Dissecting Heterogeneity in GWAS Meta-analysis



Lerato Elaine Magosi

Keble College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

© Lerato Elaine Magosi, Oxford, England, 2017

Contents

List of Abbreviations vi

List of Figures ix

List of Tables xiv

Chapter 1 1

1.1 A brief historical note on meta-analysis 1

1.2 An introduction to exploring heterogeneity in GWAS meta-analysis 4

1.2.1 Importance of dissecting heterogeneity in GWAS meta-analysis 4

1.2.2 Typical sources of heterogeneity in GWAS meta-analysis 8

1.2.3 Popular approaches for conducting meta-analysis 21

1.2.4 Common heterogeneity measures in GWAS meta-analysis 28

1.3 Summary 47

1.4 General aims and outline 48

Chapter 2 51

2.1 Hypothesis / Research needs 51

2.2 Methods 55

2.2.1 Identifying systematic heterogeneity patterns in GWAS meta-analysis 55

2.2.2 Empirically evaluating the *M* statistic's null distribution 62

2.2.3 A comparative analytical power analysis of *M* and the Q-statistic to identify outlier studies showing systematic heterogeneity in GWAS meta-analysis 63

2.2.4 Assessing *M*'s power to identify outlier studies showing systematic heterogeneity patterns over a background of substantial marker-specific (random) heterogeneity 65

2.2.5 Explaining heterogeneity in the CARDIoGRAMplusC4D meta-analysis 66
2.3 Results 67
2.3.1 Aim 1: Simulation studies to examine the properties of the *M* statistic 67
2.3.2 Aim 2: Identifying systematic heterogeneity patterns in a GWAS meta-analysis of coronary artery disease 82
2.4 Discussion 119
2.4.1 Rationale for investigating systematic heterogeneity in GWAS meta-analysis 119
2.4.2 Assumptions and properties of the *M* statistic 120
2.4.3 Applying the *M* statistic to a meta-analysis of CAD risk 124
2.4.4 Advantages and limitations of the *M* approach 128
2.4.5 Alternative approaches 132
2.4.6 Overlapping approaches 133
2.5 Chapter summary 136
Chapter 3 138
3.1 Hypothesis/ Research needs 138
3.2 Methods 144
3.2.1 Bayesian statistical inference 144
3.2.2 A Bayesian random-effects model for genetic association meta-analysis 151
3.2.3 Constructing a GWAS derived empirically informed heterogeneity prior 154
3.2.4 Performing Bayesian and frequentist random-effects meta-analyses 159
3.2.5 Calculating credible intervals for Bayesian average genetic effect estimates 159
3.2.6 Computing coverage probabilities of average genetic effect estimates produced by Bayesian and classical random-effects meta-analyses 160
3.2.7 Calculating the extent of bias in Bayesian and classical heterogeneity estimates 160

3.2.8 Calculating mean squared errors associated with Bayesian and classical heterogeneity estimates 161
3.3 Results 162
3.3.1 Aim 1: Simulation studies to compare the precision of Bayesian and frequentist estimates for average genetic effects and heterogeneity 162
3.3.2 Aim 2: Comparing the precision of Bayesian and frequentist average genetic effect estimates in a meta-analysis of coronary artery disease 177
3.4 Discussion 187
3.4.1 Rationale for improving average genetic effect estimates in smaller genetic association meta-analyses 187
3.4.2 Advantages and limitations of employing GWAS derived empirical heterogeneity priors. 189
3.5 Chapter summary 191
Chapter 4 193
4.1 Hypothesis/ Research needs 193
4.1.1 Contemporary fixed and random-effects meta-analysis methods 199
4.1.2 Potential heterogeneity sources for inflated genetic effects at small effect loci 203
4.1.3 Performance of traditional and contemporary fixed and random-effects methods 205
4.2 Methods 208
4.2.1 A power comparison of the Han-Eskin random-effects model and conventional fixed and random-effects methods. 208
4.2.2 Identifying small effect heterogeneous loci overlooked by traditional fixed and random-effects methods, in the CARDIoGRAMplusC4D GWAS meta-analysis. 208
4.3 Results 209

4.3.1. Aim 1: Simulation studies to assess the power of the Han-Eskin random-effects model in relation to traditional fixed and random-effects meta-analysis methods. 209
4.3.2 Aim 2: A re-appraisal of the CARDIoGRAMplusC4D meta-analysis of coronary artery disease, to identify small effect heterogeneous risk loci overlooked by traditional fixed and random-effects methods. 221
4.4 Discussion 229
4.4.1 Improving power to detect small effect genetic associations amidst moderate-to-high heterogeneity 229
4.4.2 Advantages, limitations and alternative methods 234
4.5 Chapter summary 235
Chapter 5 237
5.1 General Discussion 237
5.1.1 Identifying systematic heterogeneity patterns in genetic association meta-analyses .. 239
5.1.2 Improving inference on average genetic effects and heterogeneity in smaller genetic association meta-analyses 243
5.1.3 Identifying small effect heterogeneous loci overlooked by traditional fixed and random-effects methods 245
5.1.4 Future directions of research 249
5.2 Final remarks 252
References 253
Contributions of Collaborators 265
Appendix 266

Abstract

DISSECTING HETEROGENEITY IN GWAS META-ANALYSIS

Lerato Elaine Magosi, Keble College

DPhil Thesis, Michaelmas Term, 2017

Statistical heterogeneity refers to differences among results of studies combined in a meta-analysis beyond that expected by chance. On the one hand, excessive heterogeneity can diminish power to discover genetic signals; on the other, moderate heterogeneity can reveal important biological differences among studies. Given its double-edged nature, this thesis dissects heterogeneity in genetic association meta-analyses from three vantage points. First, a novel multi-variant statistic, *M* is proposed to detect genome-wide (systematic) heterogeneity patterns in genetic association meta-analyses. This was motivated by the limited availability of appropriate methodology to measure the impact of heterogeneity across genetic signals, since traditional metrics (Q, I^2 and τ^2) measure heterogeneity at individual variants. Second, given that meta-analyses comprising small numbers of studies typically report imprecise summary effect estimates; GWAS-derived empirical heterogeneity priors are used to improve precision in estimation of average genetic effects and heterogeneity in smaller meta-analyses (e.g. ≤ 10 studies). Third, a critical evaluation of the Han-Eskin random-effects model shows how it can identify small effect heterogeneous loci overlooked by traditional fixed and random-effects methods.

This work draws attention to the existence of genome-wide heterogeneity patterns, to reveal systematic differences among the ascertainment criteria of participating studies in a meta-analysis

of coronary disease (CAD) risk. Furthermore, simulation studies with the Han-Eskin random-effects model revealed inflated genetic signals at small effect loci when heterogeneity levels were high. However, it did reveal an additional CAD risk variant overlooked by traditional meta-analysis methods.

We therefore recommend a holistic approach to exploring heterogeneity in meta-analyses which assesses heterogeneity of genetic effects both at individual variants with traditional statistics and across multiple genetic signals with the *M* statistic. Furthermore, it is critically important to review forest plots for small effect loci identified using the Han-Eskin random-effects model amidst moderate-to-high heterogeneity ($I^2 \geq 40\%$).

Acknowledgements

I would like to thank my supervisors Professor Martin Farrall and Dr. Jemma C. Hopewell for their guidance throughout this project. Furthermore, I would like to thank my family, friends and fellow lab members for their support.

Contributions

Through the work done in this thesis my specific contributions to the field comprise:

1. Devising a novel multi-variant statistic, *M* to search for genome-wide heterogeneity patterns in genetic association meta-analyses. The *M* statistic has been made available through the following paper:

   Magosi, L.E., Goel, A., Hopewell, J.C., Farrall, M., and CARDIoGRAMplusC4D Consortium (2017). Identifying systematic heterogeneity patterns in genetic association meta-analysis studies. PLoS Genetics *13*, e1006755.

2. Developing the getmstatistic R® package and Stata® command to avail the *M* statistic to a wider audience.

   R package:

   getmstatistic: Quantifying Systematic Heterogeneity in Meta-analysis

   Lerato E. Magosi

   R package version 0.1.1. 2017.

   URL: https://cran.r-project.org/web/packages/getmstatistic/index.html

   Stata command:

   GETMSTATISTIC: Stata module to Quantify Systematic Heterogeneity in Meta-Analysis

Lerato E. Magosi

Distribution date: 20170518.

URL: https://ideas.repec.org/c/boc/bocode/s458347.html

3. The development of GWAS-derived empirical heterogeneity priors as an alternative method to improve inference on average genetic effects and heterogeneity in smaller genetic association meta-analyses.
4. Identifying an additional coronary artery disease risk variant overlooked by traditional fixed and random-effects methods.
5. Highlighting a vulnerability of the Han-Eskin random-effects method to inflated genetic signals.

List of Abbreviations

ABO	Alpha 1-3-N-acetylgalactosaminyltransferase and alpha 1-3-galactosyltransferase
ACF	Autocorrelation factor
ADAMTS9	ADAM metallopeptidase with thrombospondin type 1 motif 9
AIC	Akaike information criterion
APOE	Apolipoprotein E
BMI	Body mass index
CAD	Coronary artery disease
CARDIoGRAMplusC4D	Coronary ARtery DIsease Genome wide Replication and Meta-analysis (CARDIoGRAM) plus The Coronary Artery Disease (C4D) Genetics
CDKAL1	CDK5 regulatory subunit associated protein 1 like 1
CRAN	Comprehensive R archive network
DGI	Diabetes Genetics Initiative
ECDF	Empirical Cumulative Distribution Function
ESR	Externally studentized residuals
ESS	Effective sample size
FE	Fixed-effect
FTO	Fat mass and obesity-associated gene
FUSION	Finland-United States Investigation on NIDDM Genetics
GIANT	Genetic Investigation of ANthropometric Traits

GRB14	Growth factor receptor bound protein 14
GREB1	Growth regulation by oestrogen in breast cancer 1
GRR	Genotype relative risk
GSDMA	Gasdermin A
GSDMB	Gasdermin B
GWAS	Genome-wide association study
GWA	Genome-wide association
HDAC	Histone deacetylase
HDAC9	Histone deacetylase 9
hg19	Human reference genome 19
HOX13	Homeobox C13
IEC	International Endogene Consortium
IG	Inverse-gamma distribution
IGF2	Insulin-like growth factor 2
IGF2R	Insulin-like growth factor 2 receptor
INSIG2	Insulin induced gene 2
ITPR2-SSPN	Inositol 1,4,5-triphosphate receptor, type 2 – Sarcospan
KCNQ1	Potassium voltage-gated channel subfamily Q member 1
KLF14	Kruppel-like factor 14
LD	Linkage disequilibrium
LDL	Low density lipoprotein
LPA	Lipoprotein(a)
LPAL2	Lipoprotein(a)-like 2
LYPLAL1	Lysophospholipase-like 1

MCMC	Markov Chain Monte-Carlo
MCSE	Monte-Carlo standard error
METASTROKE	Collaboration to confirm and identify novel ischaemic stroke variants
MI	Myocardial infarction
MM	Method-of-Moments
OR	Odds ratio
PEPD	Peptidase D
PMF1	Polyamine modulated factor 1
PPAR-γ2	Peroxisome proliferator-activated receptor gamma 2 receptor
RE	Random-effects
REML	Restricted maximum likelihood
RSPO3	R-spondin 3
SD	Standard deviation
SE	Standard error
SLC22A3	Solute carrier family 22 member 3
SNP	Single nucleotide polymorphism
SPRE	Standardized predicted random-effects
SSC	Statistical software components
TCF7L2	Transcription factor-7–like 2
VEGFA	Vascular endothelial growth factor A
WEB	Weibull distribution
WTCCC	Wellcome Trust Case Control Consortium

List of Figures

Chapter 1

Figure 1.1: A brief timeline highlighting the evolution of meta-analysis to GWAS meta-analysis.

