We excluded monomorphic SNPs, SNPs with minor allele count (MAC) less than or equal to 3 (MAC = MAF $*$ N; MAF multiplied by sample size), and SNPs with low imputation accuracy as represented by information content measures ($\hat{r^2} < 0.3$ in MACH, or BIMBAM; $proper_info < 0.4$ in IMPUTE (see Section 1.2.3.1)[27, 58, 64]).

2.5.3.5 Genome-wide sex-difference analysis

In the genome-wide sex-difference analysis, we selected SNPs according to the p-value for sex-difference using an adjusted version of Welch's T-test[97] (see Sec-

tion 1.3.1.4). In the adjusted version, the SE of the difference was adjusted using the Spearman rank correlation coefficient calculated across all SNPs genome-wide (see Section 1.3.1.4 and (1.21)). For each SNP, we used (1.19) and set β_1 to the effect estimate in men and β_2 to the effect estimate in women that resulted from the meta-analysis of the follow-up studies. For SE_{diff}, we used (1.21), setting SE_1 to the standard error in men and SE_2 to the standard error in women, and using the Spearman rank correlation coeffecient calculated between β_1 and β_2 across all SNPs genome-wide as $r_{1,2}$. We then used a normal approximation to calculate the sex-difference p-value from the resulting T-statistic[97].

The correlation coeffecient $r_{1,2}$ ranged from 0.04 to 0.18 across phenotypes, which likely reflects some relatedness between men and women within the studies, although it could also represent a high number of associated SNPs (perhaps of very small effect). To correct for multiple testing, the sex-difference p-values for each of the $\approx 2.8M$ SNPs and each of the nine traits were concatenated, totaling $\approx 25.2M$ sex-difference p-values. Benjamini-Hochberg FDR[103] (see Section 1.3.3.3) was applied to the full set of $\approx 25.2M$ p-values to control FDR across all SNPs and phenotypes. This selection procedure did not yield any significant associations at either a 5% FDR or a 20% FDR.

2.5.3.6 Sex-specific multi-stage analysis

Sex-specific standard errors and p-values from each participating study were GC corrected[76] using the λ_{GC} values calculated for the genome-wide results for each phenotype. Following GC correction, β estimates were meta-analyzed using the inverse-variance fixed effect method as implemented in METAL[99] (see Section 1.3.1.3).

We also performed sample-size weighted Z-score meta-analysis (see Section 1.3.1.2) as a QC measure, but it yielded similar results and thus only the results of the inverse-variance are reported.In the discovery stage, a total of 2,971,914 SNPs were analysed in each of the 18 meta-analyses (9 phenotypes , each in both men and women).

For each SNP, we annotated the genetic position in cM by using the genetic map data from HapMap release 22[24]. For SNPs without available genetic position data in HapMap release 22, we approximated the genetic position by calculating the weighted average of the genetic positions of the two nearest HapMap SNPs, one from the 3' and one from the 5' direction, with the weights chosen to be proportional to the inverse of the physical distance (in bp) from the SNP to each of the two HapMap SNPs.

To correct for multiple testing, we concatenated the p-values resulting from all 18 meta-analyses together in yield a list of 50,586,560 p-values (i.e. $2sexes \times 9phenotypes \times \approx 2.8M$SNPs). We then controlled for FDR at a 5% level, simultaneously correcting for the multiple comparisons made across the SNPs, strata, and phenotypes[103]. This procedure identified 20,215 SNPs with $q - value < 0.05$. This list of SNPs was then pruned to obtain a list of independent SNPs for follow-up using genetic distance. In this procedure, we first took the SNP with the lowest q-valueand defined it as the lead SNP for a locus, then proceeded through the remaining SNPs sorted in ascending q-valueorder, associating any SNP within $0.2cM$ of an existing locus as a member of that locus and identifying any SNP that does not belong to an existing locus as a lead SNP for a new locus. This procedure yielded 619 independent loci, the lead SNPs of which were all carried forward to follow-up.

2.5.3.7 Follow-up and joint meta-analyses

For each SNP selected for follow-up and specifically for the phenotype that the SNP was selected for, we conducted sex-specific follow-up meta-analyses using the same statistical models as the discovery stage (additive genetic effect, same phenotype transformations and covariates, fixed effect model). We conducted the follow-up meta-analyses in two stages. In the first stage, we performed two meta-analyses: one on the set of *in silico* studies with genome-wide data and another on the set of MetaboChip studies. We then performed a 2-way meta-analysis of the results of those two analyses.

Sex-specific standard errors and p-values were also GC corrected in the same manner as in the discovery stage for follow-up studies with genome-wide data. For the MetaboChip studies, in which genome-wide data was not available, a subset of 4,427 SNPs present on the MetaboChip due to selection for QT-interval associations (QT-SNPs) were used to calculate λ_{GC} in order to avoid overestimating λ_{GC} due to previously known associations with anthropometric traits. QT-SNPs were used for λ_{GC} calculation since they were a priori the least likely to be associated with anthropometric traits.

Next, a joint meta-analysis combining the results of discovery and follow-up stage was conducted, again using the inverse-variance fixed effects method. SNP associations with a joint $p(discoveryandfollow-upcombined) < 5\times10^{-8}$ were considered genome-wide significant. Among these, loci close to and correlated with (within 1Mb and $r^2 \geq 0.2$) previously published anthropometric trait associated SNPs[3, 54, 55, 70] were identified as "near" previously published loci, and otherwise were identified as novel associations.

We then tested for sex-difference using only the follow-up data and in a similar manner to the discovery stage, except without using the adjusted $SE_{difference}$, since that could be biased by the small number of SNPs in the metabochip. This resulted in a p-value for sex-difference ($p_{sexdiff}$) for each of the associated SNPs.

Among the genome-wide significant SNP, we characterized the strength of evidence for sex-difference by using a Bonferroni correction (see Section 1.3.3.1 to obtain a threshold for significant $p_{sexdiff}$ and a nominal threshold of 0.05 for suggestive sex-difference. Using this procedure, SNPs with $p_{sexdiff} < 0.05/205$ were identified as having significant sex-difference and those with $p_{sexdiff} \leq 0.05$ were identified as having suggestive sex-difference, while the remaining SNPs with $p_{sexdiff} > 0.05$ were considered as having no evidence of sex-difference.

2.5.3.8 Age-stratified sex-specific meta-analysis

For the genome-wide significant signals that showed significant or suggestive sex difference in the SNP effect, each study partner of the discovery stage conducted another association analysis stratified by both sex and age group ($\geq 50 years$, $< 50 years$) but otherwise using the same models as described above.

2.5.3.9 Association with other phenotypes

The lead marker of the *2q24.3* locus, rs6717858, is associated with WHRadjBMI in women has also showed suggestive associations in the Global Lipids Genetic Consortium (GLGC) for HDL-C and TG[330] and in the Meta-Analyses of Glucose and Insulin-related traits Consortium (MAGIC) for FI[69]. The T allele of rs6717858 was associated with decreased levels of HDL-C ($p = 2.3 \times 10^{-6}$) and an increase in TG ($p = 1.1 \times 10^{-5}$) and FI ($p = 1.1 \times 10^{-5}$) in women, but not in men ($p = 0.88$, $p = 0.08$, and $p = 0.16$, respectively).

The women-specific association of rs11743303 with WHRadjBMI at the *5q11.2* locus also had suggestive sex-specific association with TG ($p = 1.1 \times 10^{-5}$ in women and $p = 0.08$ in men)[330].

While the association of rs1358980 at *6p21.1* had a women-specific effect on WHRadjBMI, the association of rs1358980 with HDL-C ($p = 1.5 \times 10^{-5}$) and TG ($p = 8.8 \times 10^{-7}$) was specific to men (p-values in women were $p = 0.02$ and $p = 0.02$, respectively).

At *3p25.1*, rs4684854 was associated with WHRadjBMI in women and also showed suggestive association in GLGC with total cholesterol in women ($p = 1.7 \times 10^{-5}$) but also had similar association in men ($p = 4.4 \times 10^{-4}$).

2.5.4 Discussion

2.5.4.1 Summary of findings

To interrogate the sex specific genetic architecture of anthropometric traits in autosomes, we conducted sex-specific meta-analysis of genome-wide association results and follow-up partly utilizing the Metabochip including over 270,000 individuals from 92 Studies from the GIANT consortium. We found six novel associations for height, WHRadjBMI, and WCadjBMI which also displayed significant or suggestive evidence for sex difference.

Few GWAS efforts have interrogated possible genotype-by-sex interaction. Those studies that have typically interrogate the top genome-wide significant findings only. We were interested to determine whether such an approach misses true loci with sex specific effects. While there is power in sex-combined scans to detect signals that are apparent in one sex but less pronounced in the other (CED), sex-stratified analyses would have better power to identify signals that completely disappear in one sex or signals that show opposite effect direction for men and women (OED)

Thus, the extent to which we may have previously missed sexually dimorphic signals would largely depend on whether signals with opposite effect direction (OED) exist. In our analyses, we did not detect any OED signals across nine anthropometric traits in our investigation controlling the false discovery rate of 5%.

2.5.5 Conclusion

Our investigation underscores the importance of considering sex-differences in an interrogation of the genetic architecture of anthropometric traits and has employed sex-specific analyses to identify several novel associations with human anthropometric traits. The numerous sex-specific associations found for theWCadjBMI, WHRadjBMI, and height phenotypes suggest a strong influence of sex on the genetics of those traits,

while our failure to detect sex-specific associations in BMI suggest the possibility of a less sexually dimorphic pattern in the genetics of overall obesity.

Chapter 3

Evaluation of summary-statistics meta-analyses compared to individual-level analyses

Combining individual level data into a single analysis (mega-analysis) has been undertaken for some GWAS meta-analyses[331–333], although many more large-scale analyses have performed meta-analyses of summary statistics[3–5, 54, 55, 57, 65, 68–71, 149, 208, 236–238, 258, 259, 331, 334–336] using one of the methods described in Section 1.3.1. This may be owing more to political and privacy concerns inherent in sharing such low-level data rather than to a technical advantage of one method over the other, but to my knowledge no comparison between the various methods has been made, so it may be difficult for study designers to evaluate which method would be preferred for future work. In this chapter I aim to compare the performance of various methods of meta-analysis and mega-analysis is order to characterise the relative merits of meta-analysis and mega-analysis in the context of large-scale GWAS under various conditions.

3.1 Meta/mega-analysis in homogeneous populations

In a realistic analysis of any kind, heterogeneity is likely to be present due to at least some degree of population stratification[331]. However, it may be informative to first

examine meta- and mega-analysis methods under a condition with no population-based stratification present in order to determine the baseline power of the methods themselves without the addition of corrections for stratification which complicate the analyses, although they are typically applied in such meta-analyses[67].

In a homogeneous population in which no correction for population stratification is required, simulations of a single SNP can be used to evaluate various methods of meta- and mega-analysis since there is no need to model LD patterns or to use correction methods that are based on a large set of SNPs with no true trait association (see Section 1.2.5.1). For these simulations, the total number of samples was fixed at 2,500 while other parameters including the effect size of the SNP, the number of studies the samples were broken into for meta-analysis, and the degree of genotyping uncertainty was varied. For each iteration of the simulation, a single common variant was chosen to be the effect SNP from the set of 2,786 SNPs on chromosome 21 with average MAF between 40%–50% in the HapMap phase 3[46] CEU population. The genotypes for the effect SNP were then simulated for each sample by randomly assigning genotype based on the genotype frequencies from the HapMap phase 3[46] CEU population. A quantitative phenotype was also simulated conditional on the simulated effect SNP genotype: for each sample within a population, a random value was drawn from a normal distribution as the base phenotype value, and an additive genetic model was used to add the simulated effect size (effect per allele) to that base value for each copy of the effect allele present at the effect SNP to yield the final phenotype value for that simulated individual.

When varying the number of studies, the total sample of 2,500 samples was divided into $N_{studies}$ studies each comprising at least $N_{samples} = \lfloor \frac{2{,}500}{N_{studies}} \rfloor$ (the fraction of the total 2,500 sample size rounded down to the nearest integer). In order to bring the total sample size across all studies to exactly 2,500, one additional sample was added to k of the studies, where $k = 2,500 - N_{samples} \times N_{studies}$.

When varying the degree of genotype uncertainty, the genotypes were simulates as above, and then a procedure was used to simulate calling error. Briefly, the simulated true genotype was first mapped onto a cartesian plane with each point located a radius of 1 from the origin, with the two homozygous genotypes mapped to a points at $(0, 1)$ and $(1, 0)$ and the heterozygous genotype mapped to a point at $(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$. Next, error was introduced into the position on both the x and y axes by drawing two error values from a normal distribution with mean 0 and standard error equal to the degree of genotype uncertainty. The probability of belonging to each of the three clusters was then approximated using a simplistic genotype calling algorithm based on the relative distance to each of the three true genotype cluster centres, with the highest probability assigned to the closest genotype cluster and the lowest probability to the farthest genotype cluster. The formula also had the property that if the point was at the midpoint between two clusters, the probability of belonging to those two genotypes would be equal in the resulting genotype uncertainty data.

Meta-analysis: Fisher's method In the meta-analysis using Fisher's method[91], the genotype and phenotype data was tested for association within each simulated study using the missing data likelihood score test method implemented in SNPTEST[27]. The β and SE from each study were then used to calculate a p-value for each study, and that p-value was then pooled using Fisher's combined probability test[90] (described in Section 1.3.1.1), using a custom GNU R[41] script to generate a single p-value according to Equation 1.11, which was then recorded as the output p-value from this method.

Meta-analysis: WZ In the meta-analysis using the WZ method[93], the genotype and phenotype data was tested for association within each simulated study using the missing data likelihood score test method implemented in SNPTEST[27]. The β and SE from each study were then used to calculate a p-value and that p-value along with the sample size was used in a WZ meta-analysis[93] (see

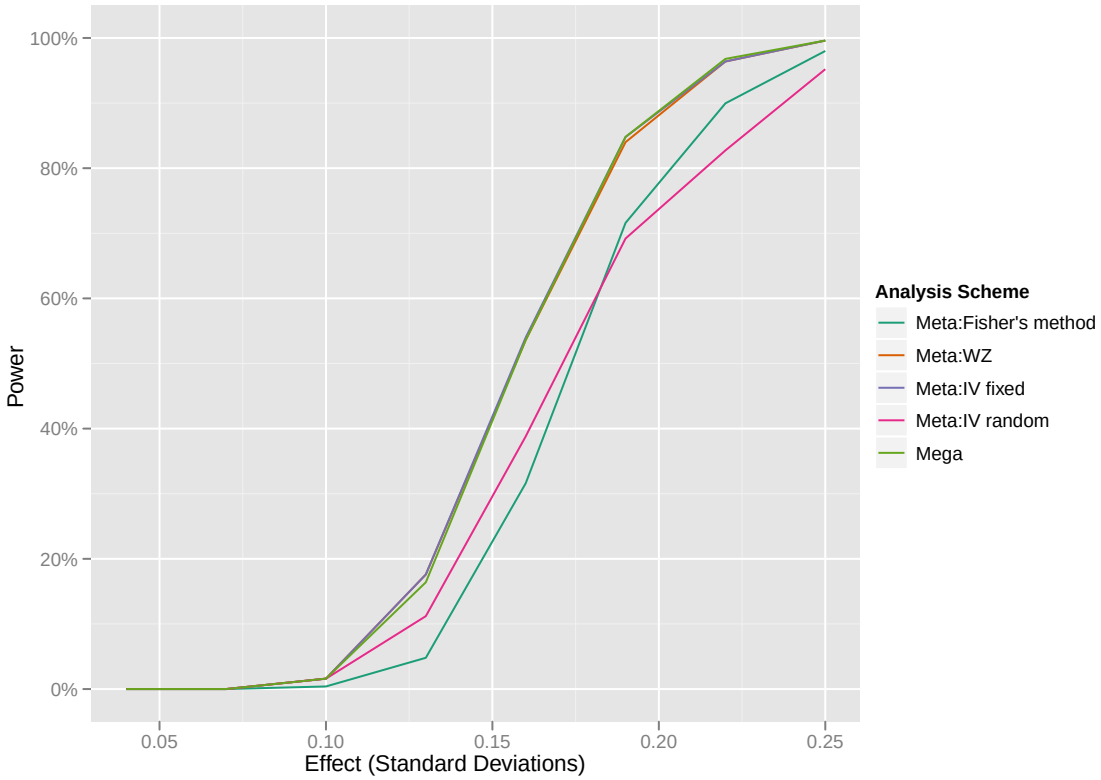
Section 1.3.1.2) implemented in METAL[99]. The resulting p-value was recorded as the output p-value from this method.

Meta-analysis: IV fixed In the meta-analysis using the IV fixed-effects method[94], the genotype and phenotype data was tested for association within each simulated study using the missing data likelihood score test method implemented in SNPTEST[27]. The β and SE from each study were then used in a IV fixed-effects meta-analysis[94] (see Section 1.3.1.3) implemented in GWAMA[337]. The resulting p-value was recorded as the output p-value from this method.

Meta-analysis: IV random In the meta-analysis using the IV random-effects method[96], the genotype and phenotype data was tested for association within each simulated study using the missing data likelihood score test method implemented in SNPTEST[27]. The β and SE from each study were then used in a IV random-effects meta-analysis[96] (see Section 1.3.1.3) implemented in GWAMA[337]. The resulting p-value was recorded as the output p-value from this method.

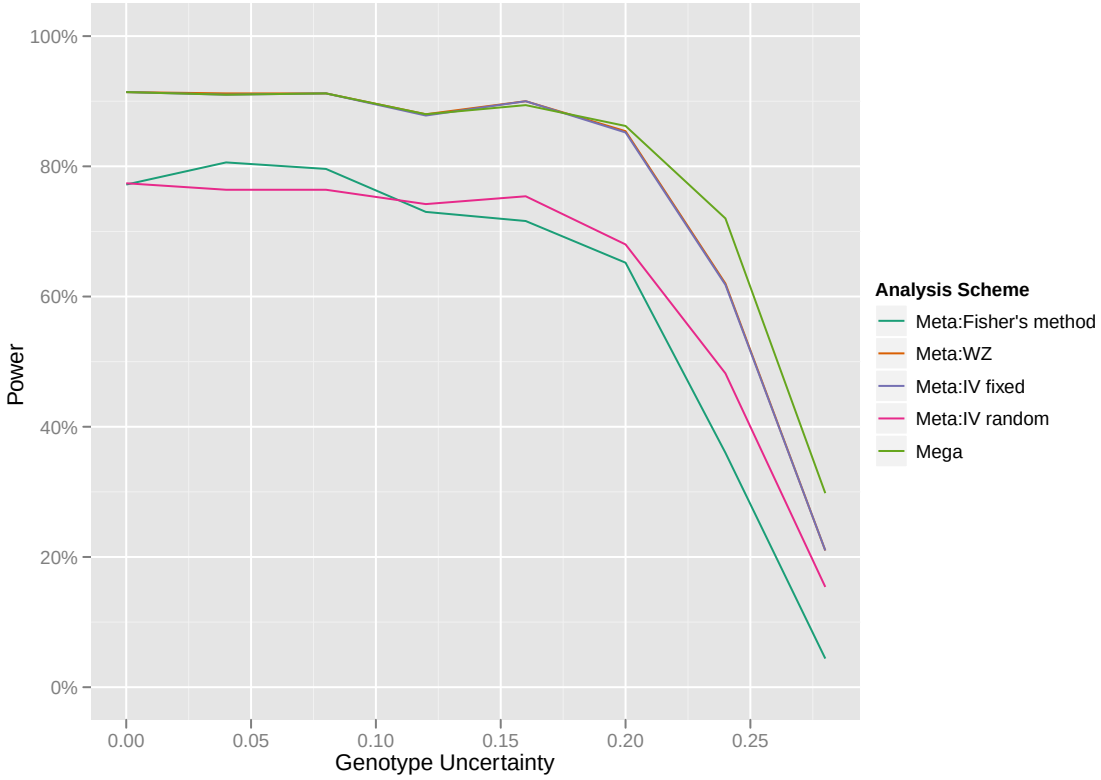
Mega-analysis In the mega-analysis, the genotype and phenotype data from each study were first merged into a single data set. The combined data were then tested for association using the missing likelihood score test method implemented in SNPTEST[27]. The resulting p-value was recorded as the output p-value from this method. No corrections were performed based on study or population identity since the simulated populations across all studies were all drawn from the same population.

Figure 3.1: Power of five different analysis methods based on simulations of five homogeneous populations of 500 samples each across different effect sizes.