Figure 1.2: Forest plots illustrating moderate heterogeneity of study effect-size estimates (log odds ratios) at rs4420638 on chromosome 19 ($I^2 = 46.2\%$) and rs2891168 on chromosome 9 ($I^2 = 57.58\%$).

Figure 1.3: Fixed and random-effects normal probability plots for rs4420638 on chromosome 19.

Chapter 2

Figure 2.1: A normal probability plot of *SPRE* statistics at the 9p21 coronary artery disease locus (rs2891168) on chromosome 9.

Figure 2.2: Empirical and theoretical distributions of the *M* statistic.

Figure 2.3: A comparative power analysis of *M* and Cochran's Q to detect systematic heterogeneity.

Figure 2.4: The power of the *M* statistic to detect systematic outlier studies.

Figure 2.5: Power to detect systematic outlier studies in the presence of variable background heterogeneity.

Figure 2.6: Forest plot of *M* statistics summarizing systematic patterns of heterogeneity among studies in the CARDIOGRAMplusC4D GWAS meta-analysis.

Figure 2.7: Systematic heterogeneity in the CARDIoGRAMplusC4D meta-analysis can be explained by differences in age-of-CAD onset, CAD family-history and ancestry.

Figure 2.8: Inferring ancestry groups of studies in the CARDIoGRAMplusC4D meta-analysis from summary allele frequencies of lead variants at GWAS-significant loci (46).

Figure 2.9: Forest plot of *M* statistics computed across 214 loci (FDR < 5%), summarizing systematic patterns of heterogeneity among studies in the CARDIOGRAMplusC4D GWAS meta-analysis.

Figure 2.10: A scatterplot of *M* statistics computed across 214 FDR-defined loci (FDR < 5%) in the CARDIoGRAMplusC4D meta-analysis.

Figure 2.11: Forest plot of effect-size estimates (log odds ratios) at rs2891168 (chromosome 9) highlighting locus specific heterogeneity among studies in the CARDIOGRAMplusC4D GWAS meta-analysis ($I^2 = 57.69\%$).

Figure 2.12: Forest plot of effect-size estimates (log odds ratios) at rs6689306 (chromosome 1) highlighting locus specific heterogeneity among studies in the CARDIOGRAMplusC4D GWAS meta-analysis ($I^2 = 40.57\%$).

Figure 2.13: Heatmap of λ_{meta} heterogeneity statistics comparing effect-size estimates at 46 GWAS-signficant lead variants, across pairs of studies in the CARDIoGRAMplusC4D meta-analysis.

Chapter 3

Figure 3.1: An illustration of Markov chain Monte-Carlo diagnostics following a Bayesian random-effects meta-analysis of 5 studies with an inverse gamma heterogeneity prior, $\tau^2 \sim IG(15.83, 1/0.01)$.

Figure 3.2: An illustration of the use of visual inspection and Akaike information criterion to identify the best fitting theoretical distribution to an empirical distribution of heterogeneity estimates.

Figure 3.3: A comparison of the precision of Bayesian and frequentist average genetic effect estimates.

Figure 3.4: A comparison of the accuracy of Bayesian and frequentist heterogeneity estimates.

Figure 3.5. A comparison of the extent of bias in heterogeneity estimates produced by Bayesian and classical random-effects meta-analyses.

Figure 3.6: Accuracy of Bayesian heterogeneity estimates is greater near the centre of the prior distribution for heterogeneity.

Figure 3.7. A comparison of mean squared errors of heterogeneity estimates afforded by Bayesian and classical genetic association meta-analyses.

Figure 3.8: A comparison of coverage probabilities of average genetic effects estimates provided by Bayesian and classical random-effects meta-analyses.

Figure 3.9: Identifying the best fitting theoretical distribution to an empirical distribution of heterogeneity estimates in a random subset of studies from the CARDIoGRAMplusC4D meta-analysis of coronary artery disease.

Figure 3.10. A comparison of the precision of Bayesian and classical average genetic effect estimates in the CARDIoGRAMplusC4D meta-analysis of coronary artery disease, using a randomly selected subset of studies.

Figure 3.11: Identifying the best fitting theoretical distribution to an empirical distribution of heterogeneity estimates sourced from the full complement of studies (48) in the CARDIoGRAMplusC4D meta-analysis of coronary artery disease.

Figure 3.12. A comparison of the precision of Bayesian and classical average genetic effect estimates in the CARDIoGRAMplusC4D meta-analysis of coronary artery disease, using the full complement of 48 studies.

Chapter 4

Figure 4.1: Forest plots of study effect-size estimates at rs17022021 (chromosome 3) in the CARDIoGRAMplusC4D meta-analysis, illustrating an inflated genetic association where the summary effect-size estimate is inflated by an outlier study showing extreme effects.

Figure 4.2: A forest plot of effect-size estimates (log odds ratios) at rs688359 (chromosome 6) highlighting locus specific heterogeneity among studies in the CARDIoGRAMplusC4D meta-analysis, $I^2 = 34.06\%$.

Figure 4.3: Regional association plots of the *LPA* locus highlighting rs688359.

Figure 4.4: A regional association plot centred on the rs688359 (chromosome 6) lead variant.

List of Tables

Chapter 2

Table 2.1: A comparison of the means and standard deviations for the theoretical and empirical null distributions of *M*.

Table 2.2: Empirical type-1 error rates and power of *M* to detect an outlier study showing no association with the phenotype of interest across the variants examined, at threshold $\alpha = 0.05$.

Table 2.3: Number of variants included in each effect-size bin for 3 meta-analysis simulation scenarios comprising 50, 25 and 10 variants.

Table 2.4: Lead variants at 46 genome-wide significant loci (p-value $\leq 5 \times 10^{-08}$) selected for computation of *M* statistics in the CARDIoGRAMplusC4D GWAS meta-analysis.

Table 2.5: Studies showing substantial systematic heterogeneity patterns in the CARDIoGRAMplusC4D GWAS meta-analysis at threshold, $\alpha = 0.05$.

Table 2.6: Ancestry, age-of-disease onset and family-history of disease grouping for studies in the CARDIoGRAMplusC4D GWAS meta-analysis.

Table 2.7: Meta-regression of *M* statistics in the CARDIoGRAMplusC4D GWAS meta-analysis employing East Asian ancestry, age-of-CAD onset and CAD family-history as covariates.

Table 2.8: FDR-defined lead variants (214) selected for a polygenic *M* analysis of the CARDIoGRAMplusC4D GWAS meta-analysis (FDR ≤ 5%, $P < 5 \times 10^{-05}$).

Table 2.9: Studies showing systematic heterogeneity patterns across 214 FDR-defined lead variants (FDR ≤ 5%) in the CARDIoGRAMplusC4D GWAS meta-analysis at, $\alpha = 0.05$.

Table 2.10: Meta-regression of *M* statistics computed across 214 FDR-defined lead variants in the CARDIoGRAMplusC4D GWAS meta-analysis (FDR ≤ 5%).

Table 2.11: Meta-regression of effect-sizes in rs2891168 and rs6689306 employing East Asian ancestry, early age-of-CAD onset and family-history of CAD as covariates.

Chapter 3

Table 3.1: AIC statistics assessing the fit of six theoretical distributions to an empirical distribution of heterogeneity estimates.

Table 3.2. Empirically informed inverse-gamma heterogeneity priors of four genetic association meta-analysis scenarios.

Chapter 4

Table 4.1: A comparative power analysis of the Han-Eskin random-effects model and conventional fixed and random-effects methods to detect small effect genetic associations in the absence of heterogeneity.

Table 4.2: A comparative power analysis of the Han-Eskin random-effects model and conventional fixed and random-effects methods to detect heterogeneous genetic associations.

Table 4.3: A comparison of the Han-Eskin random-effects model's vulnerability to inflated genetic signals in relation to traditional fixed and random-effects methods.

Table 4.4: A re-appraisal of the CARDIoGRAMplusC4D meta-analysis of coronary artery disease with the Han-Eskin random-effects model, to identify loci overlooked by traditional fixed and random-effects methods.

Chapter 1

1.1 A brief historical note on meta-analysis

Meta-analysis relates to the use of statistical methods to quantitatively combine and compare results from similar yet independent studies. The term 'meta-analysis' was devised by an American statistician named Gene Glass, to capture the sentiment with which a meta-analysis represents an analysis of previously conducted analyses (Glass, 1976). Though the term 'meta-analysis' is fairly recent, mathematical approaches to compare and combine independent observations date back to 17th century works by Blaise Pascal and Pierre di Fermat on probability theory and gambling (Athanasiou and Darzi, 2011; O'rourke, 2007).

The first application of meta-analysis to summarise and assess the consistency of evidence across multiple clinical studies came about in the early 1900s when Karl Pearson, a British statistician was asked to evaluate the effectiveness of a typhoid (enteric) fever vaccine (Simpson and Pearson, 1904). Pearson noted the 'extreme irregularity' in the results being combined i.e. heterogeneity, and was dissatisfied with the effectiveness of the vaccine owing to 'low correlation values' between immunity and inoculation against enteric fever. Consequently, he (Pearson) recommended that an additional experiment be performed using volunteers where every second participant would receive inoculation (Shannon, 2016; Simpson and Pearson, 1904).

By the new millennium (2000s), meta-analysis had become a popular approach for resolving conflicting findings among genetic association studies of candidate genes (Ioannidis et al., 2001; Weedon et al., 2003). The availability of cheaper genotyping methods and the completion of the catalogue of human genetic variation around the mid-2000s spurred on a proliferation of genome-wide association studies (International HapMap Consortium, 2005). Confronted yet again with inconsistent replication of findings (Colhoun et al., 2003), geneticists turned to meta-analysis to summarise results across multiple genome-wide association studies (Ioannidis et al., 2007a; Zeggini et al., 2008). Today, meta-analysis is regularly performed to discover genetic signals underlying complex traits (Figure 1.1).

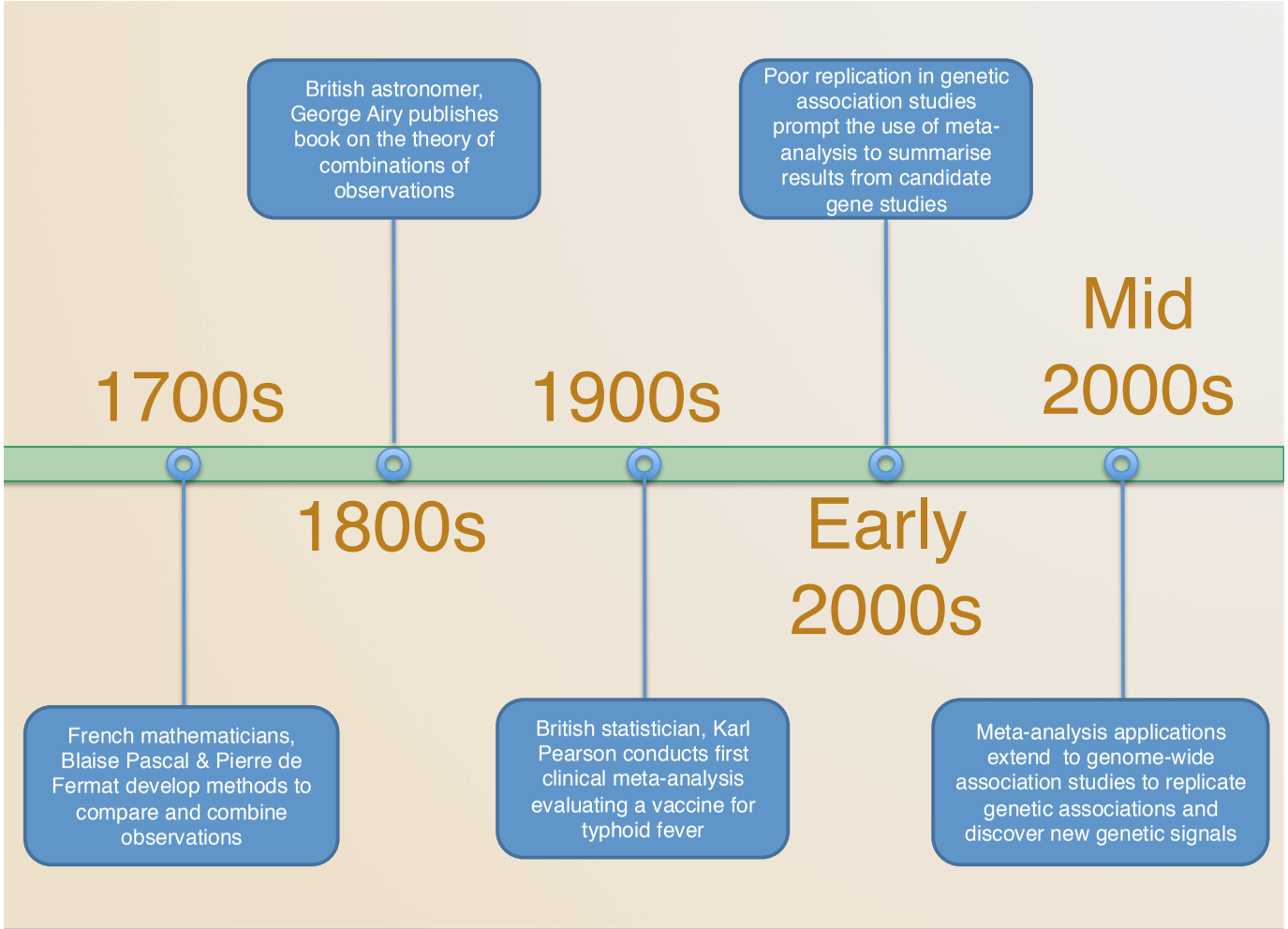


Figure 1.1: A brief timeline highlighting the evolution of meta-analysis to GWAS meta-analysis. Source: Lerato Magosi (Farrall and Hopewell labs).

1.2 An introduction to exploring heterogeneity in GWAS meta-analysis

The following section provides a rationale for dissecting heterogeneity in meta-analyses of genome-wide association studies (GWAS) i.e. GWAS meta-analyses, highlighting typical sources of heterogeneity as well as common methods to measure and account for heterogeneity in GWAS meta-analysis.

1.2.1 Importance of dissecting heterogeneity in GWAS meta-analysis

Common genetic variants underlying complex traits e.g. obesity and coronary artery disease, predominantly confer small-to-moderate effects (per-allele disease odds ratios < 1.2 or trait variance < 0.2%) (Bush and Moore, 2012; Yang et al., 2011), as such large sample sizes - on the order of (tens of) thousands of cases and controls for binary traits - are typically required for their reliable detection (Zondervan and Cardon, 2004). For example, over 20,000 cases and an equal number of controls would be required to have at least 80% power to detect (at $P < 5E-08$) a coronary artery disease risk locus with an effect-size (per-allele odds ratio) of 1.1, this is assuming a risk allele frequency of 30% and coronary artery disease prevalence estimate of 5% (CARDIoGRAMplusC4D Consortium et al., 2013). Given the challenge in marshalling sample sizes of such magnitude in stand-alone genome-wide association studies; GWAS meta-analyses came about as a way to attain larger sample sizes, and with that better power to reproducibly

discover common risk variants conferring smaller effects in complex traits. As noted by Spencer and colleagues (Spencer et al., 2009):

> It is becoming clear that many loci harbouring common alleles affecting common diseases will have effect sizes in the range 1.1– 1.2, and our simulations demonstrate that there is almost no power to detect these in studies of the size currently underway. As has already been shown empirically (Barrett et al., 2008; Zeggini et al., 2008) these loci can be found by meta-analyses and follow up in larger samples of GWA findings. (Spencer et al, 2009: 5)

Along with improved power to detect susceptibility variants, the larger sample sizes attained through meta-analysis of multiple studies allow the genetic contribution of risk variants to be estimated more precisely than could be done in stand-alone studies. Through imputation, GWAS meta-analyses can examine a wider array of variants than would ordinarily be available for single GWAS (Zeggini and Ioannidis, 2009).

Further to affording geneticists an opportunity to summarise evidence across multiple genome-wide association studies, GWAS meta-analyses importantly provide a platform to inspect the consistency or lack thereof (i.e. heterogeneity) of genetic association signals, across diverse study populations (Ioannidis et al., 2007a). Statistical heterogeneity refers to the degree of variability among the results of participating studies in a meta-analysis, beyond that expected by chance. In GWAS meta-analyses, statistical heterogeneity can present as differences in the strength and/or direction of association of variants under investigation (Dickersin and Berlin, 1992). Given that GWAS meta-analyses commonly examine millions of genetic variants at a time, my hypothesis is that statistical heterogeneity can act both locally at individual variants (i.e. marker-specific heterogeneity patterns) and globally across multiple variants (i.e. systematic or genome-wide heterogeneity patterns). In contrast to marker-specific heterogeneity patterns that arise when study-effects vary randomly at an individual variant, genome-wide heterogeneity

patterns would occur when genetic effects differ systematically from study-to-study, across a panel of variants examined in a meta-analysis.

On the one hand, excessive heterogeneity can mask genetic signals and diminish the power of a meta-analysis to detect genetic associations. For example, work done by Moonesinghe and colleagues on the reproducibility of heterogeneous genetic signals revealed that the sample size needed to detect (at $P < 1E-07$) a genetic variant with a risk allele frequency of 40% and an odds ratio (OR) of 1.2, at 80% power, was approximately 10,000 in the absence of heterogeneity, $\tau^2 = 0$ and near 100,000 in its presence, $\tau^2 = 0.007$ (Moonesinghe et al., 2008). Moonesinghe and colleagues further demonstrated with the aid of simulation experiments that there are extreme levels of heterogeneity beyond which employing larger sample sizes, in the hopes of improving power, would provide little help in detecting heterogeneous genetic signals (Moonesinghe et al., 2008). On the other hand, moderate dispersion of study-effects can provide important biological insight. For example, in Heid and colleague's meta-analysis to identify genetic variants associated with waist-hip ratio (Heid et al., 2010), the exploration of gender-related heterogeneity by stratified analyses revealed genetic variants that were more strongly associated with body fat distribution in women than men (*LYPLAL1*, *HOXC13*, *ADAMTS9*, *RSPO3*, *VEGFA*, *GRB14* and *ITPR2-SSPN*). The exploration of gender as a source of heterogeneity accentuated the role of sexual dimorphism in body fat distribution (Heid et al., 2010).

GWAS meta-analyses are generally conducted retrospectively wherein the eligibility criteria and study design features of the component studies would have been decided by the respective primary investigators independently; accordingly, heterogeneity should be expected

and explored among the GWAS datasets being combined in a meta-analysis. Differences in study design and ascertainment criteria, which commonly contribute to statistical heterogeneity include: phenotype definition, ancestry, gender, family-history of disease, age-of-disease onset, genotyping platform, quality control criteria and random chance. Further to heterogeneity brought about by methodological or clinical differences among studies in a meta-analysis; heterogeneity can also be induced by variable linkage disequilibrium (LD) patterns, epistasis or gene-environment interactions (Evangelou and Ioannidis, 2013). These sources of heterogeneity will be discussed briefly in turn citing examples from the literature.

1.2.2 Typical sources of heterogeneity in GWAS meta-analysis

Phenotype-based heterogeneity: Phenotype-based heterogeneity can arise when genetic variants under investigation are differentially associated with sub-phenotypes of the trait of interest. For example, in the 2015 CARDIoGRAMplusC4D GWAS meta-analysis to identify common and low frequency risk variants for coronary artery disease, the *ABO* and *HDAC* loci showed marked differences in their association with myocardial infarction (MI) and the broader coronary artery disease (CAD) phenotype (acute coronary syndrome, chronic stable angina, coronary stenosis > 50% in at least one coronary vessel). Whereas the *ABO* locus was more strongly associated with MI [Lead variant: rs532436, chromosome 9, OR_{CAD} = 1.08 (95% CI: 1.06–1.11), $P = 1.19 \times 10^{-11}$; OR_{MI}= 1.12 (95% CI: 1.09–1.14), $P = 2.31 \times 10^{-17}$], sub-phenotype analysis indicated that the *HDAC* locus preferentially associated with the broad CAD phenotype [Lead variant: rs2107595, chromosome 7, OR_{CAD} = 1.08 (95% CI: 1.05–1.10), $P = 8.05 \times 10^{-11}$; OR_{MI}= 1.05 (95% CI: 1.02–1.07), $P = 3.43 \times 10^{-4}$]. The dissection of phenotype-based heterogeneity underscored the role of the *ABO* locus in MI aetiology and highlighted *HDAC* as a susceptibility locus for atherosclerosis (CARDIoGRAMplusC4D Consortium, 2015).