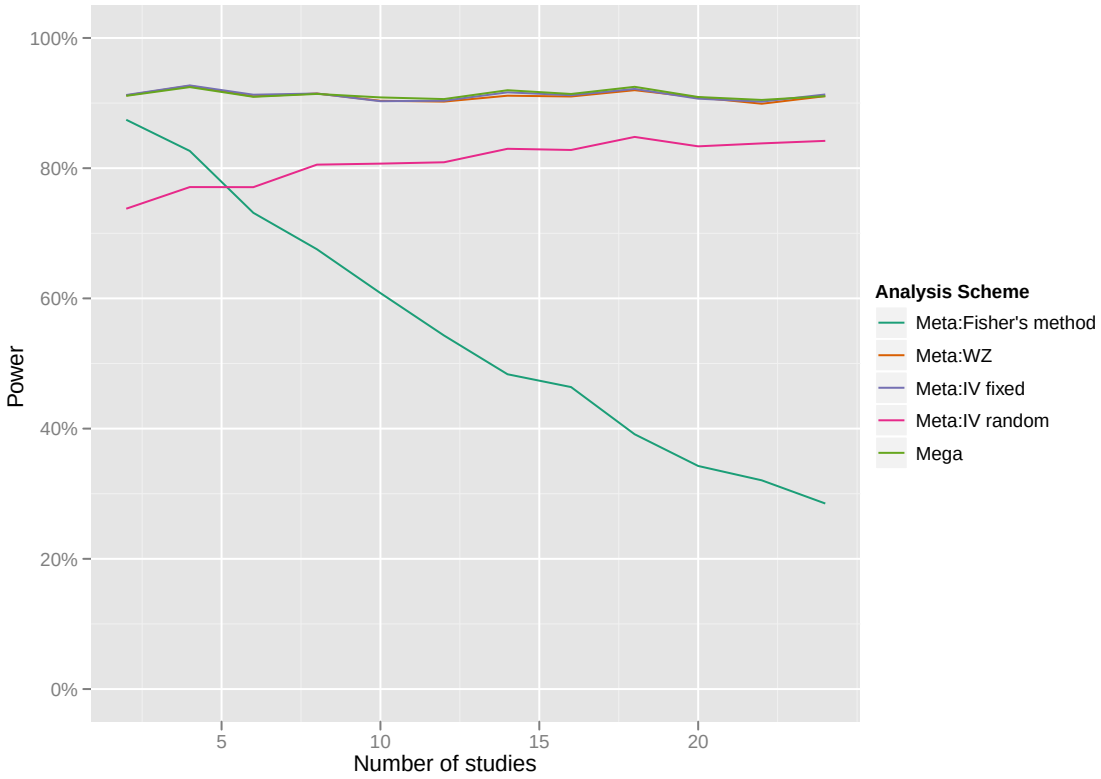
The power analysis across effect sizes shows (Figure 3.1) that the mega-analysis has approximately the same power as the meta-analysis methods based on WZ[93] and IV fixed-effects methods[94], while the meta-analysis based on Fisher's method[91] and the IV random-effects method[96] had somewhat less power. The three top methods reached 80% power at an effect size of ≈ 0.185 standard deviations, while Fisher's method[91] did not reach 80% power until an effect size of ≈ 0.203 and the IV random-effects method did not reach 80% power until the effect size reached ≈ 0.214.

Figure 3.2: Power of five different analysis methods based on simulations of five homogeneous populations of 500 samples each across varying degrees of genotype uncertainty.



The power analysis across varying degrees of genotype uncertainty shows (Figure 3.2) that as the degree of genotype uncertainty increases and the power curve moves into the linear region, the mega-analysis has about 10% more power than any of the meta-analysis methods.

Figure 3.3: Power of five different analysis methods based on simulations of 2,500 samples split into varying numbers of studies for meta-analysis.



The power analysis across varying number of studies (Figure 3.3) shows clearly that the power of Fisher's method[91] of meta-analysis declines substantially as the number of studies increases. The power of the mega-analysis as well as both the WZ[93] and IV fixed effects meta-analyses[94] appear to remain approximately equal across the range of $N_{studies}$ simulated here. The limitation on $N_{studies}$ was SNPTEST[27], whose tests require large sample size and therefore does not run tests for $N_{samples} < 100$.

3.2 Meta/mega-analyses in the presence of within-study population stratification

When population stratification is present with a study, the power of the various meta-analysis methods could be different based on the method used to correct the

population stratification. Some of these methods, such as GC (Section 1.2.5.2) and PCA (Section 1.2.5.3), require a large set of markers with no true association with the phenotype. Therefore, in order to simulate realistic analyses in the presence of population stratification it is necessary to simulate, in addition to an effect SNP, a large number of SNPs without true trait association.

For these simulations, the number of simulated studies was fixed at five, each one comprising 500 samples for a total of 2,500 samples. In order to simulate realistic patterns of LD, reference haplotypes and fine-scale recombination maps[338] from HapMap phase 3[46] were used to simulate a realistic set of population control genotypes using HAPGEN2[89], being careful to ensure that each of the HAPGEN2 random seeds was unique by running no more than one instance per second and checking the random seed values for collisions after running all iterations. Simulation of a full set of genome-wide SNPs would be computationally intensive to run in numerous simulation iterations, so as a compromise a smaller set of markers consisting of all 19,306 markers with genotypes in HapMap phase 3[46] on chromosome 21 was used to represent a genome-wide set of SNPs. For each iteration of the simulation, a common variant was chosen to be the effect SNP from the set of 2,786 SNPs with average MAF between 40%–50% across the four HapMap phase 3[46] populations used in the simulation.

A quantitative phenotype was then simulated for each of several different effect sizes, dependent on genotype at the effect SNP. For each sample within a population, a random value was drawn from a normal distribution with mean (μ) and standard deviation (SD) given in Table 3.1 and an additive genetic model was used to add the effect size (effect per allele) to that value for each copy of the effect allele present at the effect SNP.

The genotype and phenotype data generated within each iteration and effect size were then tested for association using each of several methods:

Table 3.1: Population mean and standard deviation used in simulations.

HapMap phase 3[46] Population	Mean	Standard Deviation
GIH	160	11
MXL	165	12
CEU	170	10
CEU+TSI	175	9
TSI	180	8

Meta-analysis: Fisher's method & GCC In the meta-analysis using Fisher's method[91] & GCC, the genotype and phenotype data was tested for association within each simulated study using the missing data likelihood score test method implemented in SNPTEST[27]. The β and SE from each study were then used to calculate a p-value for each SNP and λ_{GC} was calculated based on the p-values of all SNPs according to the method described in Section 1.2.5.2. That GC parameter was then used to correct the p-value for each SNP according to Equation 1.7, and the p-values from each study were then pooled using Fisher's combined probability test[90] (described in Section 1.3.1.1), using a custom GNU R[41] script to generate a single p-value for each SNP according to Equation 1.11, which were then recorded as the output p-values from this method.

Meta-analysis: WZ & GCC In the meta-analysis using the WZ method[93] & GCC, the genotype and phenotype data was tested for association within each simulated study using the missing data likelihood score test method implemented in SNPTEST[27]. The β and SE from each study were then used to calculate a p-value for each SNP and λ_{GC} was calculated based on the p-values of all SNPs according to the method described in Section 1.2.5.2. That GC parameter was then used to correct the p-value for each SNP according to Equation 1.7, and the p-value and sample size were used in a WZ[93] meta-analysis (see Section 1.3.1.2) implemented in METAL[99]. Finally, the p-values resulting from the meta-analysis across all SNPs were used to calculate an overall λ_{GC}, which was then used to generate a corrected p-value for each SNP, again according to

Equation 1.7, which were recorded as the output p-values from this method.

Meta-analysis: IV fixed & GCC In the meta-analysis using the IV fixed-effects method[94] & GCC, the genotype and phenotype data was tested for association within each simulated study using the missing data likelihood score test method implemented in SNPTEST[27]. The β and SE from each study were then used to calculate a p-value for each SNP and λ_{GC} was calculated based on the p-values of all SNPs according to the method described in Section 1.2.5.2. That GC parameter was then used to correct the SE for each SNP according to Equation 1.8, and the β and SE were then used in a IV fixed-effects meta-analysis[94] (see Section 1.3.1.3) implemented in GWAMA[337]. Finally, the β and SE resulting from the meta-analysis across all SNPs were again used to calculate an overall λ_{GC} and that was used to generate a corrected p-value for each SNP, again according to Equation 1.8, and the resulting p-values were recorded as the output p-values from this method.

Meta-analysis: IV random & GCC In the meta-analysis using the IV random-effects method[96] & GCC, the genotype and phenotype data was tested for association within each simulated study using the missing data likelihood score test method implemented in SNPTEST[27]. The β and SE from each study were then used to calculate a p-value for each SNP and λ_{GC} was calculated based on the p-values of all SNPs according to the method described in Section 1.2.5.2. That GC parameter was then used to correct the SE for each SNP according to Equation 1.8, and the β and SE were then used in a IV random-effects meta-analysis[96] (see Section 1.3.1.3) implemented in GWAMA[337]. Finally, the β and SE resulting from the meta-analysis across all SNPs were again used to calculate an overall λ_{GC} and that was used to generate a corrected p-value for each SNP, again according to Equation 1.8, which were recorded as the output p-values from this method.

Mega-analysis: GCC In the mega-analysis using GCC, the genotype and phenotype data from each study were first merged into a single data set. The combined data were then tested for association using the missing likelihood score test method implemented in SNPTEST[27]. The β and SE output from SNPTEST for each SNP was then used to calculate a p-value for each SNP and λ_{GC} was calculated based on the p-values of all SNPs according to the method described in Section 1.2.5.2. That GC parameter was then used to correct the p-value for each SNP according to Equation 1.7, which were finally recorded as the output p-values from this method.

Mega-analysis: population covariates In the mega-analysis using population identifiers as covariates, the genotype and phenotype data from each study were first merged into a single data set. The combined data were then tested for association using the missing likelihood score test method implemented in SNPTEST[27] and including a population identifier as a covariate. The resulting p-values were recorded as the output p-values from this method.

Mega-analysis: population covariates & GCC In the mega-analysis using population identifiers as covariates with GCC, the genotype and phenotype data from each study were first merged into a single data set. The combined data were then tested for association using the missing likelihood score test method implemented in SNPTEST[27] using a population identifier as a covariate. The β and SE output from SNPTEST were then used to calculate a p-value for each SNP and λ_{GC} was calculated based on the p-values of all SNPs according to the method described in Section 1.2.5.2. That GC parameter was then used to correct the p-value for each SNP according to Equation 1.7, which were recorded as the output p-values from this method.

Mega-analysis: PCA with three PCs In the mega-analysis using PCA with three PCs, the genotype and phenotype data from each study were first merged into a single data set and converted to PED format for processing by Eigensoft[83],

which was then used to calculate the first three PCs based on the full data set. The combined data were then analysed by SNPTEST for genotype-phenotype association using the three PCs as covariates, using the missing likelihood score test method[27]. The resulting p-values were recorded as the output p-values from this method.

Mega-analysis: PCA with three PCs & GCC In the mega-analysis using PCA with three PCs with GCC, the genotype and phenotype data from each study were first merged into a single data set and converted to PED format for processing by Eigensoft[83], which was then used to calculate the first three PCs based on the full data set. The combined data were then analysed by SNPTEST for genotype-phenotype association using the three PCs as covariates, using the missing likelihood score test method[27]. The β and SE output from SNPTEST were then used to calculate a p-value for each SNP and λ_{GC} was calculated based on the p-values of all SNPs according to the method described in Section 1.2.5.2. That GC parameter was then used to correct the p-value for each SNP according to Equation 1.7, which were recorded as the output p-values from this method.

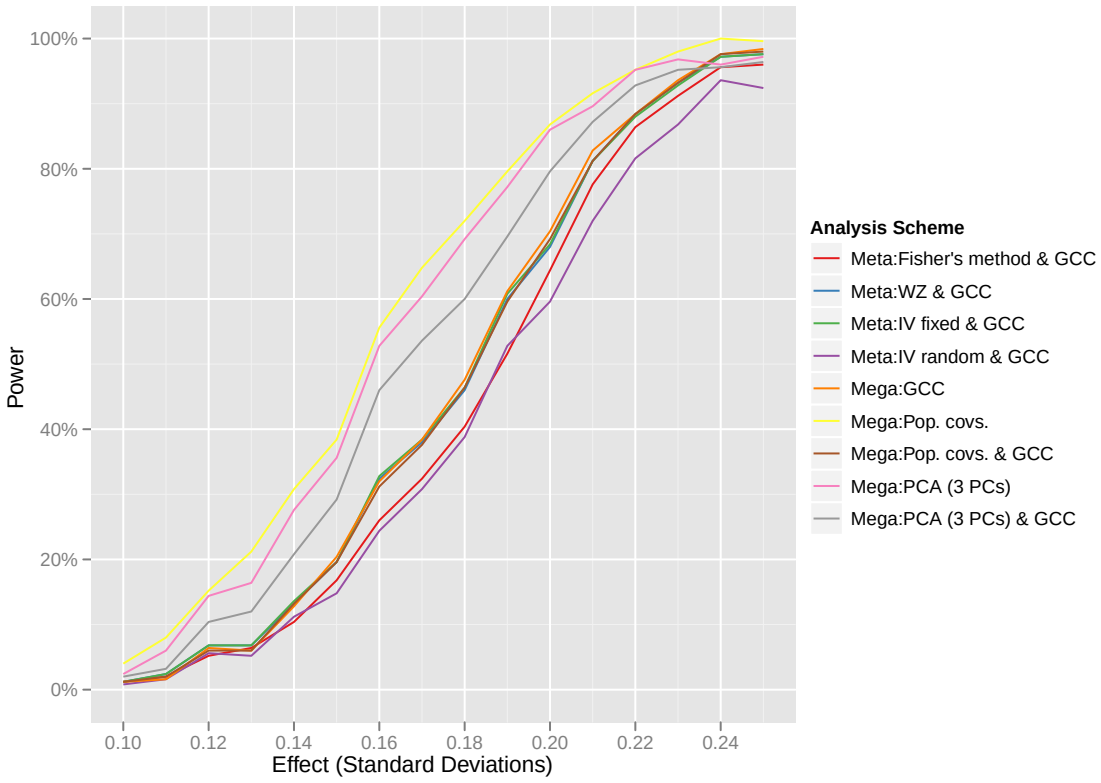
Each of the above analysis methods resulted in a list of SNPs with a p-value for each. These analysis results were post-processed by sorting the list of SNPs by p-value and then assigning each SNP into an independent locus (using the pruning procedure described in Section 1.3.2.1). A locus was defined using a threshold of 0.2cM based on the HapMap phase 2[20] genetic map[338]. The lead marker (representing the independent locus) of the effect SNP was then identified and the p-value of that locus checked to see if it exceeded the GWS threshold of $p < 5 \times 10^{-8}$. If it did, this was recorded as a true positive event, and if not it was recorded as a false negative event (type II error). Next, a list of independent loci excluding the effect SNP locus that exceeded the GWS threshold of $p < 5 \times 10^{-8}$ was generated and the count of those loci was recorded as the number of false positive events (type I errror). Finally,

the list of independent loci was counted and that count, minus the one true effect locus, was recorded as the number of true negative events.

For each analysis method and effect size combination, the power was determined by dividing the number of true positive events divided by the total number of true events (equal to the number of iterations since there was a single effect locus simulated for each). The specificity was also calculated as the number of true negative events divided by the sum of true negative events plus false positive events. Power and specificity were plotted against effect size for each analysis method (Figures 3.4 & 3.5).

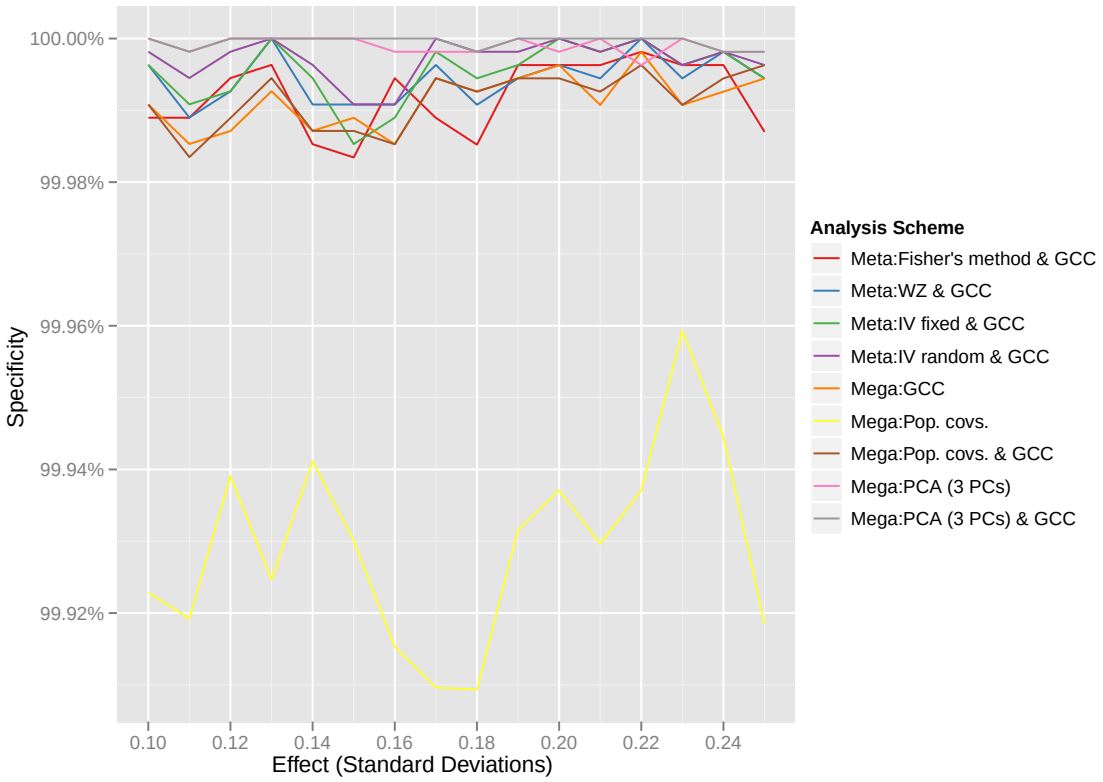
The results of the simulation show that across all effect sizes, the mega-analysis using population covariates appeared to be the most powerful (Figure 3.4). While this

Figure 3.4: Power of nine different analysis methods based on simulations of five populations of 500 samples with 10% within-study heterogeneity.



is true, it is important to note that this analysis does not include a correction for population stratification within studies, so the assumptions of that analysis are that there is no population stratification within studies. However, since all five constituent studies were simulated with substantial population stratification included each study, this analysis would be likely to have a high false-positive rate. Indeed, as can be seen in the plot of specificity versus effect size, this analysis has a substantially lower specificity (and therefore a much higher false-positive rate) than any of the other analysis methods (Figure 3.5) in this simulation.

Figure 3.5: Specificity of the nine different analysis methods based on simulations of five populations of 500 samples with 10% within-study heterogeneity.



After performing genomic control correction (GCC) on the mega-analysis using population covariates, the power dropped by $\approx 23\%$ (Table 3.2), which is consistent with a high degree of population stratification being found within the studies.

Excluding the mega-analyses using population covariates, the next highest power was obtained from the mega-analysis correcting for stratification using PCA. The PCA correction mega-analyses were performed both with and without an additional GCC, and the version without an additional correction had higher power than the one with an additional correction, although both had similarly good specificity (Figure 3.5). Performing the additional GC correction on the PCA corrected data resulted in a loss of power of $\approx 9.5\%$, while the specificity was not noticeably improved (Table 3.2).

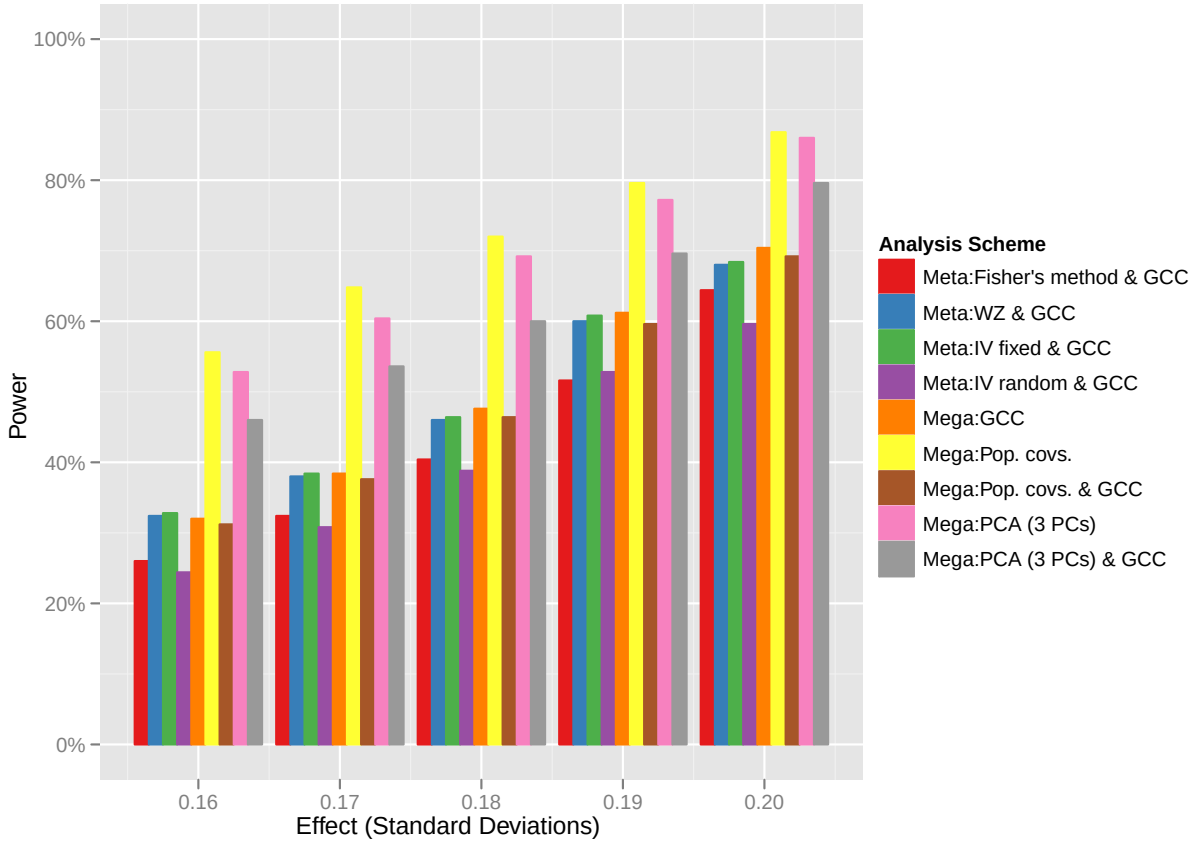
Comparing the mega-analysis with PCA correction to the best meta-analysis method, which was IV fixed effects with[94] GCC, we see that in the presence of within-study heterogeneity we have $\approx 20\%$ greater power using mega-analysis with PCA than meta-analysis with GCC when within the linear region of the power curve. See Figure 3.6 for more detail on this region of the power curve and Table 3.2 for the average power of each method within this region. Based on GWApower data[339], the approximate number of samples needed for the best meta-analysis method (IV fixed effects[94]) to make up for the loss in power over the best mega-analysis method (mega-analysis with PCA correction) under the conditions of unreported heterogeneity within studies, is ≈ 585—a meta-analysis with total sample size of $3,085$ would have approximately the same power as a mega-analysis with total sample size of $2,500$ (23.4% additional samples).

Within the linear region of the power curve a comparison between the meta-analysis methods showed that the IV fixed effects method[94] was consistently the most powerful, with the WZ method averaging only slightly less (0.48%) power and the IV random effects method[96] having a 8.08% reduction in power, which is as expected given that the simulated effects were actually fixed across studies. Fisher's method[91] had more variability (Figure 3.6), but averaged 6.4% less than the IV fixed effects method[94], making it slightly more powerful than the IV random effects method[96] on average, although it should be noted that having exactly equal sample sizes across the five studies as we do here may put Fisher's (unweighted) method at an advantage.

Table 3.2: Power of nine different analysis methods based on simulations of five populations of 500 samples with 10% within-study heterogeneity, averaged over effect sizes between 0.16–0.2 standard deviations per allele (within the linear range of the power curve).

Analysis Method	Average Power
Meta-analysis: Fisher's method & GCC	42.96%
Meta-analysis: WZ & GCC	48.88%
Meta-analysis: IV fixed & GCC	49.46%
Meta-analysis: IV random & GCC	41.28%
Mega-analysis: GCC	49.92%
Mega-analysis: population covariates	71.76%
Mega-analysis: population covariates & GCC	48.80%
Mega-analysis: PCA with 3 PCs	69.12%
Mega-analysis: PCA with 3 PCs & GCC	61.76%

Figure 3.6: Power within the linear region between 0.16–0.2 SD of nine different analysis methods based on simulations of five populations of 500 samples with 10% within-study heterogeneity.



3.3 Meta/mega-analyses in the presence of between-study population stratification

Another type of stratification in a meta/mega-analysis could occur when each study consists of an homogeneous population without any within-study heterogeneity, but with substantial heterogeneity between-studies. To compare meta- and mega-analysis methods under these conditions, an additional set of simulations were performed.

For these simulations, the same methods were used as described in Section 3.2, except in this case each of the five studies was based on a different population based on HapMap phase 3[46] haplotypes. For each of the five populations CEU, GIH, MXL, TSI, and CEU+TSI (where CEU+TSI include the haplotypes of both CEU and TSI and is therefore an average of the two), 500 samples were simulated using HAPGEN2[89], again using all 19,306 markers with genotypes in HapMap phase 3[46] on chromosome 21, and with an effect SNP randomly chosen from the set of 2,786 SNPs with average MAF between 40%–50% across the populations used in the simulation.

The results of these simulations (Figure 3.7) show that the most powerful method under these conditions are two of the weighted meta-analysis methods, with the IV fixed-effects method[94] having slightly more (1.3%) power than the WZ[93] method in the linear region of the power curve, with an average of 70.3% (Table 3.3 and Figure 3.8). Following these two methods is the mega-analysis method based on population covariates, which had 10% less power than the IV fixed-effects method[94] and the population covariates with additional GCC which had 0.7% less power than the method without additional GCC (in this case, the GCC should be unnecessary since the population covariates capture all of the heterogeneity due to population structure). The next most powerful method was the mega-analysis based on PCA, which had 14.3% less power than the IV fixed-effects meta-analysis[94] and 9.3% less power than the mega-analysis using population covariates. This intuitively makes sense because under these conditions, the population covariates (and meta-analysis

strata) perfectly capture the population stratification present in the data, while the PCA method is only able to correct based on an estimate of the differences between the populations. The PCA based mega-analysis methods also had the lowest specificity (Figure 3.9), again likely because the three PCs do not represent population differences as well as the population covariates or the stratified-by-population analyses that were effectively performed by the meta-analysis methods.

Finally, the meta-analysis using only GCC to control for population stratification was unable to detect the effect SNP in any of the simulation iterations, putting it at $\approx 0\%$ power. This appears to result from the GCC being based on extremely large values of λ_{GC}, which averaged around $\lambda_{GC} = 17.4$ (values of $\lambda_{GC} > 1$ indicate some inflation of the test statistic; see Section 1.2.5.2) because the widespread allele frequency differences between the five populations caused widespread spurious associations.

Figure 3.7: Power of nine different analysis methods based on simulations of five different populations of 500 samples each.

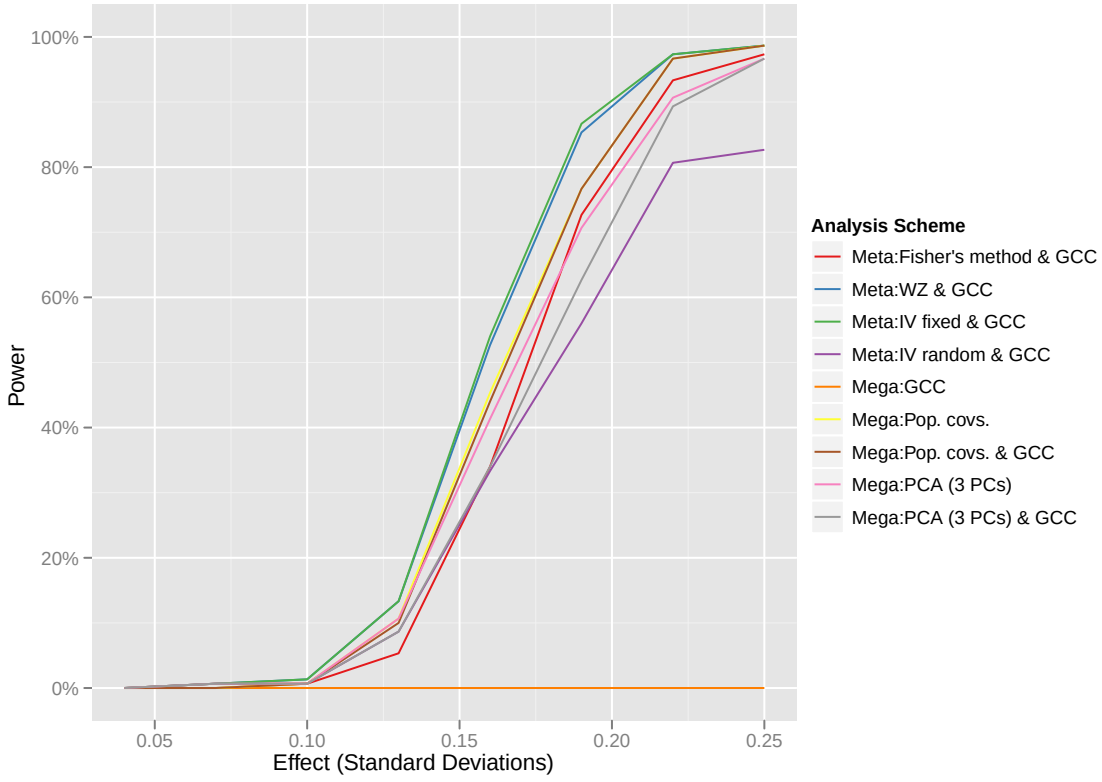


Table 3.3: Power of nine different analysis methods based on simulations of five different populations of 500 samples each, averaged over effect sizes between 0.16–0.2 standard deviations per allele (within the linear range of the power curve).

Analysis Method	Average Power
Meta-analysis: Fisher's method & GCC	53.3%
Meta-analysis: WZ & GCC	69.0%
Meta-analysis: IV fixed & GCC	70.3%
Meta-analysis: IV random & GCC	44.7%
Mega-analysis: GCC	0.0%
Mega-analysis: population covariates	61.0%
Mega-analysis: population covariates & GCC	60.3%
Mega-analysis: PCA with 3 PCs	56.0%
Mega-analysis: PCA with 3 PCs & GCC	48.3%

Figure 3.8: Power within the linear region between 0.16–0.2 SD of nine different analysis methods based on simulations of five different populations of 500 samples each.

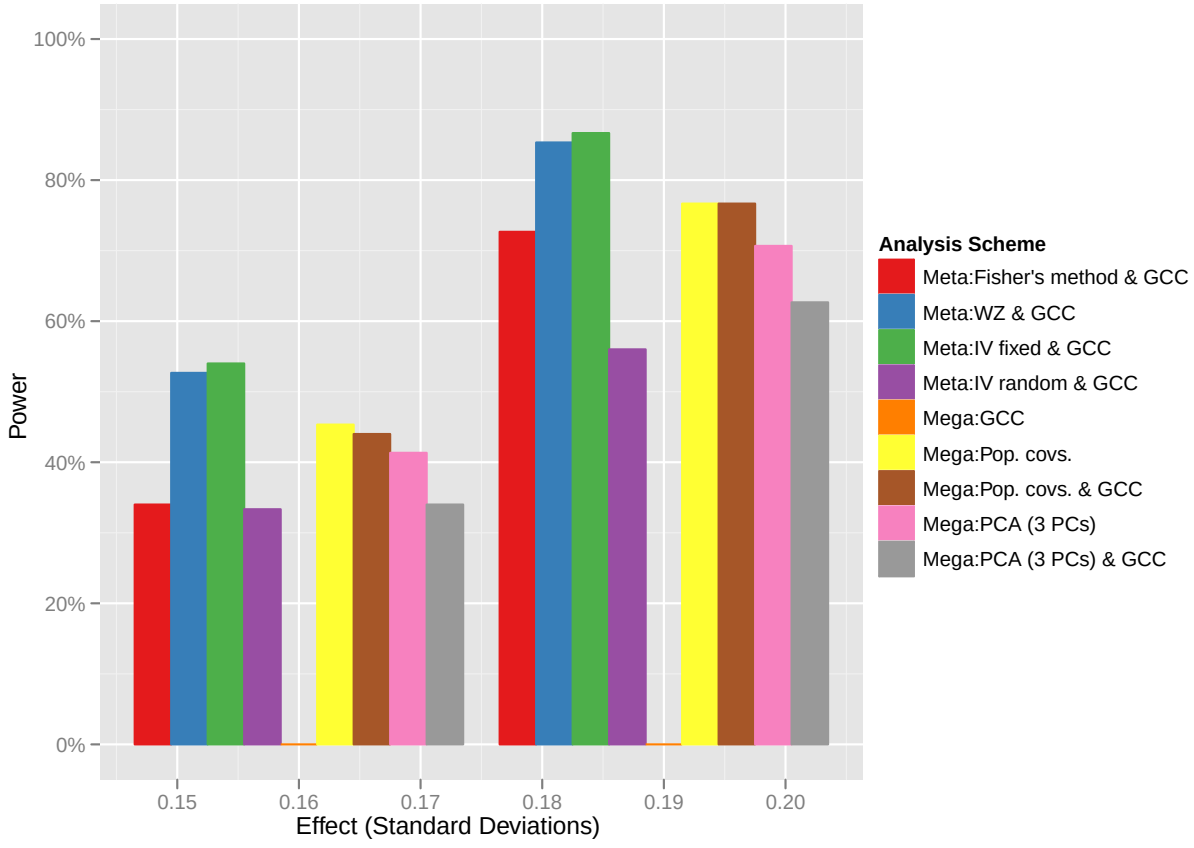
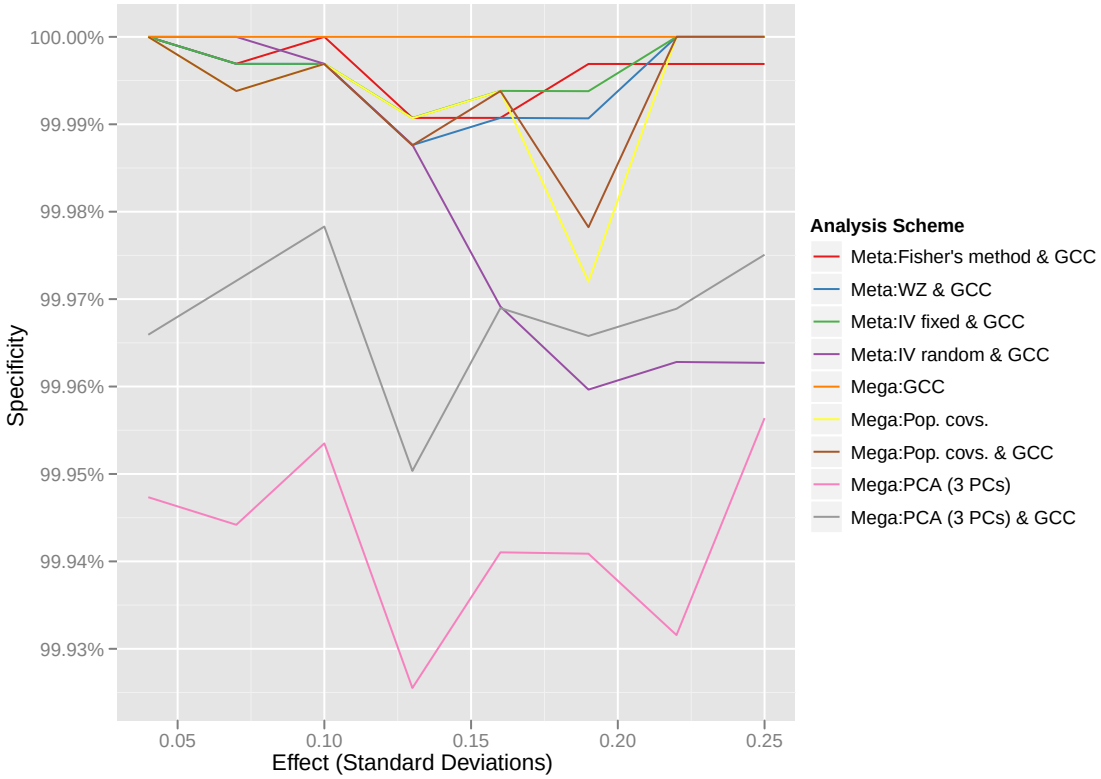


Figure 3.9: Specificity of the nine different analysis methods based on simulations of five different populations of 500 samples each.



Chapter 4

The future of large-scale genetic analyses

4.1 Mega-analysis of individual-level data

Mega-analysis of individual-level data (including genotype and phenotype data), as opposed to meta-analysis based only on summary statistics, requires a higher degree of data sharing among participating groups. However, when this data are available for analysis, greater power can be achieved (under some conditions; see Section 3), opportunities for error reduced, and more sophisticated analysis methods made possible.

In addition to the power analyses performed in Section 3, which applied only to common variants but which showed that mega-analysis has more power under some conditions, the power to analyse rare variants can also be improved by mega-analysis.

Because of the multi-level, stratified nature of meta-analysis, some rare variants will not be able to be analysed at all in a meta-analysis framework, while they could potentially be analysed in a mega-analysis. In an additive genetic association study of SNP data, a lower limit to the number of copies of the minor allele that must be observed in each individual study in order for the regression to produce stable results has been suggested to be in the range of 20–30[47]. It is suggested that meta-

analyses filter input from each study on criteria that maintain a minimum limit on the number of rare observations for each included SNP (see Section 1.3). Therefore, when employing such thresholds, a failure to observe a sufficient number of rare alleles at a variant within a particular study can result in the data from a particular study being entirely missing and therefore not able to contribute to the meta-analysis at all, effectively resulting in a limit on the MAF of SNPs that can be assessed in a meta-analysis that is proportionate to the number of studies that the meta-analysis is divided into (assuming they are of approximately equal size).

In contrast, combining individual-level data across all studies into a mega-analysis allows all available genotype data to contribute rare observations to an analysis, which should represent a substantial (in some cases going from no power to some power) increase in power to detect variants of low MAF.

There are multiple levels at which individual-level data sharing could reasonably occur. In terms of genotype data, this ranges from the raw (intensity) files through to imputed genotype uncertainty files.

4.1.1 Covariates

Analysing data from multiple studies as a single data set is not as straightforward as simply performing the same analysis that would be done on a single study on the combined data set. One issue is that, in the summary-statistics meta-analysis, individual studies were able to condition their analyses on covariates specific to each study, such as plate effects, DNA source tissue, extraction methods, or other batch effects[47]; while in the combined data set, such covariates will not be uniformly available so tests handling missing covariate data would need to be utilized in order to be able to include the study-specific covariates in the analysis.

The most easily handled covariate issue that is likely to arise is that continuous covariates are represented in different, but compatible, units. Provided that data

exchange formats included a unit specification for phenotypes, this could easily be handled using unit conversions. In fact, the pipeline we have developed for genome-wide association analysis (see Section 1.2.5.5) already contains code to automatically convert units if possible.

In addition, some categorical covariates may be present in multiple studies, but using different codings. Therefore, it is important to recognize which values are compatible across studies and can be treated as the same, and which are not compatible. For example, consider a meta-analysis in which we want to condition on the DNA source tissue as a covariate. Most studies in the analysis would probably only have a single source tissue, while one or two of the studies might have multiple source tissues. Perhaps one of the studies coded blood as '1' and saliva as '2,' while the other study might have coded another tissue as '1' and blood as '2' – clearly the values in one of the studies must be recoded so that they can be treated as a single covariate in the analysis. At the same time, none of the other studies would have included any data on source tissue since they only used one and thus would not have needed to condition on it. The source tissue from those other studies would have to be ascertained and coded in a compatible manner with those from the studies that did have variation.

Another type of covariate that may need special handling are those which are of the same type across studies, but in which values are specific to each study, such as the identifier for the lab technician who performed sample preparation. For studies that have collected the covariate, it should be sufficient to add a study-specific prefix to the identifiers in order to avoid collisions across studies. For studies that did not collect the covariate, a study-specific value can be used so that all samples for that study are treated as if they were prepared by one individual (but a different one than other studies).