The diverse sampling strategies employed by participating studies in a meta-analysis can also induce heterogeneity. For example, in a meta-analysis to investigate heterogeneity at the rs7566605 obesity risk locus (located near the *INSIG2* gene), Heid and colleagues found no overall association with obesity (body-mass-index (BMI) ≥ 30 kg/m^2 versus BMI < 30 kg/m^2; OR: 1.05; $P = 0.268$, $I^2 = 41\%$), but rather found a significant association between rs7566605 and extreme obesity when comparing increasingly obese cases with lean controls (BMI: ≥ 32.5,

35.0, 37.5, 40.0 kg/m^2 versus BMI < 25 kg/m^2; OR: 1.16, 1.18, 1.22, or 1.27; *P*-values: 0.001 to 0.003) (Heid et al., 2009). In the same light, studies which source controls from populations that were screened and filtered to exclude individuals with a personal history of the disease under investigation will tend to show greater power to detect genetic associations, than studies based on unscreened controls; as allele frequencies in studies comprising screened controls would provide a sharper contrast against the cases and therefore better power to detect genetic associations (Moskvina et al., 2005).

Diverse study designs as a source of heterogeneity: It has been suggested that combining population-based and family-based studies in a meta-analysis can contribute to heterogeneity (Gögele et al., 2012). Population-based studies (e.g. cases and unrelated controls) though powerful, can be prone to population stratification leading to inflated effects. By contrast, family-based studies (e.g. case-parent trios or case-sibling pairs) though less vulnerable to population stratification can be underpowered to detect genetic signals due to small study sizes. Even so, a comparative study of family and population-based study effect-size estimates for the same set of genetic variants, found little evidence of systematic overestimation or underestimation of genetic effects by either design (Evangelou et al., 2006).

When pooling case-control and prospective cohort (nested case-control) studies in meta-analyses conducted on diseases with high lethality e.g. MI and acute ischaemic stroke, there is a potential that survival bias might contribute to heterogeneity (Gögele et al., 2012); because cases with certain genetic risk profiles that are relevant to the trait under investigation might have been excluded in cross-sectional studies through mortality. Notably, survival bias is a form of selection bias where mortality due to expression of the trait of interest shrinks the potential case

pool (i.e. the number of individuals that can be recruited as cases) (Anderson et al., 2011). For example, survival bias was speculated to have been a contributing factor to the inability of a well powered meta-analysis based on case-control studies to replicate associations between risk variants (rs12425791 and rs11833579) on the 12p13 locus and ischaemic stroke (Wellcome Trust Case-Control Consortium and International Stroke Genetics Consortium, 2010); the associations had been previously identified by a meta-analysis based on longitudinal cohort studies (Ikram et al., 2009). However, there has been little empirical evidence confirming the role of survival bias as a substantial source of heterogeneity in GWAS meta-analysis. Moreover, large variability in the case-control ratios of studies in a meta-analysis can be another contributing factor to heterogeneity, where studies with a smaller proportion of cases would tend to have lower power to detect genetic signals leading to inconsistent genetic associations among studies in a meta-analysis.

Concomitant phenotypes as a source of heterogeneity: Genetic variants that show association with the trait of interest in some participating studies in a meta-analysis and not others can be a signal that the observed association is actually with a concomitant trait rather than the trait of interest. For example, in assessing the consistency of genetic associations at type-2 diabetes susceptibility loci, Ioannidis and colleagues found substantial heterogeneity at the rs8050136 locus near the *FTO* gene (summary OR: 1.13, 95% CI: 1.02 – 1.25, $I^2 \geq 75\%$) (Ioannidis et al., 2007a). Of the three participating studies in the meta-analysis, a statistically significant association between the rs8050136 locus and type-2 diabetes had only been detected in WTCCC (Wellcome Trust Case Control Consortium) and FUSION (Finland-United States Investigation on NIDDM Genetics) but not in DGI (Diabetes Genetics Initiative). An inspection of the ascertainment criteria of the studies revealed that DGI had strictly matched

cases and controls to adjust for body mass index, whilst WTCCC and FUSION had not. By contrast, participants in the DGI study were leaner than those included in the WTCCC and FUSION studies. Given that *FTO* is associated with increased BMI and obesity (Frayling et al., 2007), the heterogeneity among the three studies pointed to interplay between diabetes and obesity, where the apparent association between *FTO* and type-2 diabetes in obese individuals was actually a reflection of the relationship between *FTO* and increased BMI (Ioannidis et al., 2007a).

Ancestry-related heterogeneity: Ancestry-related heterogeneity is more likely to occur in multi-ethnic GWAS meta-analyses as they typically comprise studies conducted on diverse ancestry groups in different geographic areas (Evangelou and Ioannidis, 2013). For example, substantial ancestry-related heterogeneity was uncovered in a recent trans-ethnic meta-analysis for type-2 diabetes involving populations of East and South Asian, European, and Mexican ancestry (DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium et al., 2014). Of the 69 loci examined, 3 preferentially associated with type-2 diabetes in populations of East Asian and European descent. The **rs3786897 (*PEPD*)** locus showed significantly stronger associations with type-2 diabetes, in studies based on individuals of East Asian descent ($OR_{European}$ = 1.02, 95% CI: 0.98-1.06, P = 0.33; $OR_{EastAsian}$ = 1.17, 95% CI: 1.10-1.24, P = 3.5 x 10^{-7}; $OR_{SouthAsian}$ = 1.04, 95% CI: 0.99-1.10, P = 0.15; $OR_{Mexican}$ = 0.90, 95% CI: 0.76-1.07, P = 0.23; Cochran's Q p-value = 5.5 x 10^{-4}), whilst the **rs13233731 (*KLF14*)** locus was associated with increased risk of type-2 diabetes in the European ancestry group ($OR_{European}$ = 1.10, 95% CI: 1.06-1.13, P = 1.7 x 10^{-7}; $OR_{EastAsian}$ = 0.99, 95% CI: 0.94-1.04, P = 0.65; $OR_{SouthAsian}$ = 0.99, 95% CI: 0.94-1.04, P = 0.61; $OR_{Mexican}$ = 1.07, 95% CI: 0.93-1.23, P = 0.34; Cochran's Q p-value = 6.4 x 10^{-4}). Another locus, which showed a stronger association with type-2 diabetes in the

European ancestry group was, the **rs7903146 (*TCF7L2)*** locus ($OR_{European}$ = 1.40, 95% CI: 1.35-1.46, $P = 1.9 \times 10^{-59}$; $OR_{EastAsian}$ = 1.18, 95% CI: 1.03-1.35, $P = 1.6 \times 10^{-2}$; $OR_{SouthAsian}$ = 1.25, 95% CI: 1.19-1.32, $P = 3.6 \times 10^{-19}$; $OR_{Mexican}$ = 1.16, 95% CI: 0.97-1.38, $P = 0.11$; Cochran's Q p-value = 5.5×10^{-4}) (DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium et al., 2014).

The growing trend towards large meta-analyses conducted by international consortia suggests that ancestry-related heterogeneity will be an increasingly common occurrence in GWAS meta-analysis, and should be among typical heterogeneity sources explored by meta-regression. Explorations of ancestry-related heterogeneity are usually based on demographic information furnished by participating studies in a meta-analysis. Noting that genetic information for individuals tends to cluster by geographic origin or ancestry (Jorde and Wooding, 2004), and similar to the identification of population outliers in stand-alone GWAS (via principal component analysis), my hypothesis is that genetic distances computed from summary allele frequencies of variants independently associated with the trait of interest can be used to verify demographic information in a meta-analysis, and reveal outlier studies showing systematic allele frequency differences.

Population stratification: Population stratification can contribute to heterogeneity when studies based on particular ancestral sub-populations are more represented in the cases of a meta-analysis than others (Evangelou and Ioannidis, 2013; Hirschhorn et al., 2002). Such that, alleles which occur at a higher frequency among ancestral populations over-represented in the cases, potentially due to genetic drift or natural selection, will appear to be associated with the trait of interest leading to false positive genetic signals and inflated effects (Astle and Balding,

2009; Colhoun et al., 2003). Simulation studies (Marchini et al., 2004) have shown that the presence of structure even at modest levels as found in populations of East Asian or European descent can contribute to inflated genetic signals, especially when sample sizes are large. Genomic correction can be used to adjust for global population structure in a GWAS meta-analysis (Evangelou and Ioannidis, 2013). Essentially, study-level genomic correction entails down weighting i.e. "expanding" the standard errors of studies showing inflated association test statistics; a simple way this can be achieved is by multiplying the standard errors of studies with the square root of the genomic inflation factor. The genomic inflation factor, λ is defined as ratio of the median value of observed association statistics across the genome to the expected median for a chi-squared distribution with a single degree of freedom (Devlin and Roeder, 1999).

Variable linkage disequilibrium patterns as a source of heterogeneity: It is customary for genetic association studies to detect markers (tagging variants) in LD with the causal variant(s) rather than to identify the causal variants themselves. In a meta-analysis, heterogeneity may arise when the frequency of tagging variants and the magnitude of the genetic effects they confer differ among ancestry groups with the: prevalence of the trait, environmental risks and LD patterns (Ioannidis et al., 2007a; Ntzani et al., 2012). Notably, populations of African descent generally have more haplotype diversity and lower levels of LD than populations of European and Asian ancestry (Gabriel et al., 2002) (International HapMap Consortium et al., 2007). Variation in LD patterns at the population level can be caused by genetic drift and admixture (Shifman et al., 2003). Further to that, the lower levels of LD in populations of African descent could be attributed to them having accumulated a larger number of recombination events over

time. Consequently, genetic effects detected by tagging variants in weak LD with functional polymorphisms (e.g. in low LD populations) may be systematically weaker than those found in high LD populations.

Additionally, heterogeneity may arise when causal variants have distinct markers with varying LD levels in different ancestry groups. One example came to light in a meta-analysis investigating heterogeneity at type-2 diabetes susceptibility loci identified in studies based on individuals of Finnish, Swedish and British (UK) descent (Ioannidis et al., 2007a). Among the 11 variants examined, marked heterogeneity was observed at the weakly associated rs9300039 risk locus located at an intergenic region on chromosome 11. Of the 3 GWAS datasets (DGI, FUSION, WTCCC) included in the meta-analysis, 2 had used different tagging variants to detect the rs9300039 risk locus; DGI had used a multi-marker tag whilst WTCCC had employed rs1514823 (r^2 = 0.97 with rs9300039) (Scott et al., 2007). Suggesting that, a portion of the observed variation at the rs9300039 risk locus might have been a result of variable LD levels between the marker and causal variants. In such scenarios, genotyping nearby variants on the same gene (i.e. fine-mapping studies) might be helpful in identifying variants directly associated with the trait of interest at the locus in question (Ioannidis et al., 2007a; Salanti et al., 2005).