Other covariates may be similar, but not directly interchangeable across studies. For example, there are several different technologies for directly measuring central fat content, including DXA scans, bioimpedance measurements, or skin fold measurements.

In some cases, it may be possible to create formulas to approximately convert these direct measures into compatible units (in this case, perhaps using them to estimate a percentage of central fat content).

Finally, the study itself could be included as a covariate to control for the large-scale batch effect of the entire study.

Population stratification correction, which may be performed separately in some of the individual studies of a summary-statistics meta-analysis, needs to be performed systematically for the entire individual-level meta-analysis. To facilitate this, methods for population stratification correction, such as the PCA methods described in Section 1.2.5.3, need to be extended to work with meta-analysis genotype data, which will typically have an extremely large number of samples and includes imputed data containing genotype uncertainties rather than called genotypes.

4.1.2 Privacy & informed consent

Perhaps the most significant obstacle to the sharing of individual-level data are privacy concerns, but there are technological measures that could potentially be used to address the concerns and make large-scale individual-level analyses across studies possible. Those same measures could also be part of a system that vastly improves the experience of subject participants.

4.1.2.1 Data protection

Individuals taking part in genetic research studies such as genome-wide association studies have done so under a variety of agreements made under informed consent (IC) which restrict the dissemination of their private data[340]. As a result, individual studies have diverse practices for protecting data as well as varying limitations on what data may be published, shared with other researchers, or used for research purposes other than those of the original study[340]. Laws in some jurisdictions also

provide specific protections applicable to genetic data gathered and/or used in those areas[341, 342].

4.1.2.2 Informed consent

Recently, critiques of the current practice in IC for genetic studies have been made, largely due to the fact that the required anonymisation of the study participants prevents much in the way of useful results from being returned to them[343]. Furthermore, it has become common practice to continue to re-analyse data after the initial experiment (e.g. in a meta-analysis), but typically the subjects receive no additional information about the new studies being done using their data[342].

In one sense, the old model for IC is no longer working since it relies on anonymisation to protect privacy, but as the genetic data itself is becoming increasingly individually identifiable as it becomes more and more dense, so that anonymity can be betrayed by the data itself.

Perhaps the creation of an online community could provide a way to give subjects much more control over what analyses are done with their genetic data and what privacy restrictions are in place to protect them. Not only could the principles behind IC be restored by giving participants control over and access to their own data, but researchers could potentially gain access to far more subjects through massively increased data sharing, and subjects would be easier to contact to request additional phenotyping or data collection throughout their lives.

4.1.2.3 Publishing of analysis results

There are obvious privacy concerns to making individual-level data from a study publicly available, as the genetic data alone can be used to identify a specific person or family[342], but even publishing the summary statistics from a study may trigger privacy concerns in that, given the genotypes of an individual, it is possible to estimate

the likelihood that the individual is a member of a particular study group, given only summary statistics such as allele frequencies or genotype counts from the genome-wide marker set used in that study[344].

By extension, it should also be possible to apply similar analyses to discover members in meta-analysis summary statistics, although increasing the number of individuals in the study while keeping the number of markers constant would tend to obscure the results by significantly reducing the confidence level. It may even be possible to make predictions regarding the inclusion of an individual in a study by looking at results (such as β, SE, or p-values) alone, without disseminating any explicit information about frequency, though this would be best facilitated in straightforward studies of simple categorical traits such as case/control studies, as the inclusion of quantitative phenotypes and/or covariates would make divining any information about individual participants exponentially more difficult.

One GWA case-control study on T2D[68] removed their analysis results from public dissemination on the web due to these concerns because it is an area that is not well understood.It would therefore be an interesting and useful area of research to explore more formally the amount of information contained in various sets of summary statistics, in order to facilitate more informed decisions about what results can be safely made public without violating the privacy of study participants.

4.1.3 Trusted analysis platform

Technological means could be used to enable the analysis (or meta-analysis) of individual level data while protecting the privacy of participants by preventing data used in the analyses from being directly accessed by anyone – even the analysts who are performing the analysis.

4.1.3.1 Platform architecture

Using cryptographic techniques including public key encryption and trusted secure co-processors, a system could be created that can perform meta-analysis of individual-level data wherein that data is never made available outside of the software used for analysis. Similar techniques are used in the financial industry to protect customer personal identification number (PIN) information, to store account data, and to allow for analysis of private customer transaction data for anti-money laundering purposes by providing pathways for analysis of data from different secure sources (effectively a meta-analysis of data from sources that do not trust each other or who are not allowed to trust each other with the private data)[345], so the creation of such a system for genetic analysis does not need to be built from the ground up, but can be based largely on existing technologies, though the scale and scope of the secure data storage and secure computation requirements for genetic analyses will likely be vastly larger and will therefore require a good deal of retooling.

The heart of most trusted computing platforms is a secure co-processor. A secure co-processor is a hardware device that is factory sealed with a private encryption key built-in to tamper-resistant hardware. They typically support outbound authentication to securely attest to the contents of their program memory. The hardware is typically signed at the time of manufacture by a private key held by the manufacturer, attesting to the fact that the hardware is a legitimate device. This signature provides the root certificate in a chain of certificates that can be used to trust that the software running on the co-processor is really what it is purported to be[345].

A trusted analysis platform could be built on top of this hardware to support analysis of data that could otherwise not be shared for joint analyses due to privacy concerns. Because such analyses would be computationally intensive, it would likely be infeasible to implement them entirely on (existing) secure hardware, but it may be possible to use virtualisation to enable a hybrid approach that extends the trust afforded

the secure co-processor to also trust virtual machine images running on the main system processors[346–348]. The virtual machine images themselves could be built by a community involving representatives of groups advocating for each study or on behalf of study participants, and ultimately the images could be cryptographically signed by a set of individuals entrusted by those groups to audit the software source code to ensure data protection limitations are being maintained within them.

4.1.3.2 Data

Using such a platform, representatives from a participating study (or indeed representatives of the participants themselves) could prepare their data for inclusion in the system by using a provided tool to transform it into a structured format including meta-information and flags regarding the extent of privacy protection required for the data, then encrypting it using the public key of the trusted analysis platform, signing it with a certificate belonging to the study, and uploading it to the analysis platform where it is then securely stored in a central encrypted database. From that point on, the only device capable of decrypting the stored data would be one authorised by the secure co-processor at the core of the trusted platform.

4.1.3.3 Analysis pipelines and modules

Analysts could then set up analysis pipelines that passed the data through a series of processing and analysis modules. During each step of the pipeline, the data would be accessed from the encrypted database, decrypted on the secure co-processor and re-encrypted to the internal key of an virtual machine analysis module whose image has been authenticated by the secure co-processor. Each module would perform data processing and/or analysis on the data and then encrypt and return the results to the secure co-processor. The secure co-processor would store the encrypted results in the database before moving on to the next step in the analysis pipeline. At the end of the pipeline, analysts could perform results queries to access generated plots,

statistics, or even input data; but access to this data would be restricted depending on the privacy flags provided by the constituent studies.

The analysis modules could be based entirely on open-source software [1] such that anyone could download the source, verify for themselves that there are no attempts to circumvent protections to allow either input or output data to be removed from the module except by sending it in encrypted form to the secure platform from which it was received, and compile the source to machine code to verify that the resulting image matches the signature of the image being used as a component to the system.

A set of auditors, perhaps from the study or consortium level or appointed by some sort of participant advocacy groups, could perform verification of the behaviour of analysis modules and subsequently sign them with their personal certificates in order to relieve the burden of auditing from each individual study. Individual studies, groups of participants, or even individual participants could set privacy settings that specify what types of modules are allowed to access their data, as well as requiring verification from specific auditors or from some number of verifications from a specific set of auditors to have signed components of the system in order to access data.

Since trust of analysis modules would be based on trust of groups of individual auditors, it would ultimately be up to the auditors to decide what software to trust, so it is possible that partially-closed-source software could be used as well, assuming that sufficient auditors representing the studies in a given analysis approved of the

[1] Building a system entirely on open-source software may seem restrictive, but the only way to really be certain of what a program does is to read the source. Of course, not everyone needs to be able to have access to the source to make that possible, but an argument can be made that analyses based on closed-source software are not very scientific. The crux of the issue comes down to reproducibility, in that it is crucial that published scientific results include a description of the methods used in enough detail for someone else to reproduce them independently. Historically, this meant a detailed set of steps to be followed in the laboratory, but today much of our methods are computational, and in reality, the methods we use are perfectly represented by the code itself, for the code is, after all, a set of steps for the computer to follow. It seems all too common for a methods section to contain a step which basically says 'we used software X to analyse the data.' Sometimes that software is closed-source and though, often (but not always), it does come with a citation, rarely does the cited paper even begin to describe what the program actually does in enough detail for another programmer to actually implement it independently—if the author was willing to do that, it would probably be open-source software already!

use of said software (whether by examining its closed source or through some other means).

The platform could also allow for data revocation or for updating to a new version of a study's meta-information, flags, or data that had previously been uploaded by signing a request with the same certificate used to upload the original data. Updates to permission flags could propagate to results already present in the system, such that results that were previously unavailable could become available (or vice-versa) and results that had been generated with old data would automatically be marked as being out-of-date in the instance when a new version of the data was uploaded.

The trusted analysis platform could enable analyses that would be otherwise impossible to perform due to legal concerns surrounding sharing of private subject data. It would allow studies to collectively analyse data without having to trust each other or any individual analyst with the data. It could be based upon cryptographic standards and hardware protections that meet some of the highest levels of security (US NIST FIPS 140-1 Security Level 4)[349], such that any attempts to steal the private key of the secure co-processor in order to attain unlimited access to the data would result in the destruction of the private key, permanently rendering useless all of the results and data in the secure database (though, of course, it would be possible to start again with a new system, having each study re-upload their data, and re-running all analyses).

Additionally, the trusted platform can record and provide digitally signed attestation to the complete chain of software used to run a particular analysis along with the results, which should provide a greater degree of confidence in the published results as well as allowing reproduction of the analyses by third parties.

4.1.3.4 Informed consent within a trusted platform

Though a bit more difficult to implement, a trusted analysis platform could be designed not as a monolithic system but rather as a distributed (but probably federated) system that itself could be incorporated into a larger online system for managing private genetic data along with the analysis results that come out of that data. It could be part of a new model for IC that protects privacy through encryption rather than anonymisation, and that can provide stakeholders in the data (such as the participants or their relatives) with the results of analyses performed using their data, with increased control over what analyses can be done on their data (whether through setting privacy policies or by making day-to-day decisions), and with direct access to their own genetic data. Of course, direct access to their data would only be useful if provided along with analysis and visualisation tools to make that access useful and meaningful, but a number of commercial ventures, such as 23andMe (`http://23andme.com/`), deCODEme (`http://decodeme.com/`, Navigenics (`http://www.navigenics.com/`), and Knome (`http://www.knome.com/`) have demonstrated that there is a demand for providing personal genetic data, along with interpretation, to individuals. It would seem to make a lot of sense for study participants to be able to access similar information for free as part of the reward for participation in a study, and replacing the veil of anonymity required by IC practices of the past with a technological system based on encryption to protect privacy could make participation in studies more rewarding for participants as well as researchers.

4.1.3.5 Cost of a trusted analysis platform

The cost of a trusted analysis platform capable of providing for the analysis of genetic data while cryptographically maintaining the privacy of the data is dependent on many factors, the most significant of which is what level of protection is needed.

Implementation of a system in which all private data analyses occurs only on cryptographic co-processors in physically protected memory would require the development of new trusted computing hardware with increased computational and memory resources, and would likely be quite difficult to scale. If, however, it is acceptable for the data to be analysed within a secure virtual machine it would be much easier to scale up by running additional virtual machine instances, a number of which could potentially share the same hardware trusted computing module. The drawback would be that in order to be processed by the virtual machine, the unencrypted data would at some point reside in main (unprotected) memory. Although the virtual machine could keep data about the format and location of data structures kept in main memory private by storing them within the trusted module, and could minimize the amount of time and the amount of unencrypted data that is kept in main memory at any one time, it would still be theoretically possible for an attacker with physical access to the hardware to intercept all data writes to main memory and analyze it to reconstruct portions of the unencrypted data.

In order to ensure security of a system based on trusted virtual machines, the hardware it runs on would have to be physically secured by more traditional means. This is likely to be an acceptable compromise since existing private data such as traditional medical records are kept in paper form and thus security of such records is provided primarily from physical security.

The cost of such a virtual machine based system in terms of hardware would be comparable to current systems used to store and analyse genetic data, since a trusted virtual machine could be run on commodity hardware with the additional of a trusted platform module as a peripheral device, which could potentially be shared between multiple virtual machines running in a cluster environment. An example of such a peripheral device is the IBM PCIe Cryptographic Coprocessor[350], which had a list price of $14,408 as of August 2010[351]. Another substantial cost would be the initial cost of software research & development to develop the platform for genetic analysis

including a framework for running analyses as well interfaces for analysis module verification.

References

[1] L.A. Hindorff, P. Sethupathy, H.A. Junkins, E.M. Ramos, J.P. Mehta, F.S. Collins, et al. "Potential etiologic and functional implications of genome-wide association loci for human diseases and traits". In: *Proceedings of the National Academy of Sciences* 106.23 (2009), p. 9362.

[2] M.I. McCarthy, G.R. Abecasis, L.R. Cardon, D.B. Goldstein, J. Little, J.P.A. Ioannidis, et al. "Genome-wide association studies for complex traits: consensus, uncertainty and challenges". In: *Nature Reviews Genetics* 9.5 (2008), pp. 356–369.

[3] Cecilia M. Lindgren, Iris M. Heid, Joshua C. Randall, Claudia Lamina, Valgerdur Steinthorsdottir, Lu Qi, et al. "Genome-Wide Association Scan Meta-Analysis Identifies Three Loci Influencing Adiposity and Fat Distribution". In: *PLoS Genet* 5.6 (June 2009), e1000508.

[4] Ruth J F Loos, Cecilia M Lindgren, Shengxu Li, Eleanor Wheeler, Jing Hua Zhao, Inga Prokopenko, et al. "Common variants near MC4R are associated with fat mass, weight and risk of obesity". In: *Nat Genet* 40.6 (June 2008), pp. 768–775.

[5] Cristen J Willer, Elizabeth K Speliotes, Ruth J F Loos, Shengxu Li, Cecilia M Lindgren, Iris M Heid, et al. "Six new loci associated with body mass index highlight a neuronal influence on body weight regulation." eng. In: *Nat Genet* 41.1 (2009), pp. 25–34. ISSN: 1546-1718 (Electronic).

[6] D. Altshuler and M. Daly. "Guilt beyond a reasonable doubt". In: *Nature genetics* 39.7 (2007), pp. 813–815.

[7] P.M. Visscher, W.G. Hill, and N.R. Wray. "Heritability in the genomics era—concepts and misconceptions". In: *Nature Reviews Genetics* 9.4 (2008), pp. 255–266.

[8] A.J. Walley, J.E. Asher, and P. Froguel. "The genetic contribution to non-syndromic human obesity". In: *Nature Reviews Genetics* 10.7 (2009), pp. 431–442.

[9] M. Hofker and C. Wijmenga. "A supersized list of obesity genes". In: *Nature Genetics* 41.2 (2009), pp. 139–140.

[10] T Strachan and Andrew P. Read. *Human molecular genetics 3*. 3rd ed. London: Garland Press, 2004. ISBN: 0815341849 (pbk.)

[11] Gregor Mendel. “Experiments in Plant Hybridization”. In: *Proceedings of the Brünn Natural History Society* IV (1865). Trans. William Bateson. Ed. Roger Blumberg, 1996.

[12] A.O.T. MacLeod and M. McCarty. “Studies of the chemical nature of the substance inducing transformation of pneumococcal types. induction of transformation by a deoxyribonucleic acid fraction isolated from pneumococcus type iii”. In: *Journal of Experimental Medicine* 79 (1944), pp. 137–158.

[13] AD Hershey and M. Chase. “Independent functions of viral protein and nucleic acid in growth of bacteriophage”. In: *Journal of General Physiology* 36.1 (1952), pp. 39–56.

[14] J.D. Watson and F.H.C. Crick. “A Structure for Deoxyribose Nucleic Acid”. In: *Nature* 171.4356 (1953), pp. 737–738.

[15] J.D. Watson and F.H.C. Crick. “Genetical Implications of the Structure of Deoxyribonucleic Acid”. In: *Nature* 171.4361 (1953), pp. 964–967.

[16] Trudy F. C. Mackay, Eric A. Stone, and Julien F. Ayroles. “The genetics of quantitative traits: challenges and prospects”. In: *Nat Rev Genet* 10.8 (Aug. 2009), pp. 565–577.

[17] T.F.C. Mackay. “The genetic architecture of quantitative traits”. In: *Annual Review of Genetics* 35.1 (2001), pp. 303–339.

[18] ES Lander and D. Botstein. “Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps”. In: *Genetics* 121.1 (1989), pp. 185–199.

[19] K.L. Gunderson, F.J. Steemers, G. Lee, L.G. Mendoza, and M.S. Chee. “A genome-wide scalable SNP genotyping assay using microarray technology”. In: *Nature genetics* 37.5 (2005), pp. 549–554.

[20] K.A. Frazer, D.G. Ballinger, D.R. Cox, D.A. Hinds, L.L. Stuve, R.A. Gibbs, et al. “A second generation human haplotype map of over 3.1 million SNPs”. In: *Nature* 449.7164 (2007), pp. 851–861.

[21] David J. Balding. “A tutorial on statistical methods for population association studies”. In: *Nat Rev Genet* 7.10 (Oct. 2006), pp. 781–791.

[22] RC Lewontin and K. Kojima. “The evolutionary dynamics of complex polymorphisms”. In: *Evolution* 14.4 (1960), pp. 458–472. ISSN: 0014-3820.

[23] J.N. Hirschhorn and M.J. Daly. “Genome-wide association studies for common diseases and complex traits”. In: *Nature Reviews Genetics* 6.2 (2005), pp. 95–108.

[24] D. Altshuler, L.D. Brooks, A. Chakravarti, F.S. Collins, M.J. Daly, P. Donnelly, et al. “A haplotype map of the human genome”. In: *Nature* 437.7063 (2005), pp. 1299–1320.

[25] P. Hedrick and S. Kumar. “Mutation and linkage disequilibrium in human mtDNA.” In: *European journal of human genetics: EJHG* 9.12 (2001), p. 969. ISSN: 1018-4813.

[26] C.A. Anderson, F.H. Pettersson, J.C. Barrett, J.J. Zhuang, J. Ragoussis, L.R. Cardon, et al. “Evaluating the effects of imputation on the power, coverage, and cost efficiency of genome-wide SNP platforms”. In: *The American Journal of Human Genetics* 83.1 (2008), pp. 112–119.

[27] J. Marchini, B. Howie, S. Myers, G. McVean, and P. Donnelly. "A new multipoint method for genome-wide association studies by imputation of genotypes". In: *Nature genetics* 39.7 (2007), pp. 906–913.

[28] *1000 Genomes Project*. URL: http://www.1000genomes.org/.

[29] James A. Morris, Joshua C. Randall, Julian B. Maller, and Jeffrey C. Barrett. "Evoker: a visualization tool for genotype intensity data". In: *Bioinformatics* (2010), btq280.

[30] J.M. Korn, F.G. Kuruvilla, S.A. McCarroll, A. Wysoker, J. Nemesh, S. Cawley, et al. "Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs". In: *Nature Genetics* 40.10 (2008), pp. 1253–1260.

[31] Chris Spencer. *CHIAMO (v0.2.1)*. 2008. URL: https://mathgen.stats.ox.ac.uk/genetics_software/chiamo/chiamo.html.

[32] J. Hua, D.W. Craig, M. Brun, J. Webster, V. Zismann, W. Tembe, et al. "SNiPer-HD: improved genotype calling accuracy by an expectation-maximization algorithm for high-density SNP arrays". In: *Bioinformatics* 23.1 (2007), p. 57.

[33] Y.Y. Teo, M. Inouye, K.S. Small, R. Gwilliam, P. Deloukas, D.P. Kwiatkowski, et al. "A genotype calling algorithm for the Illumina BeadArray platform". In: *Bioinformatics* 23.20 (2007), p. 2741.

[34] Y. Xiao, M.R. Segal, YH Yang, and R.F. Yeh. "A multi-array multi-SNP genotyping algorithm for Affymetrix SNP microarrays". In: *Bioinformatics* 23.12 (2007), p. 1459.

[35] N. Rabbee and T.P. Speed. "A genotype calling algorithm for affymetrix SNP arrays". In: *Bioinformatics* 22.1 (2006), pp. 7–12.

[36] Shaun Purcell. *PLINK (v1.07)*. 2009. URL: http://pngu.mgh.harvard.edu/~purcell/plink/.

[37] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M.A.R. Ferreira, D. Bender, et al. "PLINK: a tool set for whole-genome association and population-based linkage analyses". In: *The American Journal of Human Genetics* 81.3 (2007), pp. 559–575.

[38] Shaun Purcell. *PLINK (v1.06)*. 2009. URL: http://pngu.mgh.harvard.edu/~purcell/plink/.

[39] Jonathan Marchini. *SNPTEST (v1.1.5)*. 2007. URL: https://mathgen.stats.ox.ac.uk/genetics_software/snptest/snptest.html.

[40] Z. Kutalik, T. Johnson, M. Bochud, V. Mooser, P. Vollenweider, G. Waeber, et al. "Methods for testing association between uncertain genotypes and quantitative traits". In: *Biostatistics* 12.1 (2011), p. 1.

[41] R Development Core Team. *R: A Language and Environment for Statistical Computing*. ISBN 3-900051-07-0. R Foundation for Statistical Computing. Vienna, Austria, 2009. URL: http://www.R-project.org.

[42] Hadley Wickham. *ggplot2: elegant graphics for data analysis*. 1st ed. Vol. 6991. Use r. New York: Springer, 2009. ISBN: 9780387981406 (softcover : alk. paper).

[43] Hadley Wickham. "Reshaping Data with the reshape Package". In: *Journal of Statistical Software* 21.12 (Sept. 2007), pp. 1–20. ISSN: 1548-7660.

[44] Cynthia A. Brewer. *Color Brewer*. 2009. URL: http://www.colorbrewer.org/.

[45] Wellcome Trust Case Control Consortium. "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls." eng. In: *Nature* 447.7145 (2007), pp. 661–678. ISSN: 1476-4687 (Electronic).

[46] D.M. Altshuler, R.A. Gibbs, L. Peltonen, S.F. Schaffner, F. Yu, E. Dermitzakis, et al. "Integrating common and rare genetic variation in diverse human populations". In: *Nature* 467 (2010), pp. 52–58.

[47] BM Neale and S. Purcell. "The positives, protocols, and perils of genome-wide association." In: *Am J Med Genet B Neuropsychiatr Genet* (2008).

[48] Curt Stern. "The Hardy-Weinberg Law". In: *Science*. New Series 97.2510 (1943), pp. 137–138. ISSN: 00368075.

[49] J.K. Wittke-Thompson, A. Pluzhnikov, and N.J. Cox. "Rational inferences about departures from Hardy-Weinberg equilibrium". In: *The American Journal of Human Genetics* 76.6 (2005), pp. 967–986.

[50] J.E. Wigginton, D.J. Cutler, and G.R. Abecasis. "A note on exact tests of Hardy-Weinberg equilibrium". In: *The American Journal of Human Genetics* 76.5 (2005), pp. 887–893.

[51] National Institute of Standards, Technology (US), C. Croarkin, P. Tobias, C. Zey, and International SEMATECH. *Engineering statistics handbook*. The Institute, 2001.

[52] J. P. Royston. "An Extension of Shapiro and Wilk's W Test for Normality to Large Samples". In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 31.2 (1982), pp. 115–124. ISSN: 00359254.

[53] G.E. Dallal and L. Wilkinson. "An analytic approximation to the distribution of Lilliefors's test statistic for normality". In: *American Statistician* (1986), pp. 294–296.

[54] Iris M Heid, Anne U Jackson, Joshua C Randall, Thomas W Winkler, Lu Qi, Valgerdur Steinthorsdottir, et al. "Meta-analysis identifies 13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution". In: *Nat Genet* 42.11 (Nov. 2010), pp. 949–960.

[55] Elizabeth K Speliotes, Cristen J Willer, Sonja I Berndt, Keri L Monda, Gudmar Thorleifsson, Anne U Jackson, et al. "Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index". In: *Nat Genet* 42.11 (Nov. 2010), pp. 937–948.

[56] S. Van Gestel, J.J. Houwing-Duistermaat, R. Adolfsson, C.M. van Duijn, and C. Van Broeckhoven. "Power of selective genotyping in genetic association analyses of quantitative traits". In: *Behavior Genetics* 30.2 (2000), pp. 141–146.

[57] T.M. Frayling, N.J. Timpson, M.N. Weedon, E. Zeggini, R.M. Freathy, C.M. Lindgren, et al. "A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity". In: *Science* 316.5826 (2007), p. 889.

[58] B. Servin and M. Stephens. "Imputation-based analysis of association studies: candidate regions and quantitative traits". In: *PLoS Genet* 3.7 (2007), e114.

[59] P. Scheet and M. Stephens. "A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase". In: *The American Journal of Human Genetics* 78.4 (2006), pp. 629–644.

[60] DY Lin, Y. Hu, and BE Huang. "Simple and efficient analysis of disease association with missing genotype data". In: *The American Journal of Human Genetics* 82.2 (2008), pp. 444–452.

[61] D.L. Nicolae. "Testing untyped alleles (TUNA)-applications to genome-wide association studies". In: *Genetic epidemiology* 30.8 (2006).

[62] B.L. Browning and S.R. Browning. "A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals". In: *The American Journal of Human Genetics* 84.2 (2009), pp. 210–223.

[63] BN Howie, P. Donnelly, and J. Marchini. "A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association". In: *PLoS Genet* (2009).

[64] Yun Li, Cristen Willer, Serena Sanna, and Gonçalo Abecasis. "Genotype Imputation". In: *Annual Review of Genomics and Human Genetics* 10.1 (2009). PMID: 19715440, pp. 387–406.

[65] Inga Prokopenko, Claudia Langenberg, Jose C Florez, Richa Saxena, Nicole Soranzo, Gudmar Thorleifsson, et al. "Variants in MTNR1B influence fasting glucose levels". In: *Nat Genet* 41.1 (Jan. 2009), pp. 77–81.

[66] J. Marchini and B. Howie. "Genotype imputation for genome-wide association studies". In: *Nat Rev Genet* 11.7 (2010), pp. 499–511.

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

[68] E. Zeggini, LJ Scott, R. Saxena, and BF Voight. "Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes". In: *Nat Genet* 40 (2008), pp. 638–645.

[69] Josee Dupuis, Claudia Langenberg, Inga Prokopenko, Richa Saxena, Nicole Soranzo, Anne U Jackson, et al. "New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk". In: *Nature Genetics* 42.2 (Feb. 2010), pp. 105–116.

[70] Hana Lango Allen, Karol Estrada, Guillaume Lettre, Sonja I. Berndt, Michael N. Weedon, Fernando Rivadeneira, et al. "Hundreds of variants clustered in genomic loci and biological pathways affect human height". In: *Nature* 467.7317 (Oct. 2010), pp. 832–838.

[71] M.N. Weedon, H. Lango, C.M. Lindgren, C. Wallace, D.M. Evans, M. Mangino, et al. "Genome-wide association analysis identifies 20 loci that influence adult height". In: *Nature Genetics* 40.5 (2008), pp. 575–583.

[72] Y. Guan and M. Stephens. "Bayesian variable selection regression for genome-wide association studies, and other large-scale problems". In: *Ann. Appl. Stat* (2011).

[73] G. Lettre, C. Lange, and J.N. Hirschhorn. "Genetic model testing and statistical power in population-based association studies of quantitative traits". In: *Genetic Epidemiology* 31.4 (2007).

[74] J. Marchini, L.R. Cardon, M.S. Phillips, and P. Donnelly. "The effects of human population structure on large genetic association studies". In: *Nature genetics* 36 (2004), pp. 512–517.

[75] D.G. Clayton, N.M. Walker, D.J. Smyth, R. Pask, J.D. Cooper, L.M. Maier, et al. "Population structure, differential bias and genomic control in a large-scale, case-control association study". In: *Nature genetics* 37.11 (2005), pp. 1243–1246.

[76] B. Devlin and K. Roeder. "Genomic Control for Association Studies". In: *Biometrics* 55.4 (1999), pp. 997–1004.

[77] B. Devlin, K. Roeder, and L. Wasserman. "Genomic Control, a New Approach to Genetic-Based Association Studies". In: *Theoretical Population Biology* 60.3 (2001), pp. 155–166.

[78] B. Devlin, S.A. Bacanu, and K. Roeder. "Genomic control to the extreme". In: *Nature genetics* 36.11 (2004), pp. 1129–1130.

[79] M.L. Freedman, D. Reich, K.L. Penney, G.J. McDonald, A.A. Mignault, N. Patterson, et al. "Assessing the impact of population stratification on genetic association studies". In: *Nature genetics* 36.4 (2004), pp. 388–393.

[80] J.E. Jackson, J. Wiley, and W. InterScience. *A user's guide to principal components*. 1991.

[81] D. Reich, A.L. Price, and N. Patterson. "Principal component analysis of genetic data". In: *Nature Genetics* (2008).

[82] A.L. Price, N.J. Patterson, R.M. Plenge, M.E. Weinblatt, N.A. Shadick, and D. Reich. "Principal components analysis corrects for stratification in genome-wide association studies". In: *Nature genetics* 38 (2006), pp. 904–909.

[83] Nick Patterson, Alkes L Price, and David Reich. "Population Structure and Eigenanalysis". In: *PLoS Genet* 2.12 (2006), e190.

[84] X. Zhu, S. Li, R.S. Cooper, and R.C. Elston. "A unified association analysis approach for family and unrelated samples correcting for stratification". In: *The American Journal of Human Genetics* 82.2 (2008), pp. 352–365.

[85] Yun Li, Cristen J Willer, Jun Ding, Paul Scheet, and Gonçalo R Abecasis. "MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes". In: *Genet Epidemiol* 34.8 (2010), pp. 816–34.

[86] R Mott, C J Talbot, M G Turri, A C Collins, and J Flint. "A method for fine mapping quantitative trait loci in outbred animal stocks". In: *Proc Natl Acad Sci U S A* 97.23 (2000), pp. 12649–54.

[87] Cristen J Willer, Serena Sanna, Anne U Jackson, Angelo Scuteri, Lori L Bonnycastle, Robert Clarke, et al. "Newly identified loci that influence lipid concentrations and risk of coronary artery disease". In: *Nat Genet* 40.2 (2008), pp. 161–9.

[88] J. Zheng, Y. Li, G.R. Abecasis, and P. Scheet. "A comparison of approaches to account for uncertainty in analysis of imputed genotypes". In: *Genetic Epidemiology* 35.2 (2011), pp. 102–110.

[89] Z. Su, J. Marchini, and P. Donnelly. "HAPGEN2: simulation of multiple disease SNPs." In: *Bioinformatics* (2011).

[90] M C Whitlock. "Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach". In: *J Evol Biol* 18.5 (2005), pp. 1368–73.

[91] S.R.A. Fisher. *Statistical methods for research workers*. 5. Genesis Publishing Pvt Ltd, 1932.

[92] Frederick Mosteller and R. A. Fisher. "Questions and Answers". In: *The American Statistician* 2.5 (Oct. 1948), pp. 30–31.

[93] S.A. Stouffer, E.A. Suchman, L.C. Devinney, S.A. Star, and R.M. Williams Jr. "The American soldier: adjustment during army life.(Studies in social psychology in World War II, Vol. 1.)." In: (1949).

[94] B WOOLF. "On estimating the relation between blood group and disease". In: *Ann Hum Genet* 19.4 (1955), pp. 251–3.

[95] A.J. Sutton and K.R. Abrams. "Bayesian methods in meta-analysis and evidence synthesis". In: *Statistical Methods in Medical Research* 10.4 (2001), p. 277.

[96] R. DerSimonian, N. Laird, et al. "Meta-analysis in clinical trials". In: *Control Clin Trials* 7.3 (1986), pp. 177–188.

[97] B.L. Welch. "THE GENERALIZATION OF'STUDENT'S'PROBLEM WHEN SEVERAL DIFFERENT POPULATION VARLANCES ARE INVOLVED". In: *Biometrika* 34.1-2 (1947), pp. 28–35.

[98] FE Satterthwaite. "An approximate distribution of estimates of variance components". In: *Biometrics Bulletin* (1946), pp. 110–114.

[99] Cristen J Willer, Yun Li, and Gonçalo R Abecasis. "METAL: fast and efficient meta-analysis of genomewide association scans". In: *Bioinformatics* 26.17 (2010), pp. 2190–1.

[100] N. Risch and K. Merikangas. "The future of genetic studies of complex human diseases". In: *Science* 273.5281 (1996), p. 1516.

[101] Xiaoyi Gao, Lewis C Becker, Diane M Becker, Joshua D Starmer, and Michael A Province. "Avoiding the high Bonferroni penalty in genome-wide association studies". In: *Genet Epidemiol* 34.1 (2010), pp. 100–5.

[102] F. Dudbridge and A. Gusnanto. "Estimation of significance thresholds for genomewide association scans". In: *Genetic epidemiology* 32.3 (2008), p. 227.

[103] Y. Benjamini and Y. Hochberg. "Controlling the false discovery rate: a practical and powerful approach to multiple testing". In: *Journal of the Royal Statistical Society. Series B (Methodological)* (1995), pp. 289–300.

[104] J.D. Storey and R. Tibshirani. "Statistical significance for genomewide studies". In: *Proceedings of the National Academy of Sciences* 100.16 (2003), pp. 9440–9445.

[105] M.E. Tabangin, J.G. Woo, C. Liu, T.G. Nick, and L.J. Martin. "Comparison of false-discovery rate for genome-wide and fine mapping regions". In: *feedback* (2008).

[106] K. Strimmer. "A unified approach to false discovery rate estimation". In: *BMC Bioinformatics* 9.1 (2008), p. 303.

[107] L.A. Hindorff, H.A. Junkins, P.N. Hall, J.P. Mehta, and T.A. Manolio. *A Catalog of Published Genome-Wide Association Studies*. Aug. 2011. URL: `http://www.genome.gov/gwastudies`.

[108] A. Scuteri, S. Sanna, W.M. Chen, M. Uda, G. Albai, J. Strait, et al. "Genome-wide association scan shows genetic variants in the FTO gene are associated with obesity-related traits". In: *PLoS Genet* 3.7 (2007), e115.

[109] T Pischon, H Boeing, K Hoffmann, M Bergmann, M B Schulze, K Overvad, et al. "General and abdominal adiposity and risk of death in Europe". In: *N Engl J Med* 359.20 (2008), pp. 2105–20.

[110] I Baik, A Ascherio, E B Rimm, E Giovannucci, D Spiegelman, M J Stampfer, et al. "Adiposity and mortality in men". In: *Am J Epidemiol* 152.3 (2000), pp. 264–71.

[111] Salim Yusuf, Steven Hawken, Stephanie Ounpuu, Leonelo Bautista, Maria Grazia Franzosi, Patrick Commerford, et al. "Obesity and the risk of myocardial infarction in 27,000 participants from 52 countries: a case-control study". In: *Lancet* 366.9497 (2005), pp. 1640–9.

[112] Karoline Schousboe, Gonneke Willemsen, Kirsten O Kyvik, Jakob Mortensen, Dorret I Boomsma, Belinda K Cornes, et al. "Sex differences in heritability of BMI: a comparative study of results from twin studies in eight countries". In: *Twin Res* 6.5 (2003), pp. 409–21.

[113] KM Rose, B. Newman, EJ Mayer-Davis, and JV Selby. "Genetic and behavioral determinants of waist-hip ratio and waist circumference in women twins." In: *Obesity research* 6.6 (1998), p. 383.

[114] Christian Dina, David Meyre, Sophie Gallina, Emmanuelle Durand, Antje Körner, Peter Jacobson, et al. "Variation in FTO contributes to childhood obesity and severe adult obesity". In: *Nat Genet* 39.6 (2007), pp. 724–6.

[115] Thomas Gerken, Christophe A Girard, Yi-Chun Loraine Tung, Celia J Webby, Vladimir Saudek, Kirsty S Hewitson, et al. "The obesity-associated FTO gene encodes a 2-oxoglutarate-dependent nucleic acid demethylase". In: *Science* 318.5855 (2007), pp. 1469–72.

[116] John C Chambers, Paul Elliott, Delilah Zabaneh, Weihua Zhang, Yun Li, Philippe Froguel, et al. "Common genetic variation near MC4R is associated with waist circumference and insulin resistance". In: *Nat Genet* 40.6 (2008), pp. 716–8.

[117] Stephen O'Rahilly and I Sadaf Farooqi. "Human obesity: a heritable neurobehavioral disorder that is highly sensitive to environmental conditions". In: *Diabetes* 57.11 (2008), pp. 2905–10.

[118] M C Zillikens, M Yazdanpanah, L M Pardo, F Rivadeneira, Y S Aulchenko, B A Oostra, et al. "Sex-specific genetic effects influence variation in body composition". In: *Diabetologia* 51.12 (2008), pp. 2233–41.

[119] P.A. Fujita, B. Rhead, A.S. Zweig, A.S. Hinrichs, D. Karolchik, M.S. Cline, et al. "The UCSC Genome Browser database: update 2011". In: *Nucleic Acids Research* 39.suppl 1 (2011), p. D876.

[120] R.J. Pruim, R.P. Welch, S. Sanna, T.M. Teslovich, P.S. Chines, T.P. Gliedt, et al. "LocusZoom: regional visualization of genome-wide association scan results". In: *Bioinformatics* 26.18 (2010), p. 2336.

[121] Gregory R Steinberg, Bruce E Kemp, and Matthew J Watt. "Adipocyte triglyceride lipase expression in human obesity". In: *Am J Physiol Endocrinol Metab* 293.4 (2007), E958–64.

[122] J.G. Smith, J.K. Lowe, S. Kovvali, J.B. Maller, J. Salit, M.J. Daly, et al. "Genome-wide association study of electrocardiographic conduction measures in an isolated founder population: Kosrae". In: *Heart Rhythm* 6.5 (2009), pp. 634–641.

[123] Kim D Pruitt, Tatiana Tatusova, and Donna R Maglott. "NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins". In: *Nucleic Acids Res* 35.Database issue (2007), pp. D61–5.

[124] Shiro Maeda, Shuichi Tsukada, Akio Kanazawa, Akihiro Sekine, Tatsuhiko Tsunoda, Daisuke Koya, et al. "Genetic variations in the gene encoding TFAP2B are associated with type 2 diabetes mellitus". In: *J Hum Genet* 50.6 (2005), pp. 283–92.

[125] Shuichi Tsukada, Yasushi Tanaka, Hiroshi Maegawa, Atsunori Kashiwagi, Ryuzo Kawamori, and Shiro Maeda. "Intronic polymorphisms within TFAP2B regulate transcriptional activity and affect adipocytokine gene expression in differentiated adipocytes". In: *Mol Endocrinol* 20.5 (2006), pp. 1104–11.

[126] Kazuhiro Ikeda, Hiroshi Maegawa, Satoshi Ugi, Yukari Tao, Yoshihiko Nishio, Shuichi Tsukada, et al. "Transcription factor activating enhancer-binding protein-2beta. A negative regulator of adiponectin gene expression". In: *J Biol Chem* 281.42 (2006), pp. 31245–53.

[127] M A Rahman, H Nelson, H Weissbach, and N Brot. "Cloning, sequencing, and expression of the Escherichia coli peptide methionine sulfoxide reductase gene". In: *J Biol Chem* 267.22 (1992), pp. 15549–51.

[128] J Moskovitz, S Bar-Noy, W M Williams, J Requena, B S Berlett, and E R Stadtman. "Methionine sulfoxide reductase (MsrA) is a regulator of antioxidant defense and lifespan in mammals". In: *Proc Natl Acad Sci U S A* 98.23 (2001), pp. 12920–5.

[129] Sarah de Ferranti and Dariush Mozaffarian. "The perfect storm: obesity, adipocyte dysfunction, and metabolic consequences". In: *Clin Chem* 54.6 (2008), pp. 945–55.

[130] Y.S. Aulchenko, S. Ripatti, I. Lindqvist, D. Boomsma, I.M. Heid, P.P. Pramstaller, et al. "Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts". In: *Nature genetics* 41.1 (2008), pp. 47–55.

[131] V J Carey, E E Walters, G A Colditz, C G Solomon, W C Willett, B A Rosner, et al. "Body fat distribution and risk of non-insulin-dependent diabetes mellitus in women. The Nurses' Health Study". In: *Am J Epidemiol* 145.7 (1997), pp. 614–9.