Allelic and locus heterogeneity: Inconsistent genetic signals may also occur as a result of allelic and locus heterogeneity, particularly in multi-ethnic meta-analyses. When multiple polymorphisms on the same gene give rise to the same phenotype, as is the case with allelic heterogeneity, or functional polymorphisms at different loci induce the same phenotype, as found in locus heterogeneity, the strength and significance of genetic associations between the variants examined and the trait of interest may vary from study-to-study (Salanti, 2005); one reason being that, functional polymorphisms underlying a trait of interest in one ancestry group

may be different from or have lower allele frequencies than those in another ancestry group. Notably, the contribution of variable effect allele frequencies or genetic heterogeneity to inconsistent genetic signals was highlighted in studies attempting to replicate genetic associations of type-2 diabetes susceptibility loci (e.g. *KCNQ1, TCF7L2* and *CDKAL1*) across various ancestry groups (McCarthy, 2008; Morris, 2011; Unoki et al., 2008). In particular, the *KCNQ1* risk locus was found to be significantly associated with type-2 diabetes in genome-wide association (GWA) studies primarily based on East Asian populations (Unoki et al., 2008), but fell below the significance threshold set for replication in studies conducted on individuals of European descent (McCarthy, 2008). The fact that the risk allele at the *KCNQ1* locus is more common in East Asians than in Europeans (40% in East Asians versus 5% in Europeans) made it easier for studies based on East Asian populations to detect the association between *KCNQ1* and type-2 diabetes, even though their sample size was smaller than that of the European-descent studies (East Asian ancestry study < 200 cases; European ancestry studies > 4500 cases) (McCarthy, 2008; Unoki et al., 2008). In line with the latter point, calculations by McCarthy have shown that the power to detect variants conferring effects of similar magnitude can vary considerably when those variants have different risk allele frequencies (McCarthy, 2008). In particular, McCarthy showed that for similarly sized studies, each with 2000 cases and 2000 controls, the power to detect a variant with an effect-size (per-allele odds ratio) of 1.35 would be over 80% in scenarios where the minor allele frequency was approximately 40%, as was the case with the *KCNQ1* risk allele in East Asians, but would drastically fall to 1% where the minor allele frequency was ~ 5% as was found in Europeans (McCarthy, 2008).

In contrast to the *KCNQ1* locus, the risk allele for the *TCF7L2* locus is more common in European ancestry populations than in East Asian ones (30% in Europeans versus 3% in East

Asians); as a result, studies conducted on individuals of European descent are likely to be better powered to detect the genetic association between *TCF7L2* and type-2 diabetes than studies based on East Asian populations (McCarthy, 2008). With regards to the *CDKAL1* locus, stronger associations with type-2 diabetes were observed in populations of Japanese American and Native Hawaiian descent than in other American sub-ancestry groups (African Americans, European Americans and Latinos) (Morris, 2011). The allelic heterogeneity at the *CDKAL1* locus was speculated to be a result of the presence of multiple causal variants, some of which could be specific to populations of Japanese American and Native Hawaiian descent, or perhaps variable LD patterns among the ancestry groups (Morris, 2011). Genetic heterogeneity has the potential to dilute the power to detect genetic signals in meta-analyses based on diverse ancestral populations (Asimit et al., 2012; Fisher and Lewis, 2008). To optimize power for discovery of genetic associations in the presence of genetic heterogeneity, trans-ethnic meta-analysis methods have been developed which factor in the expected similarity of allele frequencies among closely related ancestral populations, and allow for heterogeneity among distantly related ancestry groups (Morris, 2011).

Winner's curse: The winner's curse phenomenon or "Jack pot effect" can be a contributing factor to heterogeneity in a meta-analysis when there is modest agreement between the results of the first study to publish a significant association with the trait of interest and subsequent research (Ioannidis et al., 2001; Salanti et al., 2005); this is because power calculations to detect genetic signals are usually based on effect-sizes reported by previous studies. Thus, overestimation of the strength of genetic associations in earlier studies can lead to subsequent studies underestimating the sample sizes required to detect the underlying risk variants for the trait of interest (Salanti et al., 2005). Upwardly biased effect-size estimates in earlier studies tend

to be more common in scenarios where the first study comprises relatively few participants (a couple hundred cases and controls) and/or the magnitudes of the genetic effects estimated are small (Hirschhorn et al., 2002; Ioannidis et al., 2001).

One example of the "jack pot effect" was noted by Hirschhorn and colleagues in studies investigating an association between a non-synonymous mutation (Pro-12-Ala) on the peroxisome proliferator-activated gamma 2 receptor (*PPARγ2*) and decreased risk for developing type-2 diabetes (Hirschhorn et al., 2002). An earlier study by Deeb and colleagues had reported a large effect association (OR: 4.35, 95% CI: 1.24− 15.3, $P = 0.028$) between *PPARγ2* (Pro-12-Ala) and type-2 diabetes in populations of Japanese American ancestry (Deeb et al., 1998). However, in a subsequent similarly sized study (91 cases and 54 controls) there was little evidence to support the *PPARγ2* (Pro-12-Ala) polymorphism as a protective variant for type-2 diabetes in individuals of Italian descent (OR: 0.66, 95% CI: 0.36 − 1.24, $P = 0.22$) (Mancini et al., 1999); one might speculate that the play of chance might have had a role as well. The association between *PPARγ2* (Pro-12-Ala) and decreased risk for type-2 diabetes was later confirmed by several larger studies (Altshuler et al., 2000; Mori et al., 2001). In contrast to the initial report by Deeb and colleagues, the findings of the larger studies were that the *PPARγ2* (Pro-12-Ala) polymorphism modestly lowered type-2 diabetes risk (Deeb et al., 1998). The study by Altshuler and colleagues was conducted on over 3000 individuals of Finnish, Swedish and French Canadian descent and reported an effect-size estimate of (OR: 0.78, 95% CI: 0.59 − 1.05, $P = 0.045$) (Altshuler et al., 2000). Mori and colleague's study was comparable to that of Altshuler and colleagues in size (2,201 cases and 1,212 controls), and similarly reported a conservative effect-size estimate for the *PPARγ2* (Pro-12-Ala) polymorphism in individuals of Japanese ancestry (OR: 0.57, 95% CI: 0.43 − 0.75, $P = 0.000054$). Notably, there are a variety of methods that can be applied to adjust for the ascertainment bias brought about by the winner's

curse phenomenon (Bigdeli et al., 2016); essentially, these methods shrink the effect-size estimates of the most extreme genetic signals in a meta-analysis thereby reducing the impact of heterogeneity due to the "Jack pot effect".

Variable genotyping platforms: The inclusion of studies genotyped on different platforms is often mentioned as a potential source of heterogeneity in meta-analyses (Evangelou and Ioannidis, 2013; Ioannidis et al., 2007a; Thompson et al., 2011). Even though imputation is used to arrive at a common set of variants to be investigated across the participating studies, the concern is that combining data from different chips that have different error rates and were possibly produced by different manufactures may add to the variation of effects in a meta-analysis (Gögele et al., 2012). That said there has been little empirical evidence to support the inclusion of data derived from diverse genotyping platforms as a substantial source of heterogeneity in meta-analyses (Salanti et al., 2005).

Variable quality control cut-offs: Inconsistent quality control thresholds are another potential source of heterogeneity in a meta-analysis (Begum et al., 2012; Gögele et al., 2012). In that, the assembly of studies with substantial differences in the stringency of the quality control criteria employed (e.g. genotype call rate, sample missingness, differential missingness between cases and controls, heterozygosity scores, cryptic relatedness, population structure, Hardy Weinberg equilibrium deviations and imputation quality) can contribute to systematic variation in a meta-analysis. These systematic differences may result in some of the studies lowering the power of the meta-analysis to detect genetic associations rather than improving it. In the case of population stratification, such systematic variation is likely to increase the frequency of false positives in the meta-analysis (Evangelou and Ioannidis, 2013; Winkler et al., 2014).

Family-history and age-of-disease onset: Genetic enrichment is one of the strategies that can be employed to optimize the power of GWA studies to detect genetic signals; this is based on the idea that sampling cases with a greater genetic load will create a sharper contrast with the controls and in turn improve power to detect genetic associations. In accordance with the multi-locus threshold liability model, individuals that are likely to have a greater genetic load (i.e. higher accumulation of risk loci) are ones with: lower lifestyle risk profiles (e.g. for CAD: a non-smoker who is physically active and has low LDL-cholesterol), an earlier age-of-disease onset and/or a family-history of disease (Risch, 2001; Watkins and Farrall, 2006). Thus, a meta-analysis which combines genetically enriched studies- where cases might have been selected on the basis of an early age-of-disease onset (e.g. in the case of coronary disease) and/or having first-degree relatives that express the trait of interest- with non-enriched studies that have a lower genetic load is likely to see systematically stronger effects (larger odds ratios) across the loci examined, on the side of the genetically enriched studies. This systematic difference in the observed effects can be an underlying factor in the heterogeneity of a meta-analysis.

Heterogeneity related to gene-gene and gene-environment interactions: Complex traits are so called as they have both a genetic and an environmental component. As such, genetic effects conferred by variants predisposing individuals to complex traits can be modified by both other genes as well as lifestyle factors including: diet, smoking, alcohol consumption, air pollution and physical activity (Hunter, 2005). For example, recent work by Young and colleagues investigating how various lifestyle variables modify the effect of the *FTO* gene on body-mass-index in the UK Biobank dataset (a large prospective study comprising 500,000

individuals aged 45-69), uncovered independent interactions between the *FTO* (rs1421085) variant and frequency of alcohol intake ($P = 3.0 \times 10^{-4}$), overall diet ($P = 5.0 \times 10^{-6}$) and physical activity ($P = 3.1 \times 10^{-4}$) (Young et al., 2016). The environment as relates to complex traits can be broadly defined as any physical or chemical factors affecting expression of the trait of interest in the populations being studied. Differences in the lifestyle factors of study populations included in a meta-analysis can contribute to heterogeneity in the strength of the genetic associations (Salanti et al., 2005). However, the exploration of heterogeneity owing to gene-environment interactions in meta-analyses is commonly limited by the availability of anthropometric and lifestyle data for the genetic information at hand.

Gene-gene interactions are commonly acknowledged as a potential source of heterogeneity in meta-analyses (Salanti et al., 2005). However, the numerous combinations involved in the analysis of gene-gene interactions have restricted progress on dissecting heterogeneity due to epistasis; the nub has been the large potential for false positives and the burden of multiple testing (Thompson et al., 2011). That said, Hall and colleagues have recently introduced an approach which uses pathway information and protein-protein interactions to whittle down the number of variant combinations examined in gene-gene interaction analyses, thereby improving power to detect epistatic effects (Hall et al., 2015).

The impact of the different heterogeneity sources described above is typically examined in genetic association meta-analyses on a variant-by-variant basis using single-variant methods (e.g. Q, I^2, τ^2 and single-variant meta-regression), as will be discussed in section 1.2.4. However, the relative impact of these heterogeneity sources across genetic signals generally goes unreported due to the lack of appropriate methodology.