[132] Youfa Wang, Eric B Rimm, Meir J Stampfer, Walter C Willett, and Frank B Hu. "Comparison of abdominal adiposity and overall obesity in predicting risk of type 2 diabetes among men". In: *Am J Clin Nutr* 81.3 (2005), pp. 555–63.

[133] Dexter Canoy. "Distribution of body fat and risk of coronary heart disease in men and women". In: *Curr Opin Cardiol* 23.6 (2008), pp. 591–8.

[134] Marieke B Snijder, Jacqueline M Dekker, Marjolein Visser, Lex M Bouter, Coen D A Stehouwer, Piet J Kostense, et al. "Associations of hip and thigh circumferences independent of waist circumference with the incidence of type 2 diabetes: the Hoorn Study". In: *Am J Clin Nutr* 77.5 (2003), pp. 1192–7.

[135] Marieke B Snijder, Jacqueline M Dekker, Marjolein Visser, Lex M Bouter, Coen D A Stehouwer, John S Yudkin, et al. "Trunk fat and leg fat have independent and opposite associations with fasting and postload glucose levels: the Hoorn study". In: *Diabetes Care* 27.2 (2004), pp. 372–7.

[136] G W Mills, P J Avery, M I McCarthy, A T Hattersley, J C Levy, G A Hitman, et al. "Heritability estimates for beta cell function and features of the insulin resistance syndrome in UK families with an increased susceptibility to type 2 diabetes". In: *Diabetologia* 47.4 (2004), pp. 732–8.

[137] N Y Souren, A D C Paulussen, R J F Loos, M Gielen, G Beunen, R Fagard, et al. "Anthropometry, carbohydrate and lipid metabolism in the East Flanders Prospective Twin Survey: heritabilities". In: *Diabetologia* 50.10 (2007), pp. 2107–16.

[138] J V Selby, B Newman, C P Quesenberry Jr, R R Fabsitz, D Carmelli, F J Meaney, et al. "Genetic and behavioral influences on body fat distribution". In: *Int J Obes* 14.7 (1990), pp. 593–602.

[139] Anil K Agarwal and Abhimanyu Garg. "Genetic disorders of adipose tissue development, differentiation, and death". In: *Annu Rev Genomics Hum Genet* 7 (2006), pp. 175–99.

[140] Abhimanyu Garg. "Acquired and inherited lipodystrophies". In: *N Engl J Med* 350.12 (2004), pp. 1220–34.

[141] B. Rosner. *Fundamentals of biostatistics*. Duxbury Pr, 2010.

[142] P.D. Thomas, M.J. Campbell, A. Kejariwal, H. Mi, B. Karlak, R. Daverman, et al. "PANTHER: a library of protein families and subfamilies indexed by function". In: *Genome research* 13.9 (2003), p. 2129.

[143] Steven A McCarroll, Finny G Kuruvilla, Joshua M Korn, Simon Cawley, James Nemesh, Alec Wysoker, et al. "Integrated detection and population-genetic analysis of SNPs and copy number variation". In: *Nat Genet* 40.10 (2008), pp. 1166–74.

[144] Donald F Conrad, Dalila Pinto, Richard Redon, Lars Feuk, Omer Gokcumen, Yujun Zhang, et al. "Origins and functional impact of copy number variation in the human genome". In: *Nature* 464.7289 (2010), pp. 704–12.

[145] Wellcome Trust Case Control Consortium, Nick Craddock, Matthew E Hurles, Niall Cardin, Richard D Pearson, Vincent Plagnol, et al. "Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls". In: *Nature* 464.7289 (2010), pp. 713–20.

[146] Anna L Dixon, Liming Liang, Miriam F Moffatt, Wei Chen, Simon Heath, Kenny C C Wong, et al. "A genome-wide association study of global gene expression". In: *Nat Genet* 39.10 (2007), pp. 1202–7.

[147] Eric E Schadt, Cliona Molony, Eugene Chudin, Ke Hao, Xia Yang, Pek Y Lum, et al. "Mapping the genetic architecture of gene expression in human liver". In: *PLoS Biol* 6.5 (2008), e107.

[148] Valur Emilsson, Gudmar Thorleifsson, Bin Zhang, Amy S Leonardson, Florian Zink, Jun Zhu, et al. "Genetics of gene expression and its effect on disease". In: *Nature* 452.7186 (2008), pp. 423–8.

[149] Gudmar Thorleifsson, G Bragi Walters, Daniel F Gudbjartsson, Valgerdur Steinthorsdottir, Patrick Sulem, Anna Helgadottir, et al. "Genome-wide association yields new sequence variants at seven loci that associate with measures of obesity". In: *Nat Genet* 41.1 (2009), pp. 18–24.

[150] Sekar Kathiresan, Cristen J Willer, Gina M Peloso, Serkalem Demissie, Kiran Musunuru, Eric E Schadt, et al. "Common variants at 30 loci contribute to polygenic dyslipidemia". In: *Nat Genet* 41.1 (2009), pp. 56–65.

[151] Richa Saxena, Marie-France Hivert, Claudia Langenberg, Toshiko Tanaka, James S Pankow, Peter Vollenweider, et al. "Genetic variation in GIPR influences the glucose and insulin responses to an oral glucose challenge". In: *Nat Genet* 42.2 (2010), pp. 142–8.

[152] Soumya Raychaudhuri, Robert M Plenge, Elizabeth J Rossin, Aylwin C Y Ng, International Schizophrenia Consortium, Shaun M Purcell, et al. "Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions". In: *PLoS Genet* 5.6 (2009), e1000534.

[153] Elena G Bochukova, Ni Huang, Julia Keogh, Elana Henning, Carolin Purmann, Kasia Blaszczyk, et al. "Large, rare chromosomal deletions associated with severe early-onset obesity". In: *Nature* 463.7281 (2010), pp. 666–70.

[154] R G Walters, S Jacquemont, A Valsesia, A J de Smith, D Martinet, J Andersson, et al. "A new highly penetrant form of obesity due to deletions on chromosome 16p11.2". In: *Nature* 463.7281 (2010), pp. 671–5.

[155] Len A Pennacchio, Gabriela G Loots, Marcelo A Nobrega, and Ivan Ovcharenko. "Predicting tissue-specific enhancers in the human genome". In: *Genome Res* 17.2 (2007), pp. 201–11.

[156] Stephane Gesta, Matthias Blüher, Yuji Yamamoto, Andrew W Norris, Janin Berndt, Susan Kralisch, et al. "Evidence for a role of developmental genes in the origin of obesity and body fat distribution". In: *Proc Natl Acad Sci U S A* 103.17 (2006), pp. 6676–81.

[157] Ekkehart Lausch, Pia Hermanns, Henner F Farin, Yasemin Alanay, Sheila Unger, Sarah Nikkel, et al. "TBX15 mutations cause craniofacial dysmorphism, hypoplasia of scapula and pelvis, and short stature in Cousin syndrome". In: *Am J Hum Genet* 83.5 (2008), pp. 649–55.

[158] Sanne Kuijper, Annemiek Beverdam, Carla Kroon, Antje Brouwer, Sophie Candille, Gregory Barsh, et al. "Genetics of shoulder girdle formation: roles of Tbx15 and aristaless-like genes". In: *Development* 132.7 (2005), pp. 1601–10.

[159] Manvendra K Singh, Marianne Petry, Bénédicte Haenig, Birgit Lescher, Michael Leitges, and Andreas Kispert. "The T-box transcription factor Tbx15 is required for skeletal development". In: *Mech Dev* 122.2 (2005), pp. 131–44.

[160] James D Orth and Mark A McNiven. "Dynamin at the actin-membrane interface". In: *Curr Opin Cell Biol* 15.1 (2003), pp. 31–9.

[161] I Lisinski, A Schürmann, H G Joost, S W Cushman, and H Al-Hasani. "Targeting of GLUT6 (formerly GLUT9) and GLUT8 in rat adipose cells". In: *Biochem J* 358.Pt 2 (2001), pp. 517–22.

[162] N Inoue, R Watanabe, J Takeda, and T Kinoshita. "PIG-C, one of the three human genes involved in the first step of glycosylphosphatidylinositol biosynthesis is a homologue of Saccharomyces cerevisiae GPI2". In: *Biochem Biophys Res Commun* 226.1 (1996), pp. 193–9.

[163] R Watanabe, N Inoue, B Westfall, C H Taron, P Orlean, J Takeda, et al. "The first step of glycosylphosphatidylinositol biosynthesis is mediated by a complex of PIG-A, PIG-H, PIG-C and GPI1". In: *EMBO J* 17.4 (1998), pp. 877–85.

[164] D.D.C. Burkardt, J.A. Rosenfeld, M.L. Helgeson, B. Angle, V. Banks, W.E. Smith, et al. "Distinctive phenotype in 9 patients with deletion of chromosome 1q24-q25". In: *American Journal of Medical Genetics Part A* (2011).

[165] Michel Seve, Fabrice Chimienti, Séverine Devergnas, and Alain Favier. "In silico identification and expression of SLC30 family genes: an expressed sequence tag data mining strategy for the characterization of zinc transporters' tissue expression". In: *BMC Genomics* 5.1 (2004), p. 32.

[166] Sotoodeh Abhary, Kathryn P Burdon, Aanchal Gupta, Stewart Lake, Dinesh Selva, Nikolai Petrovsky, et al. "Common sequence variation in the VEGFA gene predicts risk of diabetic retinopathy". In: *Invest Ophthalmol Vis Sci* 50.12 (2009), pp. 5552–8.

[167] A.J. Coffey, F. Kokocinski, M.S. Calafato, C.E. Scott, P. Palta, E. Drury, et al. "The GENCODE exome: sequencing the complete human exome". In: *European Journal of Human Genetics* (2011).

[168] JL Ashurst, C.K. Chen, JGR Gilbert, K. Jekosch, S. Keenan, P. Meidl, et al. "The vertebrate genome annotation (Vega) database". In: *Nucleic acids research* 33.suppl 1 (2005), p. D459.

[169] Elizabeth A Carroll, Dianne Gerrelli, Stéphan Gasca, Elizabeth Berg, David R Beier, Andrew J Copp, et al. "Cordon-bleu is a conserved gene involved in neural tube formation". In: *Dev Biol* 262.1 (2003), pp. 16–31.

[170] Paul M Ridker, Guillaume Paré, Alex N Parker, Robert Y L Zee, Joseph P Miletich, and Daniel I Chasman. "Polymorphism in the CETP gene region, HDL cholesterol, and risk of future myocardial infarction: Genomewide analysis among 18 245 initially healthy women from the Women's Genome Health Study". In: *Circ Cardiovasc Genet* 2.1 (2009), pp. 26–33.

[171] Lowenna J Holt and Kenneth Siddle. "Grb10 and Grb14: enigmatic regulators of insulin action–and more?" In: *Biochem J* 388.Pt 2 (2005), pp. 393–406.

[172] Rafael S Depetris, Junjie Hu, Ilana Gimpelevich, Lowenna J Holt, Roger J Daly, and Stevan R Hubbard. "Structural basis for inhibition of the insulin receptor by the adaptor protein Grb14". In: *Mol Cell* 20.2 (2005), pp. 325–33.

[173] Gregory J Cooney, Ruth J Lyons, A Jayne Crew, Thomas E Jensen, Juan Carlos Molero, Christopher J Mitchell, et al. "Improved glucose homeostasis and enhanced insulin signalling in Grb14-deficient mice". In: *EMBO J* 23.3 (2004), pp. 582–93.

[174] Bertrand Cariou, Nadège Capitaine, Véronique Le Marcis, Nathalie Vega, Véronique Béréziat, Micheline Kergoat, et al. "Increased adipose tissue expression of Grb14 in several models of insulin resistance". In: *FASEB J* 18.9 (2004), pp. 965–7.

[175] Jacqueline M Vink, August B Smit, Eco J C de Geus, Patrick Sullivan, Gonneke Willemsen, Jouke-Jan Hottenga, et al. "Genome-wide association study of smoking initiation and current smoking". In: *Am J Hum Genet* 84.3 (2009), pp. 367–79.

[176] Francis J McMahon, Nirmala Akula, Thomas G Schulze, Pierandrea Muglia, Federica Tozzi, Sevilla D Detera-Wadleigh, et al. "Meta-analysis of genome-wide association data identifies a risk locus for major mood disorders on 3p21.1". In: *Nat Genet* 42.2 (2010), pp. 128–31.

[177] Laura J Scott, Pierandrea Muglia, Xiangyang Q Kong, Weihua Guan, Matthew Flickinger, Ruchi Upmanyu, et al. "Genome-wide association and meta-analysis of bipolar disorder in individuals of European ancestry". In: *Proc Natl Acad Sci U S A* 106.18 (2009), pp. 7501–6.

[178] Hiroyuki Sano, Simon C H Liu, William S Lane, John E Piletz, and Gustav E Lienhard. "Insulin receptor substrate 4 associates with the protein IRAS". In: *J Biol Chem* 277.22 (2002), pp. 19439–47.

[179] G Sesti, M Federici, M L Hribal, D Lauro, P Sbraccia, and R Lauro. "Defects of the insulin receptor substrate (IRS) system in human metabolic disorders". In: *FASEB J* 15.12 (2001), pp. 2099–111.

[180] Andrew I Su, Tim Wiltshire, Serge Batalov, Hilmar Lapp, Keith A Ching, David Block, et al. "A gene atlas of the mouse and human protein-encoding transcriptomes". In: *Proc Natl Acad Sci U S A* 101.16 (2004), pp. 6062–7.

[181] Julia Kzhyshkowska, A Gratchev, and S Goerdt. "Stabilin-1, a homeostatic scavenger receptor with multiple functions". In: *J Cell Mol Med* 10.3 (2006), pp. 635–49.

[182] Yuichi J Machida, Yuka Machida, Ajay A Vashisht, James A Wohlschlegel, and Anindya Dutta. "The deubiquitinating enzyme BAP1 regulates cell growth via interaction with HCF-1". In: *J Biol Chem* 284.49 (2009), pp. 34179–88.

[183] B L Tang. "ADAMTS: a novel family of extracellular matrix proteases". In: *Int J Biochem Cell Biol* 33.1 (2001), pp. 33–44.

[184] Ming-Chung Kan, Aparna Oruganty-Das, Amalene Cooper-Morgan, Guang Jin, Sharon A Swanger, Gary J Bassell, et al. "CPEB4 is a cell survival protein retained in the nucleus upon ischemia or endoplasmic reticulum calcium depletion". In: *Mol Cell Biol* 30.24 (2010), pp. 5658–71.

[185] Joel D Richter. "CPEB: a life in translation". In: *Trends Biochem Sci* 32.6 (2007), pp. 279–85.

[186] Masao Kimoto, Kohei Nagasawa, and Kensuke Miyake. "Role of TLR4/MD-2 and RP105/MD-1 in innate recognition of lipopolysaccharide". In: *Scand J Infect Dis* 35.9 (2003), pp. 568–72.

[187] Shih-Wei Lee, Jiu-Yao Wang, Yuan-Chun Hsieh, Ying-Jye Wu, Hsien-Wei Ting, and Lawrence Shih-Hsin Wu. "Association of single nucleotide polymorphisms of MD-1 gene with pediatric and adult asthma in the Taiwanese population". In: *J Microbiol Immunol Infect* 41.6 (2008), pp. 445–9.

[188] D.F. Conrad, D. Pinto, R. Redon, L. Feuk, O. Gokcumen, Y. Zhang, et al. "Origins and functional impact of copy number variation in the human genome". In: *Nature* 464.7289 (2009), pp. 704–712.

[189] Daisuke Watanabe, Kiyoshi Suzuma, Izumi Suzuma, Hirokazu Ohashi, Tomonari Ojima, Masafumi Kurimoto, et al. "Vitreous levels of angiopoietin 2 and vascular endothelial growth factor in patients with proliferative diabetic retinopathy". In: *Am J Ophthalmol* 139.3 (2005), pp. 476–81.

[190] Hussam Al-Kateb, Andrew P Boright, Lucia Mirea, Xinlei Xie, Rinku Sutradhar, Alireza Mowjoodi, et al. "Multiple superoxide dismutase 1/splicing factor serine alanine 15 variants are associated with the development and progression of diabetic nephropathy: the Diabetes Control and Complications Trial/Epidemiology of Diabetes Interventions and Complications Genetics study". In: *Diabetes* 57.1 (2008), pp. 218–28.

[191] Satoshi Nishimura, Ichiro Manabe, Mika Nagasaki, Yumiko Hosoya, Hiroshi Yamashita, Hideo Fujita, et al. "Adipogenesis in obesity requires close interplay between differentiating adipocytes, stromal cells, and blood vessels". In: *Diabetes* 56.6 (2007), pp. 1517–26.

[192] K P Claffey, W O Wilkison, and B M Spiegelman. "Vascular endothelial growth factor. Regulation by cell differentiation and activated second messenger pathways". In: *J Biol Chem* 267.23 (1992), pp. 16317–22.

[193] Q X Zhang, C J Magovern, C A Mack, K T Budenbender, W Ko, and T K Rosengart. "Vascular endothelial growth factor is the major angiogenic factor in omentum: mechanism of the omentum-mediated angiogenesis". In: *J Surg Res* 67.2 (1997), pp. 147–54.

[194] A Soukas, N D Socci, B D Saatkamp, S Novelli, and J M Friedman. "Distinct transcriptional profiles of adipogenesis in vivo and in vitro". In: *J Biol Chem* 276.36 (2001), pp. 34167–74.

[195] M Emoto, T Anno, Y Sato, K Tanabe, S Okuya, Y Tanizawa, et al. "Troglitazone treatment increases plasma vascular endothelial growth factor in diabetic patients and its mRNA in 3T3-L1 adipocytes". In: *Diabetes* 50.5 (2001), pp. 1166–70.

[196] J V Silha, M Krsek, P Sucharda, and L J Murphy. "Angiogenic factors are elevated in overweight and obese individuals". In: *Int J Obes (Lond)* 29.11 (2005), pp. 1308–14.

[197] N García de la Torre, J A H Wass, and H E Turner. "Antiangiogenic effects of somatostatin analogues". In: *Clin Endocrinol (Oxf)* 57.4 (2002), pp. 425–41.

[198] Ju-Suk Nam, Taryn J Turcotte, Peter F Smith, Sangdun Choi, and Jeong Kyo Yoon. "Mouse cristin/R-spondin family proteins are novel ligands for the

Frizzled 8 and LRP6 receptors and activate beta-catenin-dependent gene expression". In: *J Biol Chem* 281.19 (2006), pp. 13247–57.

[199] Olga Kazanskaya, Bisei Ohkawara, Melanie Heroult, Wei Wu, Nicole Maltry, Hellmut G Augustin, et al. "The Wnt signaling regulator R-spondin 3 promotes angioblast and vascular development". In: *Development* 135.22 (2008), pp. 3655–64.

[200] Motoko Aoki, Michihiro Mieda, Toshio Ikeda, Yoshio Hamada, Harukazu Nakamura, and Hitoshi Okamoto. "R-spondin3 is required for mouse placental development". In: *Dev Biol* 301.1 (2007), pp. 218–26.

[201] Vassiliki Theodorou, Melanie A Kimm, Mandy Boer, Lodewyk Wessels, Wendy Theelen, Jos Jonkers, et al. "MMTV insertional mutagenesis identifies genes, gene families and pathways involved in mammary cancer". In: *Nat Genet* 39.6 (2007), pp. 759–69.

[202] Anja M Duursma, Martijn Kedde, Mariette Schrier, Carlos le Sage, and Reuven Agami. "miR-148 targets human DNMT3b protein coding region". In: *RNA* 14.5 (2008), pp. 872–7.

[203] G. Chevillard and V. Blank. "NFE2L3 (NRF3): the Cinderella of the Cap 'n'Collar transcription factors". In: *Cellular and Molecular Life Sciences* (2011), pp. 1–12.

[204] Akira Futatsugi, Takeshi Nakamura, Maki K Yamada, Etsuko Ebisui, Kyoko Nakamura, Keiko Uchida, et al. "IP3 receptor types 2 and 3 mediate exocrine secretion underlying energy metabolism". In: *Science* 309.5744 (2005), pp. 2232–4.

[205] Jonathan C. K. Wells. "Sexual dimorphism of body composition". In: *Best Practice & Research Clinical Endocrinology & Metabolism* 21.3 (Sept. 2007), pp. 415–430.

[206] Angela Döring, Christian Gieger, Divya Mehta, Henning Gohlke, Holger Prokisch, Stefan Coassin, et al. "SLC2A9 influences uric acid concentrations with pronounced sex-specific effects". In: *Nat Genet* 40.4 (2008), pp. 430–6.