1.2.3 Popular approaches for conducting meta-analysis

Fixed-effect model: Meta-analyses are typically conducted under two types of models, namely the fixed-effect and random-effects models. Whereas the main objective under a fixed-effect model is to estimate a summary effect across the component studies of a meta-analysis, the aim of a random-effects meta-analysis is to estimate both the summary effect and the extent of heterogeneity in a meta-analysis. To compute a summary effect estimate (i.e. weighted mean), $\hat{\theta}$ across the effect-size estimates of studies included in a meta-analysis, both models assume that the observed effect in each participating study is normally distributed. Additionally, effect-size estimates of the individual studies are assigned weights corresponding to their reciprocal total observed variances; as a result, studies with larger sample sizes are assigned larger weights under the fixed-effect model. In the case of binary traits (e.g. CAD) the observed effects for studies in a GWAS meta-analysis are usually log-transformed regression coefficients from a logistic regression (e.g. per allele log odds ratios). With regards to quantitative traits, the study effect-size estimates are typically log-transformed regression coefficients from a linear regression (Evangelou and Ioannidis, 2013). Additionally, the fixed-effect model assumes that all participating studies in a meta-analysis have the same underlying effect (i.e. a common effect) and the only source of variability is sampling error. Such that, the underlying effect for a meta-analysis comprising S studies is, $\theta_1 = \theta_2 = \theta_3 = \cdots = \theta_s = \theta$ and the observed effect, y_s in the s^{th} study is given by

$$y_s \sim N(\theta, \sigma_s^2),$$

where,

y_s: observed effect for the s^{th} study

θ: underlying effect common to all participating studies in the meta-analysis

σ_s^2: sampling variance for the s^{th} study

or simply,

$$y_s = \theta + \epsilon_s \quad \text{where} \quad \epsilon_s \sim N(0, \sigma_s^2),$$

and

y_s: observed effect for the s^{th} study

θ: underlying effect common to all participating studies in the meta-analysis

ϵ_s: the error associated with estimating the common effect in the s^{th} study

σ_s^2: sampling variance for the s^{th} study.

Random-effects model: The random-effects model assumes that the underlying effect-size might be the same across all studies as would be the case for a fixed-effect model or that it might vary from study-to-study due to heterogeneity. Hence we use the singular (effect) when referring to the fixed-effect model and plural (effects) for the random-effects model (Borenstein et al., 2010). The underlying effects in a random-effects model are often assumed to be random samples from a grand population that is normally distributed, $\theta_1, \theta_2, \theta_3, \cdots, \theta_s \sim N(\theta, \tau^2)$ such that the centre of the distribution denotes the average underlying effect (i.e. summary effect) in the grand population and the spread, the amount of heterogeneity. Therefore it is important to note that estimates for the average underlying effect and extent of heterogeneity will likely be imprecise when meta-analyses comprise few studies e.g. (< 5 studies) (Higgins et al., 2009). The normal distribution is typically preferred as the distribution for the underlying effects due to its simplicity, in that it only involves two parameters being the mean and the variance. In contrast

to the fixed-effect model, the random-effects model has two sources of variability that contribute to the total observed variance in a meta-analysis. These are within-study and between-study variation also referred to as sampling variance and heterogeneity, respectively. Under a random-effects model, the observed effect for the s^{th} study in a meta-analysis comprising S studies is denoted by

$$y_s \sim N(\theta, \sigma_s^2 + \tau^2),$$

where,

y_s: observed effect for the s^{th} study

θ: mean of the distribution of the underlying effects i.e. the summary effect

σ_s^2: sampling variance for the s^{th} study

τ^2: amount of marker-specific heterogeneity in the meta-analysis

or equivalently,

$$y_s = \theta + \epsilon_s + u_s \text{ where } \epsilon_s \sim N(0, \sigma_s^2) \text{ and } u_s \sim N(0, \tau^2),$$

so that,

y_s: observed effect for the s^{th} study

θ: mean of the distribution of the underlying effects i.e. the summary effect

ϵ_s: error contribution of the sampling variance to the observed effect in the s^{th} study

σ_s^2: sampling variance for the s^{th} study

u_s: error contribution of the between-study variance to the observed effect in the s^{th} study

τ^2: amount of marker-specific heterogeneity in the meta-analysis.

The traditional random-effects model is generally less powerful at detecting genetic signals than the fixed-effect model (Thompson et al., 2011); this is because the traditional random-effects model has a conservative null hypothesis, which assumes the presence of heterogeneity under the null. Recently, contemporary random-effects models have been developed which have a relaxed null hypothesis that presumes the absence of heterogeneity under the null, thereby granting the random-effects model greater power (than the fixed-effect model) to detect genetic associations in the presence of heterogeneity (Han and Eskin, 2011). Confidence intervals for the summary effect estimate tend to be wider under the random-effects model than in the fixed-effect model; this is because the random-effects model takes into account two sources of variation (sampling variance and heterogeneity) whilst the fixed-effect model is just concerned with sampling variance. Due to the additional source of variation study weights under the random-effects model tend to be more similar in magnitude and smaller than those under the fixed-effect model. Study weights are computed as the inverse of the total observed variance such that the weight allocated to the s^{th} study under the random-effects model is

$$w_s = \frac{1}{(\sigma_s^2 + \hat{\tau}^2)},$$

where

w_s: weight for the s^{th} study

σ_s^2: sampling variance for the s^{th} study

$\hat{\tau}^2$: estimate of the amount of marker-specific heterogeneity in the meta-analysis.

And the weight assigned to the s^{th} study under the fixed-effect model is

$$w_s = \frac{1}{(\sigma_s^2)},$$

where,

w_s: weight for the s^{th} study

σ_s^2: sampling variance for the s^{th} study.

The summary effect estimate, $\hat{\theta}$ is then computed as the inverse-variance weighted mean of the observed study effects in the meta-analysis,

$$\hat{\theta} = \frac{\sum_{s=1}^{S} w_s y_s}{\sum_{s=1}^{S} w_s}.$$

Given that study weights under the random-effects model are generally smaller, the variance of the summary effect estimate under the random-effects model will often be larger than that under the fixed-effect model leading to wider confidence intervals. The variance of a summary effect estimate, $V_{\hat{\theta}}$ is computed as the reciprocal of the sum of the weights such that,

$$V_{\hat{\theta}} = \frac{1}{\sum_{s=1}^{S} w_s}.$$

Then invoking the central limit theorem, the upper and lower bounds of the 95% confidence interval surrounding the summary effect estimate are given by

$$LB = \hat{\theta} - 1.96\, x\, \sqrt{V_{\hat{\theta}}}\ ,$$

and

$$UB = \hat{\theta} + 1.96\, x\, \sqrt{V_{\hat{\theta}}}\ .$$

The fact that GWAS meta-analyses typically examine millions of genetic variants at a time can complicate the decision of whether to use a fixed or a random-effects model. In that, a proportion of variants examined in a meta-analysis may have the same allelic effects across all populations whilst others may have allelic effects that vary among the studies included in the meta-analysis. One practical implication of conducting a fixed-effect meta-analysis on heterogeneous variants is that the contribution of heterogeneity to the observed variation in the effects is ignored resulting in misleadingly small p-values. The indiscriminate application of random-effects meta-analysis is no help either as there may be instances where variants show heterogeneity just by chance leading to needlessly conservative p-values (Thompson et al., 2011). The popular approach in regards to GWAS meta-analyses has been to start off with a fixed-effect meta-analysis and then re-appraise variants that show statistically significant heterogeneity using a random-effects meta-analysis (CARDIoGRAMplusC4D Consortium, 2015; Hardy and Thompson, 1998). Though a rational approach Borenstein and colleagues advise against this practice on the following grounds (Borenstein et al., 2010):

1. Tests of heterogeneity are typically underpowered to diagnose heterogeneity when the meta-analysis includes few studies. As such non-significant heterogeneity tests may not always be indicative of the absence of heterogeneity.

2. The random-effects model would be a better default as it caters both for scenarios where the underlying effect might be identical across the studies included in the meta-analysis

as well as ones where there might be less consistency among the underlying effects. Especially since contemporary random-effects methods now offer improved power to detect genetic signals (Han and Eskin, 2011).

3. The two models are based on different ideas. In particular the fixed-effect model is meant to estimate a common effect that is specific to the studies included in the meta-analysis, whilst the random-effects model estimates a grand mean such that the results of the meta-analysis can be generalized beyond the study populations included in the meta-analysis.

1.2.4 Common heterogeneity measures in GWAS meta-analysis

Measures commonly used to explore heterogeneity in genetic association meta-analyses include: the Q-statistic, Higgin's inconsistency metric, I^2 and the τ^2-statistic (Begum et al., 2012; Evangelou and Ioannidis, 2013). These three measures are typically applied to detect and quantify the impact and extent of heterogeneity at individual variants in genetic association meta-analyses; and thus can be categorised as single-variant (or marker-specific) heterogeneity methods.

1.2.4.1 Detecting heterogeneity in meta-analyses

Q-statistic: The Q-statistic is popularly known as Cochran's Q in relation to William Cochran's work combining results from field experiments conducted at different agricultural stations (Cochran, 1937, 1954). However, it might be useful to remember that Cochran used a transformation of the Q-statistic rather than the Q-statistic itself to assess heterogeneity or interactions as he called them (Cochran, 1937; Hoaglin, 2016). In physics, the Q-statistic had been in use at least as early as 1932 where a ratio of the Q-statistic to its degrees of freedom, also known as the Birge ratio was used to assess the consistency of estimates of physical constants (Birge, 1932; Hedges, 2016). An earlier description of the Q-statistic can also be found in Merriman's book on least squares, which describes a general (weighted) mean for the combination of several measurements of the same quantity made with varying levels of precision (Merriman, 1909).

Computing Q-statistics: The Q-statistic forms the basis of several heterogeneity measures. It is commonly used in combination with the inspection of forest plots to diagnose the presence of heterogeneity in a meta-analysis. For a meta-analysis comprising S studies, the Q-statistic is computed as the weighted sum of the squared deviations of the individual study effects from the combined mean effect (i.e. summary effect estimate),

$$Q = \sum_{s=1}^{S} w_s \left(y_s - \hat{\theta}\right)^2,$$

where

$$w_s = \frac{1}{\sigma_s^2},$$

and

y_s: observed genetic effect for the s^{th} study

$\hat{\theta}$: summary effect estimate

w_s: weight assigned to the s^{th} study

σ_s^2: sampling variance for the s^{th} study.

Q-statistic null distribution: When the number and size of studies in a meta-analysis are large the distribution of the Q-statistic in the absence of heterogeneity (i.e. under the null) is approximately chi-squared, with $S - 1$ degrees of freedom, where S is the number of studies (Hoaglin, 2016; Kulinskaya and Dollinger, 2016). This is based on the assumption that the

observed study effects are normally distributed, and that the sampling variances are treated as known even though they are typically estimated from the data in each study (Hardy and Thompson, 1998; Hoaglin, 2016). To test for the presence of heterogeneity the Q-statistic is compared to a chi-square distribution in a one-tailed test, where by convention a p-value < 0.1 indicates the presence of heterogeneity (rather than p-value < 0.05) (Higgins et al., 2003; Zeggini and Ioannidis, 2009).

For meta-analyses with sparse data (e.g. small sample sizes with under 100 participants) where the observed study effects may not be normally distributed, simulation studies have shown that the null distribution of the Q-statistic varies with the type of effect measure; such that, it approximates a gamma distribution when log odds ratios are used and a chi-square distribution with fractional degrees of freedom for meta-analyses conducted using standardized mean differences (Hoaglin, 2016; Kulinskaya and Dollinger, 2015; Kulinskaya et al., 2011).

Q-statistic power: The Q-statistic's power to detect heterogeneity depends on the number of studies included in a meta-analysis, their combined weight (sum of inverse variances) and the distribution of that weight among the individual studies (Hardy and Thompson, 1998; Rücker et al., 2008). Consequently, the Q-statistic's power to detect heterogeneity is optimal when studies are equally weighted (i.e. the expectation of the Q-statistic, E(Q) is maximal). However, when the bulk of the information in a meta-analysis is provided by a few large studies (e.g. one or two studies contributing 50% or 90% of the total weight), the power of the Q-statistic degrades due to the uneven distribution of study weights. Furthermore, it is easier for the Q-statistic to detect heterogeneity when the amount of heterogeneity in a meta-analysis is high; this is because the expectation of the Q-statistic and in turn its power increases linearly with the amount of

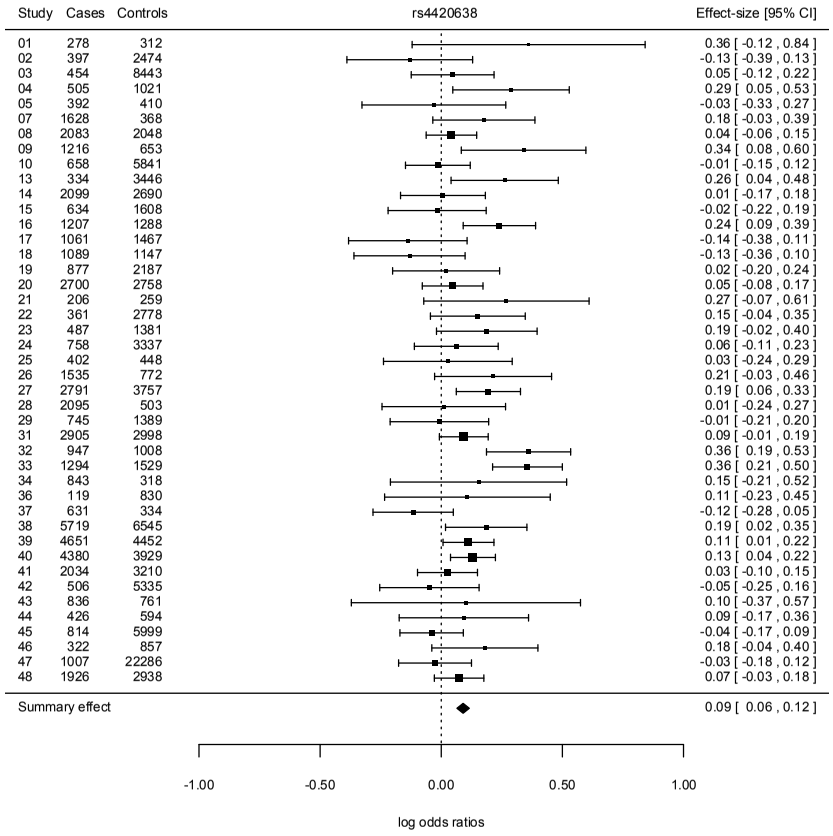
heterogeneity in a meta-analysis (Hardy and Thompson, 1998). The Q-statistic's power to detect heterogeneity depends more on the total information available in a meta-analysis (i.e. total weight) than the number of studies included; in that, there is little impact on the power to detect heterogeneity when the number of studies is increased but the total amount of information in the meta-analysis remains the same (Hardy and Thompson, 1998).

Limitations of the Q-statistic: When meta-analyses comprise few studies (e.g. < 20 studies) and/or have small sample sizes, the Q-statistic tends to have low power to detect the presence of heterogeneity (Begum et al., 2012; Evangelou and Ioannidis, 2013; Thorlund et al., 2012). This means that a non-significant result does not necessarily rule out the presence of heterogeneity. Moreover, the Q-statistic tends to be too conservative (i.e. excessively powered) when meta-analyses contain many large studies; the practical implication being that, negligible study differences of little importance in a GWAS context are sometimes reported as statistically significant (Hardy and Thompson, 1998; Higgins and Thompson, 2002). Furthermore, Q-statistic values increase with the precision and number of studies in a meta-analysis (Rücker et al., 2008). The Q-statistic's dependence on the number of studies limits its comparability across meta-analyses of different sizes (Hardy and Thompson, 1998).

Diagnosing heterogeneity by visual inspection: Further to quantitative detection by the Q-statistic, heterogeneity can be identified and described graphically using a variety of plots (Anzures-Cabrera and Higgins, 2010). GWAS meta-analyses regularly employ forest plots and normal probability plots to visually assess the consistency of study effects and identify outlying studies.

Forest plots: A forest plot provides a graphical summary of a meta-analysis displaying effect-size estimates for the studies included in the meta-analysis, their accompanying confidence intervals and an estimate of the summary effect (with its corresponding confidence interval). To highlight the contribution of each study to the meta-analysis data points denoting effect-size estimates for the individual studies are scaled such that their sizes are proportional to the weight allocated to each study. When the plot is generated using unsorted study effect-size estimates the scattered arrangement of data points often resembles a forest of lines, hence the name forest plot (Lewis and Clarke, 2001) (Figure 1.2A-B). The presence of heterogeneity in forest plots is typically indicated by little overlap among the confidence intervals of study effect estimates (Anzures-Cabrera and Higgins, 2010).

A



B

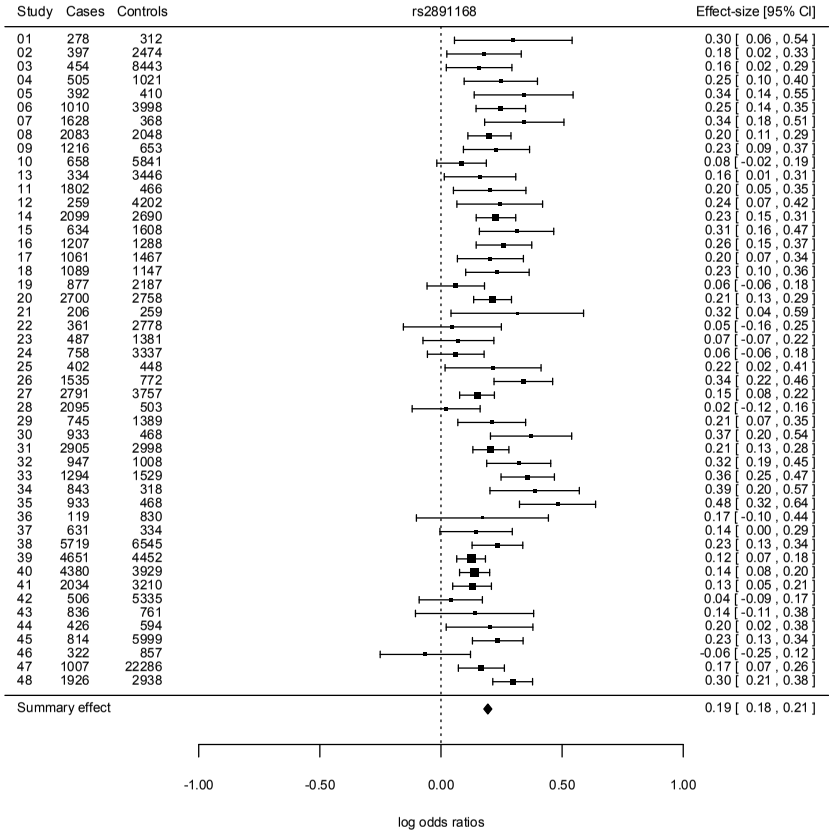


Figure 1.2: Forest plots illustrating moderate heterogeneity of study effect-size estimates (log odds ratios) at rs4420638 on chromosome 19 $(I^2 = 46.2\%)$ **and rs2891168 on chromosome 9** $(I^2 = 57.58\%)$**.** Panels A and B show dispersion of study effect-size estimates at rs4420638 and rs2891168, respectively. Filled squares denote individual study effect-size estimates with the 95% confidence intervals shown by horizontal lines; the sizes of the squares are proportional to each studies' inverse-variance weighting. A filled diamond represents the summary effect-size estimate. Source: Lerato Magosi (Farrall and Hopewell labs).

Normal probability plots: Normal probability plots are useful for identifying outlier studies and checking normality assumptions in fixed and random-effects meta-analysis models (i.e. study effect estimates and their variances are drawn from normal distributions). To explore heterogeneity, normal probability plots can be generated where the contribution of each study in a meta-analysis to the total observed variation (Q-statistic) is plotted against the quantiles of a standard normal distribution (Hardy and Thompson, 1998) (Figure 1.3A-B). For a meta-analysis with S participating studies the contribution of the s^{th} study to the Q-statistic is,

$$q_s = \sqrt{w_s}\left(y_s - \hat{\theta}\right) \quad \text{where} \quad Q = \sum_{s=1}^{S} q_s^2 \ ,$$

and

y_s: observed genetic effect for the s^{th} study

$\hat{\theta}$: summary effect estimate

w_s: weight assigned to the s^{th} study

q_s: contribution of the s^{th} study to the Q-statistic.

Presuming the number and size of studies in a meta-analysis is large a linear normality plot with a slope of 1 typically indicates that the normality assumptions have been met. A linear normality plot with deviations in the upper and/or lower tails might indicate the presence of outliers (Anzures-Cabrera and Higgins, 2010). A limitation of assessing heterogeneity solely by visual inspection is the subjectivity of the analyst. However, the combination of visual inspection and the Q-statistic test generally provides stronger evidence for the presence or absence of heterogeneity in a meta-analysis.