[207] S. Shifman, M. Johannesson, M. Bronstein, S.X. Chen, D.A. Collier, N.J. Craddock, et al. "Genome-wide association identifies a common variant in the reelin gene that increases the risk of schizophrenia only in women". In: *PLoS Genet* 4.2 (2008), e28.

[208] Benjamin F Voight, Laura J Scott, Valgerdur Steinthorsdottir, Andrew P Morris, Christian Dina, Ryan P Welch, et al. "Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis". In: *Nat Genet* 42.7 (2010), pp. 579–89.

[209] Trine Welløv Boesgaard, Anette Prior Gjesing, Niels Grarup, Jarno Rutanen, Per-Anders Jansson, Marta Letizia Hribal, et al. "Variant near ADAMTS9 known to associate with type 2 diabetes is related to insulin resistance in offspring of type 2 diabetes patients–EUGENE2 study". In: *PLoS One* 4.9 (2009), e7236.

[210] Nuria García de la Torre, Miguel A Rubio, Elena Bordiú, Lucio Cabrerizo, Eugenio Aparicio, Carmen Hernández, et al. "Effects of weight loss after bariatric

surgery for morbid obesity on vascular endothelial growth factor-A, adipocytokines, and insulin". In: *J Clin Endocrinol Metab* 93.11 (2008), pp. 4276–81.

[211] Stephane Gesta, Yu-Hua Tseng, and C Ronald Kahn. "Developmental origin of fat: tracking obesity to its source". In: *Cell* 131.2 (2007), pp. 242–56.

[212] Christian Lanctôt, Cornelius Kaspar, and Thomas Cremer. "Positioning of the mouse Hox gene clusters in the nuclei of developing embryos and differentiating embryoid bodies". In: *Exp Cell Res* 313.7 (2007), pp. 1449–59.

[213] Sophie I Candille, Catherine D Van Raamsdonk, Changyou Chen, Sanne Kuijper, Yanru Chen-Tsai, Andreas Russ, et al. "Dorsoventral patterning of the mouse coat by Tbx15". In: *PLoS Biol* 2.1 (2004), E3.

[214] Stephen O'Rahilly. "Human genetics illuminates the paths to metabolic disease". In: *Nature* 462.7271 (2009), pp. 307–14.

[215] National Institutes of Health. "Clinical Guidelines on the Identification, Evaluation, and Treatment of Overweight and Obesity in Adults–The Evidence Report. National Institutes of Health". In: *Obes Res* 6 Suppl 2 (1998), 51S–209S.

[216] Cora E Lewis, Kathleen M McTigue, Lora E Burke, Paul Poirier, Robert H Eckel, Barbara V Howard, et al. "Mortality, health outcomes, and body mass index in the overweight range: a science advisory from the American Heart Association". In: *Circulation* 119.25 (2009), pp. 3263–71.

[217] A J Stunkard, T T Foch, and Z Hrubec. "A twin study of human obesity". In: *JAMA* 256.1 (1986), pp. 51–4.

[218] H H Maes, M C Neale, and L J Eaves. "Genetic and environmental factors in relative body weight and human adiposity". In: *Behav Genet* 27.4 (1997), pp. 325–51.

[219] Amy E Taylor, Shah Ebrahim, Yoav Ben-Shlomo, Richard M Martin, Peter H Whincup, John W Yarnell, et al. "Comparison of the associations of body mass index and measures of central adiposity and fat mass with coronary heart disease, diabetes, and all-cause mortality: a study using data from 4 UK cohorts". In: *Am J Clin Nutr* 91.3 (2010), pp. 547–56.

[220] Decheng Ren, Yingjiang Zhou, David Morris, Minghua Li, Zhiqin Li, and Liangyou Rui. "Neuronal SH2B1 is essential for controlling energy and glucose homeostasis". In: *J Clin Invest* 117.2 (2007), pp. 397–406.

[221] D Huszar, C A Lynch, V Fairchild-Huntress, J H Dunmore, Q Fang, L R Berkemeier, et al. "Targeted disruption of the melanocortin-4 receptor results in obesity in mice". In: *Cell* 88.1 (1997), pp. 131–41.

[222] S O'Rahilly and I S Farooqi. "Human obesity as a heritable disorder of the central control of energy balance". In: *Int J Obes (Lond)* 32 Suppl 7 (2008), S55–61.

[223] Nancy L Heard-Costa, M Carola Zillikens, Keri L Monda, Asa Johansson, Tamara B Harris, Mao Fu, et al. "NRXN3 is a novel locus for waist circumference: a genome-wide association study from the CHARGE Consortium". In: *PLoS Genet* 5.6 (2009), e1000539.

[224] David Meyre, Jérôme Delplanque, Jean-Claude Chèvre, Cécile Lecoeur, Stéphane Lobbens, Sophie Gallina, et al. "Genome-wide association study for early-onset

and morbid adult obesity identifies three new risk loci in European populations". In: *Nat Genet* 41.2 (2009), pp. 157–9.

[225] André Scherag, Christian Dina, Anke Hinney, Vincent Vatin, Susann Scherag, Carla I G Vogel, et al. "Two new Loci for body-weight regulation identified in a joint analysis of genome-wide association studies for early-onset extreme obesity in French and german study groups". In: *PLoS Genet* 6.4 (2010), e1000916.

[226] D J Marsh, G Hollopeter, D Huszar, R Laufer, K A Yagaloff, S L Fisher, et al. "Response of melanocortin-4 receptor-deficient mice to anorectic and orexigenic peptides". In: *Nat Genet* 21.1 (1999), pp. 119–22.

[227] Thaddeus J Unger, German A Calderon, Leila C Bradley, Miguel Sena-Esteves, and Maribel Rios. "Selective deletion of Bdnf in the ventromedial and dorsomedial hypothalamus of adult mice results in hyperphagic behavior and obesity". In: *J Neurosci* 27.52 (2007), pp. 14265–74.

[228] Zhiqin Li, Yingjiang Zhou, Christin Carter-Su, Martin G Myers Jr, and Liangyou Rui. "SH2B1 enhances leptin signaling by both Janus kinase 2 Tyr813 phosphorylation-dependent and -independent mechanisms". In: *Mol Endocrinol* 21.9 (2007), pp. 2270–81.

[229] I Sadaf Farooqi, Julia M Keogh, Giles S H Yeo, Emma J Lank, Tim Cheetham, and Stephen O'Rahilly. "Clinical spectrum of obesity and mutations in the melanocortin 4 receptor gene". In: *N Engl J Med* 348.12 (2003), pp. 1085–95.

[230] Juliette Gray, Giles S H Yeo, James J Cox, Jenny Morton, Anna-Lynne R Adlam, Julia M Keogh, et al. "Hyperphagia, severe obesity, impaired cognitive function, and hyperactivity associated with functional loss of one copy of the brain-derived neurotrophic factor (BDNF) gene". In: *Diabetes* 55.12 (2006), pp. 3366–71.

[231] Ayellet V Segrè, DIAGRAM Consortium, MAGIC investigators, Leif Groop, Vamsi K Mootha, Mark J Daly, et al. "Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits". In: *PLoS Genet* 6.8 (2010).

[232] Amanda J Myers, J Raphael Gibbs, Jennifer A Webster, Kristen Rohrer, Alice Zhao, Lauren Marlowe, et al. "A survey of genetic human cortical gene expression". In: *Nat Genet* 39.12 (2007), pp. 1494–9.

[233] Joel N Hirschhorn. "Genomewide association studies–illuminating biologic pathways". In: *N Engl J Med* 360.17 (2009), pp. 1699–701.

[234] V E P P Lemmens, A Oenema, K I Klepp, H B Henriksen, and J Brug. "A systematic review of the evidence regarding efficacy of obesity prevention interventions among adults". In: *Obes Rev* 9.5 (2008), pp. 446–55.

[235] J W Anderson, E C Konz, R C Frederich, and C L Wood. "Long-term weight-loss maintenance: a meta-analysis of US studies". In: *Am J Clin Nutr* 74.5 (2001), pp. 579–84.

[236] M.N. Weedon, G. Lettre, R.M. Freathy, C.M. Lindgren, B.F. Voight, J.R.B. Perry, et al. "A common variant of HMGA2 is associated with adult and childhood height in the general population". In: *Nature genetics* 39.10 (2007), pp. 1245–1250.

[237] S. Sanna, A.U. Jackson, R. Nagaraja, C.J. Willer, W.M. Chen, L.L. Bonnycastle, et al. "Common variants in the GDF5-UQCC region are associated with variation in human height". In: *Nature Genetics* 40.2 (2008), pp. 198–203.

[238] D.F. Gudbjartsson, G.B. Walters, G. Thorleifsson, H. Stefansson, B.V. Halldorsson, P. Zusmanovich, et al. "Many sequence variants affecting diversity of adult human height". In: *Nature Genetics* 40.5 (2008), pp. 609–615.

[239] RM Malina and SB Heymsfield. "Variation in body composition associated with sex and ethnicity". In: *Human body composition. 2nd ed. Champaign, IL: Human Kinetics* (2005), pp. 271–311.

[240] M.A. McDowell and National Center for Health Statistics (US). *Anthropometric Reference Data for Children and Adults, United States, 2003-2006*. US Dept. of Health, Human Services, Centers for Disease Control, and Prevention, National Center for Health Statistics, 2008.

[241] Rachael W Taylor, Andrea M Grant, Sheila M Williams, and Ailsa Goulding. "Sex differences in regional body fat distribution from pre- to postpuberty". In: *Obesity (Silver Spring)* 18.7 (2010), pp. 1410–6.

[242] Karol Estrada, Michael Krawczak, Stefan Schreiber, Kate van Duijn, Lisette Stolk, Joyce B J van Meurs, et al. "A genome-wide association study of northwestern Europeans involves the C-type natriuretic peptide signaling pathway in the etiology of human height variation". In: *Hum Mol Genet* 18.18 (2009), pp. 3516–24.

[243] Anke Tönjes, Moritz Koriath, Dorit Schleinitz, Kerstin Dietrich, Yvonne Böttcher, Nigel W Rayner, et al. "Genetic variation in GPR133 is associated with height: genome wide association study in the self-contained population of Sorbs". In: *Hum Mol Genet* 18.23 (2009), pp. 4662–8.

[244] Yukinori Okada, Yoichiro Kamatani, Atsushi Takahashi, Koichi Matsuda, Naoya Hosono, Hiroko Ohmiya, et al. "A genome-wide association study in 19 633 Japanese subjects identified LHX3-QSOX2 and IGF1 as adult height loci". In: *Hum Mol Genet* 19.11 (2010), pp. 2303–12.

[245] Gavin J Gordon, Roderick V Jensen, Li-Li Hsiao, Steven R Gullans, Joshua E Blumenstock, William G Richards, et al. "Using gene expression ratios to predict outcome among patients with mesothelioma". In: *J Natl Cancer Inst* 95.8 (2003), pp. 598–605.

[246] Gavin J Gordon, Raphael Bueno, and David J Sugarbaker. "Genes associated with prognosis after surgery for malignant pleural mesothelioma promote tumor cell survival in vitro". In: *BMC Cancer* 11 (2011), p. 169.

[247] Elizabeth K Speliotes, Laura M Yerges-Armstrong, Jun Wu, Ruben Hernaez, Lauren J Kim, Cameron D Palmer, et al. "Genome-wide association analysis identifies variants associated with nonalcoholic fatty liver disease that have distinct effects on metabolic traits". In: *PLoS Genet* 7.3 (2011), e1001324.

[248] Anna Köttgen, Cristian Pattaro, Carsten A Böger, Christian Fuchsberger, Matthias Olden, Nicole L Glazer, et al. "New loci associated with kidney function and chronic kidney disease". In: *Nat Genet* 42.5 (2010), pp. 376–84.

[249] Yi Yu, Tushar R Bhangale, Jesen Fagerness, Stephan Ripke, Gudmar Thorleifsson, Perciliz L Tan, et al. "Common variants near FRK/COL10A1 and

VEGFA are associated with advanced age-related macular degeneration". In: *Hum Mol Genet* (2011).

[250] Mandy van Hoek, Abbas Dehghan, Jacqueline C M Witteman, Cornelia M van Duijn, André G Uitterlinden, Ben A Oostra, et al. "Predicting type 2 diabetes based on polymorphisms from genome-wide association studies: a population-based study". In: *Diabetes* 57.11 (2008), pp. 3122–8.

[251] Y Suzuki, L L Jiang, M Souri, S Miyazawa, S Fukuda, Z Zhang, et al. "D-3-hydroxyacyl-CoA dehydratase/D-3-hydroxyacyl-CoA dehydrogenase bifunctional protein deficiency: a newly identified peroxisomal disorder". In: *Am J Hum Genet* 61.5 (1997), pp. 1153–62.

[252] E G van Grunsven, E van Berkel, L Ijlst, P Vreken, J B de Klerk, J Adamski, et al. "Peroxisomal D-hydroxyacyl-CoA dehydrogenase deficiency: resolution of the enzyme defect and its molecular basis in bifunctional protein deficiency". In: *Proc Natl Acad Sci U S A* 95.5 (1998), pp. 2128–33.

[253] Lawrence True, Ilsa Coleman, Sarah Hawley, Ching-Ying Huang, David Gifford, Roger Coleman, et al. "A molecular correlate to the Gleason grading system for prostate adenocarcinoma". In: *Proc Natl Acad Sci U S A* 103.29 (2006), pp. 10991–6.

[254] Krishan K Rasiah, Margaret Gardiner-Garden, Emma J D Padilla, Gabriele Möller, James G Kench, M Chehani Alles, et al. "HSD17B4 overexpression, an independent biomarker of poor patient outcome in prostate cancer". In: *Mol Cell Endocrinol* 301.1-2 (2009), pp. 89–96.

[255] Dalila Pinto, Christian Marshall, Lars Feuk, and Stephen W Scherer. "Copy-number variation in control population cohorts". In: *Hum Mol Genet* 16 Spec No. 2 (2007), R168–73.

[256] Jeffrey M Kidd, Gregory M Cooper, William F Donahue, Hillary S Hayden, Nick Sampas, Tina Graves, et al. "Mapping and sequencing of structural variation from eight human genomes". In: *Nature* 453.7191 (2008), pp. 56–64.

[257] Eleftheria Zeggini, Michael N Weedon, Cecilia M Lindgren, Timothy M Frayling, Katherine S Elliott, Hana Lango, et al. "Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes". In: *Science* 316.5829 (2007), pp. 1336–41.

[258] Laura J Scott, Karen L Mohlke, Lori L Bonnycastle, Cristen J Willer, Yun Li, William L Duren, et al. "A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants". In: *Science* 316.5829 (2007), pp. 1341–5.

[259] and Saxena Richa, Voight Benjamin F, Lyssenko Valeriya, Burtt Noël P, de Bakker Paul I W, Chen Hong, et al. "Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels". In: *Science* 316.5829 (2007), pp. 1331–6.

[260] M Ristow, D Müller-Wieland, A Pfeiffer, W Krone, and C R Kahn. "Obesity associated with a mutation in a genetic regulator of adipocyte differentiation". In: *N Engl J Med* 339.14 (1998), pp. 953–9.

[261] Y Barak, M C Nelson, E S Ong, Y Z Jones, P Ruiz-Lozano, K R Chien, et al. “PPAR gamma is required for placental, cardiac, and adipose tissue development”. In: *Mol Cell* 4.4 (1999), pp. 585–95.

[262] E D Rosen, P Sarraf, A E Troy, G Bradwin, K Moore, D S Milstone, et al. “PPAR gamma is required for the differentiation of adipose tissue in vivo and in vitro”. In: *Mol Cell* 4.4 (1999), pp. 611–7.

[263] Tuomo Rankinen, Aamir Zuberi, Yvon C Chagnon, S John Weisnagel, George Argyropoulos, Brandon Walts, et al. “The human obesity gene map: the 2005 update”. In: *Obesity (Silver Spring)* 14.4 (2006), pp. 529–644.

[264] G J Murphy and J C Holder. “PPAR-gamma agonists: therapeutic role in diabetes, inflammation and cancer”. In: *Trends Pharmacol Sci* 21.12 (2000), pp. 469–74.

[265] I Barroso, M Gurnell, V E Crowley, M Agostini, J W Schwabe, M A Soos, et al. “Dominant negative mutations in human PPARgamma associated with severe insulin resistance, diabetes mellitus and hypertension”. In: *Nature* 402.6764 (1999), pp. 880–3.

[266] Andrea Galgani, AnaMaria Valdes, Henry A Erlich, Calvin Mano, Suzanne Cheng, Antonio Petrone, et al. “Homozygosity for the Ala allele of the PPARÎş2 Pro12Ala polymorphism is associated with reduced risk of coronary artery disease”. In: *Dis Markers* 29.5 (2010), pp. 259–64.

[267] S Kitamura, Y Miyazaki, Y Shinomura, S Kondo, S Kanayama, and Y Matsuzawa. “Peroxisome proliferator-activated receptor gamma induces growth arrest and differentiation markers of human colon cancer cells”. In: *Jpn J Cancer Res* 90.1 (1999), pp. 75–80.

[268] P Sarraf, E Mueller, W M Smith, H M Wright, J B Kum, L A Aaltonen, et al. “Loss-of-function mutations in PPAR gamma associated with human colon cancer”. In: *Mol Cell* 3.6 (1999), pp. 799–804.

[269] X Xin, S Yang, J Kowalski, and M E Gerritsen. “Peroxisome proliferator-activated receptor gamma ligands are potent inhibitors of angiogenesis in vitro and in vivo”. In: *J Biol Chem* 274.13 (1999), pp. 9116–21.

[270] Eleonora Morini, Vittorio Tassi, Daria Capponi, Ornella Ludovico, Bruno Dallapiccola, Vincenzo Trischitta, et al. “Interaction between PPARgamma2 variants and gender on the modulation of body weight”. In: *Obesity (Silver Spring)* 16.6 (2008), pp. 1467–70.

[271] Jaime A Duffield, Tony Vuocolo, Ross Tellam, Jim R McFarlane, Kate G Kauter, Beverly S Muhlhausler, et al. “Intrauterine growth restriction and the sex specific programming of leptin and peroxisome proliferator-activated receptor gamma (PPARgamma) mRNA expression in visceral fat in the lamb”. In: *Pediatr Res* 66.1 (2009), pp. 59–65.

[272] Chun-Hsin Chen, Mong-Liang Lu, Po-Hsiu Kuo, Po-Yu Chen, Chih-Chiang Chiu, Chung-Feng Kao, et al. “Gender differences in the effects of peroxisome proliferator-activated receptor Îş2 gene polymorphisms on metabolic adversity in patients with schizophrenia or schizoaffective disorder”. In: *Prog Neuropsychopharmacol Biol Psychiatry* 35.1 (2011), pp. 239–45.

[273] Heribert Schunkert, Inke R König, Sekar Kathiresan, Muredach P Reilly, Themistocles L Assimes, Hilma Holm, et al. "Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease". In: *Nat Genet* 43.4 (2011), pp. 333–8.

[274] S Girirajan, L J Elsas 2nd, K Devriendt, and S H Elsea. "RAI1 variations in Smith-Magenis syndrome patients without 17p11.2 deletions". In: *J Med Genet* 42.11 (2005), pp. 820–8.

[275] Santhosh Girirajan, Christopher N Vlangos, Barbara B Szomju, Emily Edelman, Christopher D Trevors, Lucie Dupuis, et al. "Genotype-phenotype correlation in Smith-Magenis syndrome: evidence that multiple genes in 17p11.2 contribute to the clinical spectrum". In: *Genet Med* 8.7 (2006), pp. 417–27.

[276] S Hani Najafi-Shoushtari, Fjoralba Kristo, Yingxia Li, Toshi Shioda, David E Cohen, Robert E Gerszten, et al. "MicroRNA-33 and the SREBP host genes cooperate to control cholesterol homeostasis". In: *Science* 328.5985 (2010), pp. 1566–9.

[277] Dawn M Waterworth, Sally L Ricketts, Kijoung Song, Li Chen, Jing Hua Zhao, Samuli Ripatti, et al. "Genetic variants influencing circulating lipid levels and risk of coronary artery disease". In: *Arterioscler Thromb Vasc Biol* 30.11 (2010), pp. 2264–76.

[278] Alexander Pearlman, Johnny Loke, Cedric Le Caignec, Stefan White, Lisa Chin, Andrew Friedman, et al. "Mutations in MAP3K1 cause 46,XY disorders of sex development and implicate a common signal transduction pathway in human testis determination". In: *Am J Hum Genet* 87.6 (2010), pp. 898–904.

[279] Douglas F Easton, Karen A Pooley, Alison M Dunning, Paul D P Pharoah, Deborah Thompson, Dennis G Ballinger, et al. "Genome-wide association study identifies novel breast cancer susceptibility loci". In: *Nature* 447.7148 (2007), pp. 1087–93.

[280] Pei-Hua Lu, Jie Yang, Chen Li, Mu-Xin Wei, Wei Shen, Li-ping Shi, et al. "Association between mitogen-activated protein kinase kinase kinase 1 rs889312 polymorphism and breast cancer risk: evidence from 59,977 subjects". In: *Breast Cancer Res Treat* 126.3 (2011), pp. 663–70.

[281] Antonis C Antoniou, Amanda B Spurdle, Olga M Sinilnikova, Sue Healey, Karen A Pooley, Rita K Schmutzler, et al. "Common breast cancer-predisposition alleles are associated with breast cancer risk in BRCA1 and BRCA2 mutation carriers". In: *Am J Hum Genet* 82.4 (2008), pp. 937–48.

[282] Paola Sebastiani, Nadia Solovieff, Annibale Puca, Stephen W Hartley, Efthymia Melista, Stacy Andersen, et al. "Genetic signatures of exceptional longevity in humans". In: *Science* 2010 (2010).

[283] Eli A Stahl, Soumya Raychaudhuri, Elaine F Remmers, Gang Xie, Stephen Eyre, Brian P Thomson, et al. "Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci". In: *Nat Genet* 42.6 (2010), pp. 508–14.

[284] Alexandra Zhernakova, Eli A Stahl, Gosia Trynka, Soumya Raychaudhuri, Eleanora A Festen, Lude Franke, et al. "Meta-analysis of genome-wide asso-

ciation studies in celiac disease and rheumatoid arthritis identifies fourteen non-HLA shared loci". In: *PLoS Genet* 7.2 (2011), e1002004.

[285] Andrea Poyastro Pinheiro, Cynthia M Bulik, Laura M Thornton, Patrick F Sullivan, Tammy L Root, Cinnamon S Bloss, et al. "Association study of 182 candidate genes in anorexia nervosa". In: *Am J Med Genet B Neuropsychiatr Genet* 153B.5 (2010), pp. 1070–80.

[286] Wishal D Ramdas, Leonieke M E van Koolwijk, M Kamran Ikram, Nomdo M Jansonius, Paulus T V M de Jong, Arthur A B Bergen, et al. "A genome-wide association study of optic disc parameters". In: *PLoS Genet* 6.6 (2010), e1000978.

[287] Stuart Macgregor, Alex W Hewitt, Pirro G Hysi, Jonathan B Ruddle, Sarah E Medland, Anjali K Henders, et al. "Genome-wide association identifies ATOH7 as a major gene determining human optic disc size". In: *Hum Mol Genet* 19.13 (2010), pp. 2716–24.

[288] Xiaohui Zhang, Shiqiang Li, Xueshan Xiao, Xiaoyun Jia, Panfeng Wang, Huangxuan Shen, et al. "Mutational screening of 10 genes in Chinese patients with microphthalmia and/or coloboma". In: *Mol Vis* 15 (2009), pp. 2911–8.

[289] M E Gallardo, S Rodríguez De Córdoba, A S Schneider, M A Dwyer, C Ayuso, and P Bovolenta. "Analysis of the developmental SIX6 homeobox gene in patients with anophthalmia/microphthalmia". In: *Am J Med Genet A* 129A.1 (2004), pp. 92–4.

[290] Xue Li, Valentina Perissi, Forrest Liu, David W Rose, and Michael G Rosenfeld. "Tissue-specific regulation of retinal and pituitary precursor cell proliferation". In: *Science* 297.5584 (2002), pp. 1180–3.

[291] Xia Lin, Xueyan Duan, Yao-Yun Liang, Ying Su, Katharine H Wrighton, Jianyin Long, et al. "PPM1A functions as a Smad phosphatase to terminate TGFbeta signaling". In: *Cell* 125.5 (2006), pp. 915–28.

[292] K I Kim, S H Baek, Y J Jeon, S Nishimori, T Suzuki, S Uchida, et al. "A new SUMO-1-specific protease, SUSP1, that is highly expressed in reproductive organs". In: *J Biol Chem* 275.19 (2000), pp. 14102–6.

[293] Kirsten M Sanggaard, Klaus W Kjaer, Hans Eiberg, Gudrun Nürnberg, Peter Nürnberg, Katrin Hoffman, et al. "A novel nonsense mutation in MYO6 is associated with progressive nonsyndromic hearing loss in a Danish DFNA22 family". In: *Am J Med Genet A* 146A.8 (2008), pp. 1017–25.

[294] Nele Hilgert, Vedat Topsakal, Joost van Dinther, Erwin Offeciers, Paul Van de Heyning, and Guy Van Camp. "A splice-site mutation and overexpression of MYO6 cause a similar phenotype in two families with autosomal dominant hearing loss". In: *Eur J Hum Genet* 16.5 (2008), pp. 593–602.

[295] Jian Gu, Yuanqing Ye, Margaret R Spitz, Jie Lin, Lambertus A Kiemeney, Jingliang Xing, et al. "A genetic variant near the PMAIP1/Noxa gene is associated with increased bleomycin sensitivity". In: *Hum Mol Genet* 20.4 (2011), pp. 820–6.

[296] Noel J Buckley, Rory Johnson, Chiara Zuccato, Angela Bithell, and Elena Cattaneo. "The role of REST in transcriptional and epigenetic dysregulation in Huntington's disease". In: *Neurobiol Dis* 39.1 (2010), pp. 28–39.

[297] Hui Lv, Guoqing Pan, Guopei Zheng, Xiaoying Wu, Hongzheng Ren, Ying Liu, et al. "Expression and functions of the repressor element 1 (RE-1)-silencing transcription factor (REST) in breast cancer". In: *J Cell Biochem* 110.4 (2010), pp. 968–74.

[298] Anna C Need, Deborah K Attix, Jill M McEvoy, Elizabeth T Cirulli, Kristen L Linney, Priscilla Hunt, et al. "A genome-wide study of common SNPs and CNVs in cognitive performance in the CANTAB". In: *Hum Mol Genet* 18.23 (2009), pp. 4650–61.

[299] Christopher Newton-Cheh, Toby Johnson, Vesela Gateva, Martin D Tobin, Murielle Bochud, Lachlan Coin, et al. "Genome-wide association study identifies eight loci associated with blood pressure". In: *Nat Genet* 41.6 (2009), pp. 666–76.

[300] Chen Liu, Huaixing Li, Qibin Qi, Ling Lu, Wei Gan, Ruth Jf Loos, et al. "Common variants in or near FGF5, CYP17A1 and MTHFR genes are associated with blood pressure and hypertension in Chinese Hans". In: *J Hypertens* 29.1 (2011), pp. 70–5.

[301] Ellen L Goode, Georgia Chenevix-Trench, Honglin Song, Susan J Ramus, Maria Notaridou, Kate Lawrenson, et al. "A genome-wide association study identifies susceptibility loci for ovarian cancer at 2q31 and 8q24". In: *Nat Genet* 42.10 (2010), pp. 874–9.

[302] Demetris Pillas, Clive J Hoggart, David M Evans, Paul F O'Reilly, Kirsi Sipilä, Raija Lähdesmäki, et al. "Genome-wide association study reveals multiple loci associated with primary tooth development during infancy". In: *PLoS Genet* 6.2 (2010), e1000856.

[303] Christopher A Haiman, Gary K Chen, William J Blot, Sara S Strom, Sonja I Berndt, Rick A Kittles, et al. "Genome-wide association study of prostate cancer in men of African ancestry identifies a susceptibility locus at 17q21". In: *Nat Genet* 43.6 (2011), pp. 570–3.

[304] Raman Kumar, Jantina Manning, Hayley E Spendlove, Gabriel Kremmidiotis, Ross McKirdy, Jaclyn Lee, et al. "ZNF652, a novel zinc finger protein, interacts with the putative breast tumor suppressor CBFA2T3 to repress transcription". In: *Mol Cancer Res* 4.9 (2006), pp. 655–65.

[305] Damien C Croteau-Chonka, Amanda F Marvelle, Ethan M Lange, Nanette R Lee, Linda S Adair, Leslie A Lange, et al. "Genome-wide association study of anthropometric traits and evidence of interactions with age and study year in Filipino women". In: *Obesity (Silver Spring)* 19.5 (2011), pp. 1019–27.

[306] Aurora Esquela-Kerscher and Frank J Slack. "Oncomirs - microRNAs with a role in cancer". In: *Nat Rev Cancer* 6.4 (2006), pp. 259–69.

[307] Jason M Shohet, Rajib Ghosh, Cristian Coarfa, Andrew Ludwig, Ashley L Benham, Zaowen Chen, et al. "A genome-wide search for promoters that respond to increased MYCN reveals both new oncogenic and tumor suppressor microRNAs associated with aggressive neuroblastoma". In: *Cancer Res* 71.11 (2011), pp. 3841–51.

[308] Laura Gramantieri, Francesca Fornari, Elisa Callegari, Silvia Sabbioni, Giovanni Lanza, Carlo M Croce, et al. "MicroRNA involvement in hepatocellular carcinoma". In: *J Cell Mol Med* 12.6A (2008), pp. 2189–204.

[309] Niamh Lynam-Lennon, Stephen G Maher, and John V Reynolds. "The roles of microRNA in cancer and apoptosis". In: *Biol Rev Camb Philos Soc* 84.1 (2009), pp. 55–71.

[310] Jie Xiang and Ji Wu. "Feud or Friend? The Role of the miR-17-92 Cluster in Tumorigenesis". In: *Curr Genomics* 11.2 (2010), pp. 129–35.

[311] Jennifer L Reichek, Fenghai Duan, Lynette M Smith, Donna M Gustafson, Roddy S O'Connor, Chune Zhang, et al. "Genomic and clinical analysis of amplification of the 13q31 chromosomal region in alveolar rhabdomyosarcoma: a report from the Children's Oncology Group". In: *Clin Cancer Res* 17.6 (2011), pp. 1463–73.

[312] Virginie Olive, Margaux J Bennett, James C Walker, Cong Ma, Iris Jiang, Carlos Cordon-Cardo, et al. "miR-19 is a key oncogenic component of mir-17-92". In: *Genes Dev* 23.24 (2009), pp. 2839–49.

[313] Wan-Hsin Liu, Shiou-Hwei Yeh, Cho-Chun Lu, Sung-Liang Yu, Hsuan-Yu Chen, Chien-Yu Lin, et al. "MicroRNA-18a prevents estrogen receptor-alpha expression, promoting proliferation of hepatocellular carcinoma cells". In: *Gastroenterology* 136.2 (2009), pp. 683–93.

[314] Koji Okamoto, Katsushi Tokunaga, Kent Doi, Toshiro Fujita, Hodaka Suzuki, Tetsuo Katoh, et al. "Common variation in GPC5 is associated with acquired nephrotic syndrome". In: *Nat Genet* 43.5 (2011), pp. 459–63.

[315] Manuel Comabella, David W Craig, Montse Camiña-Tato, Carlos Morcillo, Cristina Lopez, Arcadi Navarro, et al. "Identification of a novel risk locus for multiple sclerosis at 13q31.3 by a pooled genome-wide scan of 500,000 single nucleotide polymorphisms". In: *PLoS One* 3.10 (2008), e3490.

[316] Sergio E Baranzini, Joanne Wang, Rachel A Gibson, Nicholas Galwey, Yvonne Naegelin, Frederik Barkhof, et al. "Genome-wide association analysis of susceptibility and clinical phenotype in multiple sclerosis". In: *Hum Mol Genet* 18.4 (2009), pp. 767–78.

[317] Yafei Li, Chau-Chyun Sheu, Yuanqing Ye, Mariza de Andrade, Liang Wang, Shen-Chih Chang, et al. "Genetic variants and risk of lung cancer in never smokers: a genome-wide association study". In: *Lancet Oncol* 11.4 (2010), pp. 321–30.

[318] A Itoh, T Miyabayashi, M Ohno, and S Sakano. "Cloning and expressions of three mammalian homologues of Drosophila slit suggest possible roles for Slit in the formation and maintenance of the nervous system". In: *Brain Res Mol Brain Res* 62.2 (1998), pp. 175–86.

[319] YongYong Shi, XinZhi Zhao, Lan Yu, Ran Tao, JunXia Tang, YuJuan La, et al. "Genetic structure adds power to detect schizophrenia susceptibility at SLIT3 in the Chinese Han population". In: *Genome Res* 14.7 (2004), pp. 1345–9.

[320] Nicole Soranzo, Tim D Spector, Massimo Mangino, Brigitte Kühnel, Augusto Rendon, Alexander Teumer, et al. "A genome-wide meta-analysis identifies 22

loci associated with eight hematological parameters in the HaemGen consortium". In: *Nat Genet* 41.11 (2009), pp. 1182–90.

[321] International Parkinson Disease Genomics Consortium, Michael A Nalls, Vincent Plagnol, Dena G Hernandez, Manu Sharma, Una-Marie Sheerin, et al. "Imputation of sequence variants for identification of genetic risks for Parkinson's disease: a meta-analysis of genome-wide association studies". In: *Lancet* 377.9766 (2011), pp. 641–9.

[322] Lindsay M Morton, Mark P Purdue, Tongzhang Zheng, Sophia S Wang, Bruce Armstrong, Yawei Zhang, et al. "Risk of non-Hodgkin lymphoma associated with germline variation in genes that regulate the cell cycle, apoptosis, and lymphocyte development". In: *Cancer Epidemiol Biomarkers Prev* 18.4 (2009), pp. 1259–70.

[323] Dibyendu Bhattacharyya and Benjamin S Glick. "Two mammalian Sec16 homologues have nonredundant functions in endoplasmic reticulum (ER) export and transitional ER organization". In: *Mol Biol Cell* 18.3 (2007), pp. 839–49.

[324] Masayoshi Yamaguchi. "Novel protein RGPR-p117: its role as the regucalcin gene transcription factor". In: *Mol Cell Biochem* 327.1-2 (2009), pp. 53–63.

[325] S Noto, T Maeda, S Hattori, J Inazawa, M Imamura, M Asaka, et al. "A novel human RasGAP-like gene that maps within the prostate cancer susceptibility locus at chromosome 1q25". In: *FEBS Lett* 441.1 (1998), pp. 127–31.

[326] Yongtao Guan and Matthew Stephens. "Practical issues in imputation-based association mapping". In: *PLoS Genet* 4.12 (2008), e1000279.

[327] Yurii S Aulchenko, Maksim V Struchalin, and Cornelia M van Duijn. "ProbABEL package for genome-wide association analysis of imputed data". In: *BMC Bioinformatics* 11 (2010), p. 134.

[328] Yurii S Aulchenko, Stephan Ripke, Aaron Isaacs, and Cornelia M van Duijn. "GenABEL: an R library for genome-wide association analysis". In: *Bioinformatics* 23.10 (2007), pp. 1294–6.

[329] Gonçalo R Abecasis and Janis E Wigginton. "Handling marker-marker linkage disequilibrium: pedigree analysis with clustered markers". In: *Am J Hum Genet* 77.5 (2005), pp. 754–67.

[330] Tanya M Teslovich, Kiran Musunuru, Albert V Smith, Andrew C Edmondson, Ioannis M Stylianou, Masahiro Koseki, et al. "Biological, clinical and population relevance of 95 loci for blood lipids". In: *Nature* 466.7307 (2010), pp. 707–13.

[331] International Multiple Sclerosis Genetics Consortium, Wellcome Trust Case Control Consortium 2, Stephen Sawcer, Garrett Hellenthal, Matti Pirinen, Chris C A Spencer, et al. "Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis". In: *Nature* 476.7359 (2011), pp. 214–9.

[332] The Schizophrenia Psychiatric Genome-Wide Association Study (GWAS) Consortium, Stephan Ripke, Alan R Sanders, Kenneth S Kendler, Douglas F Levinson, Pamela Sklar, et al. "Genome-wide association study identifies five new schizophrenia loci". In: *Nat Genet* (2011).

[333] Psychiatric GWAS Consortium Bipolar Disorder Working Group, Pamela Sklar, Stephan Ripke, Laura J Scott, Ole A Andreassen, Sven Cichon, et al. "Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4". In: *Nat Genet* (2011).

[334] A. Strange, F. Capon, C.C.A. Spencer, J. Knight, M.E. Weale, M.H. Allen, et al. "A genome-wide association study identifies new psoriasis susceptibility loci and an interaction between HLA-C and ERAP1". In: *Nature genetics* 42.11 (2010), pp. 985–990.

[335] P.E. Stuart, R.P. Nair, E. Ellinghaus, J. Ding, T. Tejasvi, J.E. Gudjonsson, et al. "Genome-wide association analysis identifies three psoriasis susceptibility loci". In: *Nature genetics* (2010).

[336] A. Franke, D.P.B. McGovern, J.C. Barrett, K. Wang, G.L. Radford-Smith, T. Ahmad, et al. "Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci". In: *Nature genetics* (2010).

[337] Reedik Mägi and Andrew P Morris. "GWAMA: software for genome-wide association meta-analysis". In: *BMC Bioinformatics* 11 (2010), p. 288.

[338] Gilean A T McVean, Simon R Myers, Sarah Hunt, Panos Deloukas, David R Bentley, and Peter Donnelly. "The fine-scale structure of recombination rate variation in the human genome". In: *Science* 304.5670 (2004), pp. 581–4.

[339] C.C.A. Spencer, Z. Su, P. Donnelly, and J. Marchini. "Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip". In: *PLoS genetics* 5.5 (2009), e1000477.

[340] BP Fuller, MJ Kahn, PA Barr, L. Biesecker, E. Crowley, J. Garber, et al. "Privacy in genetics research." In: *Science (New York, NY)* 285.5432 (1999), p. 1359.

[341] G.J. Annas. *The limits of state laws to protect genetic information*. 2001.

[342] D. Mascalzoni, A. Hicks, P. Pramstaller, and M. Wjst. "Informed consent in the genomics era". In: *PLoS Med* 5.9 (2008), e192.

[343] I.S. Kohane, K.D. Mandl, P.L. Taylor, I.A. Holm, D.J. Nigrin, and L.M. Kunkel. "Medicine: Reestablishing the researcher-patient compact". In: *Science* 316.5826 (2007), p. 836.

[344] N. Homer, S. Szelinger, M. Redman, D. Duggan, W. Tembe, J. Muehling, et al. "Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays". In: *PLoS Genetics* 4.8 (2008).

[345] Bishwaranjan Bhattacharjee, Naoki Abe, Kenneth Goldman, Bianca Zadrozny, R. Chillakuru Vamsavardhana, Marysabel del Carpio, et al. "Using secure coprocessors for privacy preserving collaborative data mining and analysis". In: *DaMoN '06: Proceedings of the 2nd international workshop on Data management on new hardware*. Chicago, Illinois: ACM, 2006, p. 1. ISBN: 1-59593-466-9.

[346] X. Chen, T. Garfinkel, E.C. Lewis, P. Subrahmanyam, C.A. Waldspurger, D. Boneh, et al. "Overshadow: a virtualization-based approach to retrofitting protection in commodity operating systems". In: *Proceedings of the 13th international conference on Architectural support for programming languages and operating systems*. ACM New York, NY, USA. 2008, pp. 2–13.

[347] R. Ta-Min, L. Litty, and D. Lie. *Splitting Interfaces: Making Trust Between Applications and Operating Systems Configurable.* University of Toronto, 2006.

[348] Tal Garfinkel, Ben Pfaff, Jim Chow, Mendel Rosenblum, and Dan Boneh. "Terra: A Virtual Machine-Based Platform for Trusted Computing". In: *Proceedings of the 19th Symposium on Operating System Principles(SOSP 2003)*. 2003.

[349] National Institute of Standards and Technology, U.S. Department of Commerce, Ronald H. Brown Secretary. "SECURITY REQUIREMENTS FOR CRYPTOGRAPHIC MODULES". In: *FEDERAL INFORMATION PROCESSING STANDARDS PUBLICATION* 140.1 (1994).

[350] International Business Machines Corp. *IBM PCIe Cryptographic Coprocessor*. URL: `http://www.ibm.com/security/cryptocards/pciecc/overview.shtml`.

[351] International Business Machines Corp. *IBM United States Hardware Announcement 110-158*. 2010. URL: `http://www-01.ibm.com/common/ssi/rep_ca/8/897/ENUS110-158/ENUS110-158.PDF`.