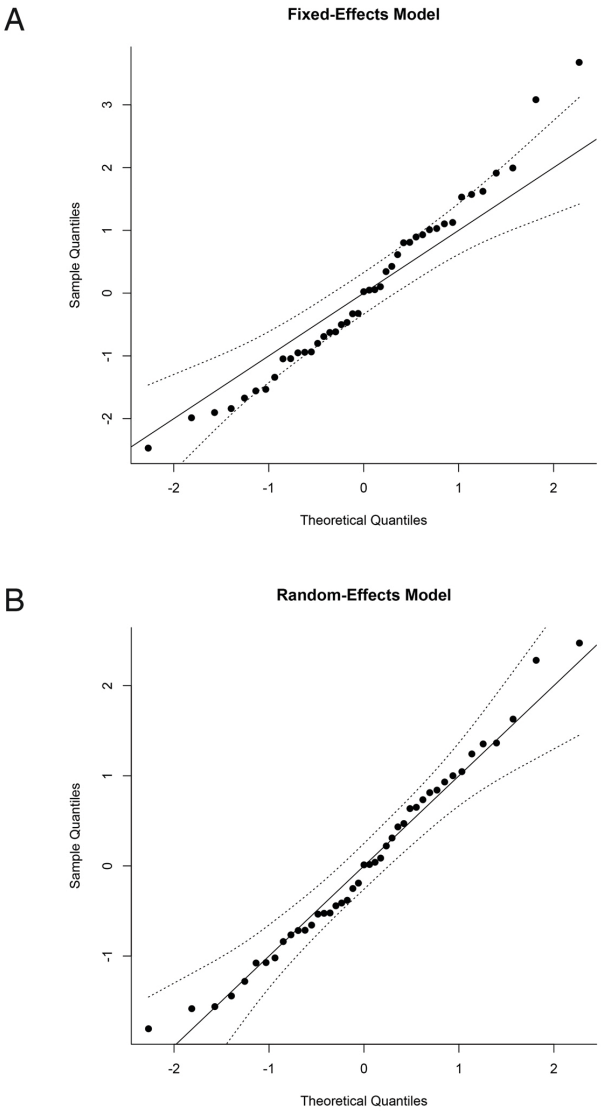


Figure 1.3: Fixed and random-effects normal probability plots for rs4420638 on chromosome 19. The solid line represents the line of identity and the broken lines denote the confidence bounds. Panel A shows a fixed-effect normal probability plot and panel B, a random-effects normal probability plot. The deviation of study-effect estimates beyond the confidence bounds in panel A suggests the underlying effects differ from study-to-study and might be better represented by a normally distributed random-effects model as shown in panel B. Source: Lerato Magosi (Farrall and Hopewell labs).

1.2.4.2 Quantifying the impact of heterogeneity in meta-analyses

I^2 metric: The I^2 metric is a descriptive heterogeneity measure that quantifies the relative increase in variability among observed study effects as a result of heterogeneity (Thompson and Higgins, 2002). The fact that the I^2 metric is expressed on a percentage scale and is easy to interpret has contributed to its wide use as a measure of heterogeneity in meta-analysis. The I^2 metric is a transformation of the Birge ratio (also called H^2),

$$H^2 = \frac{Q}{df} ,$$

$$df = number\ of\ studies - 1 ,$$

and is expressed as,

$$I^2 = \frac{H^2 - 1}{H^2} \ x \ 100 ,$$

substituting H^2 with $\frac{Q}{df}$ and 1 with $\frac{df}{df}$ yields,

$$I^2 = \frac{Q - df}{Q} \ x \ 100 ,$$

where,

Q: Q-statistic denoting the total observed variation in the study effects

df: Degrees of freedom representing the expected variation in the study effects if the only source of variability was sampling error.

Equivalently,

$$I^2 = \frac{\hat{\tau}^2}{(\hat{\tau}^2 + \sigma^2)} \; x \; 100\,,$$

where,

$\hat{\tau}^2$: estimate of the amount of heterogeneity in the meta-analysis

σ^2: sampling variance.

Typically, I^2 values of 0-25%, 25-50%, 50-75% and >75% are broadly interpreted as benchmarks for low, moderate, high and very high heterogeneity, respectively (Evangelou and Ioannidis, 2013; Higgins et al., 2003). However, the confidence intervals surrounding I^2 values can be wide ranging from low $(I^2 \leq 25\%)$ to high $(I^2 \geq 75\%)$ relative increases in the observed variation due to heterogeneity. As such, even when $(I^2 = 0\%)$ one should consider that there might be some level of inconsistency in the observed study effects. The I^2 metric's power to detect heterogeneity increases with the amount of heterogeneity in a meta-analysis (Huedo-Medina et al., 2006; Ioannidis et al., 2007b). Similar to the Q-statistic, I^2 values can be compared across meta-analyses conducted using different effect measures. Also in common with the Q-statistic, the I^2 metric has poor power to detect heterogeneity when studies are few e.g. < 10 studies (Huedo-Medina et al., 2006) (Thorlund et al., 2012); this is because the I^2 metric is a

monotone increasing function of the Q-statistic. As such, when Q-statistic values are low, as in the case of meta-analyses with few studies, I^2 values tend to be low as well, the end result being that the fraction of total observed variation due to heterogeneity is underestimated (Thorlund et al., 2012).

Moreover, values of the I^2 metric tend to show a greater bias when studies are few (e.g. < 10 studies). In particular, simulation studies have shown a positive bias in I^2 values - that is an overestimation of the percentage of observed variation due to heterogeneity - when the actual amount of heterogeneity is minimal, and a negative bias in I^2 values when the extent of heterogeneity is considerable (von Hippel, 2015). The I^2 metric depends on the precision of studies included in a meta-analysis, from the mathematical representation of the I^2 metric, $I^2 = \frac{\hat{\tau}^2}{(\hat{\tau}^2 + \sigma^2)} \, x \, 100$, one can appreciate that smaller sampling variances result in larger I^2 values. Therefore two meta-analyses with the same amount of heterogeneity but different within-study sample sizes will have different I^2 values; such that, the meta-analysis with a greater number of participants in each study will have a higher I^2 value (Rücker et al., 2008). That said, the I^2 metric addresses several shortcomings of the Q-statistic in that the I^2 metric is relatively independent of the number of participating studies in a meta-analysis. Therefore, I^2 values can be compared across meta-analyses with different numbers of studies. Furthermore, the I^2 metric informs on the relative quantity of heterogeneity in a meta-analysis rather than just indicating its presence and statistical significance (Higgins and Thompson, 2002).

Limitations I^2 metric: Given that I^2 values increase with the precision of studies in a meta-analysis. It is important to note that the I^2 metric is a relative measure that quantifies the impact of heterogeneity rather than actual amount of heterogeneity in a meta-analysis. More specifically,

the I^2 metric measures the proportion of the total observed variation that is due to heterogeneity. As described previously in section 1.2.3 the total observed variation is defined as the sum of sampling variance and heterogeneity. Consequently, if we had two genetic variants, namely variant A ($I^2 = 25\%$) and variant B ($I^2 = 75\%$) we would not be able to assert that variant B is more heterogeneous than variant A on the basis of I^2 alone, unless the two variants had the same amount of total observed variation in the study effects. Thus, I^2 values are just an indicator of what the fraction of observed total variance would be in the absence of sampling variance, they (I^2 values) do not inform on the extent of heterogeneity in a meta-analysis (Borenstein et al., 2017); direct comparisons of the amount of heterogeneity in meta-analyses can be made using the τ^2-statistic (Higgins and Thompson, 2002).

1.2.4.3 Quantifying the extent of heterogeneity in meta-analyses

τ^2-statistic: The τ^2-statistic measures the amount of dispersion in the underlying effects of studies included in a meta-analysis. τ^2 values are typically estimated as part of a random-effects meta-analysis to account for the amount of marker-specific heterogeneity. τ^2 values can be estimated using a variety of approaches including through: a restricted maximum likelihood estimate (REML) (Harville, 1977) and a moment-based estimate (DerSimonian and Laird, 1986). It should be noted that estimates of heterogeneity generated by different methods might not always be congruent (Langan et al., 2017); as such, when comparing τ^2- values across meta-analyses it might be useful to establish what algorithms were used to estimate the extent of heterogeneity and how their properties differ. Moment-based estimates of between-study variance, τ^2 can be obtained by equating the expected value of the Q-statistic to its observed value (Hardy and Thompson, 1998). For a meta-analysis with S contributing studies, the expectation of the Q-statistic is given by

$$E(Q) = (S-1) + \tau^2 \left(\sum_{s=1}^{S} w_s - \frac{\sum_{s=1}^{S} w_s^2}{\sum_{s=1}^{S} w_s} \right).$$

Then equating the Q-statistic with its expected value and rearranging terms,

$$Q = (S-1) + \tau^2 \left(\sum_{s=1}^{S} w_s - \frac{\sum_{s=1}^{S} w_s^2}{\sum_{s=1}^{S} w_s} \right),$$

$$Q - (S-1) = \tau^2 \left(\sum_{s=1}^{S} w_s - \frac{\sum_{s=1}^{S} w_s^2}{\sum_{s=1}^{S} w_s} \right),$$

$$\tau^2 = \frac{Q - (S-1)}{\left(\sum_{s=1}^{S} w_s - \frac{\sum_{s=1}^{S} w_s^2}{\sum_{s=1}^{S} w_s}\right)},$$

where

Q: Q-statistic denoting the total observed variation in the study effects

S: number of studies in the meta-analysis

w_s: weight assigned to the s^{th} study.

τ^2- values are expressed on the same scale as the effect measure used to conduct the meta-analysis (e.g. log odds ratios). This has the practical implication of limiting the comparability of τ^2- values across meta-analyses based on different effect measures (Rücker et al., 2008). In contrast to I^2, H^2 and the Q-statistic, τ^2- estimates do not increase with the precision of the studies in a meta-analysis (Higgins and Thompson, 2002; Rücker et al., 2008); as a result, τ^2- statistics are regarded as the "gold" standard for measuring the magnitude of heterogeneity in a meta-analysis (Higgins and Thompson, 2002). Similar to the I^2 and H^2 metrics, τ^2- values can be compared across meta-analyses with different numbers of studies; this is because τ^2- values do not systematically increase with the number of studies in a meta-analysis.

Limitations: Point estimates of τ^2 are generally imprecise when meta-analyses have few studies (e.g. < 10 studies) (Higgins et al., 2009; Pullenayegum, 2011). In particular, the DerSimonian and Laird τ^2- estimate has been shown on average to overestimate the amount of heterogeneity in meta-analyses when there are small numbers of studies (Borenstein et al., 2010); and to

underestimate it (heterogeneity) in meta-analyses where studies have small sample sizes, particularly when the level of heterogeneity is moderate-to-high. Restricted maximum likelihood estimates also show a negative bias when the level of heterogeneity is moderate-to-high, though to a lesser degree than DerSimonian and Laird τ^2- estimates (Langan et al., 2017). The practical implication here is that imprecise τ^2- estimates can contribute to inaccurate summary effect estimates and confidence intervals.

1.2.4.4 Explaining heterogeneity in meta-analyses

Random-effects meta-analysis accounts for the variation in the underlying effects of studies in a meta-analysis but does not explain why they differ. Meta-regression can be used to investigate whether differences among the characteristics of studies in a meta-analysis can explain some of the observed variability in the study effects; the goal being to identify study characteristics that contribute to substantially smaller or larger observed effects. Meta-regression can be carried out both under fixed-effect and random-effects models. When conducted under a random-effects model, meta-regression explores the contribution of known factors (i.e. recorded study characteristics) to heterogeneity whilst accounting for the residual variation that could not be explained by covariates included in the meta-regression. Essentially, random-effects meta-regression is an extension of random-effects meta-analysis that incorporates study characteristics as explanatory variables (Harbord and Higgins, 2008). Assuming that the underlying effects in a random-effects meta-regression are normally distributed around a linear predictor, $\theta_1, \theta_2, \theta_3, \cdots, \theta_s \sim N(\mathrm{x}_s\beta, \tau^2)$ the observed effect for the s^{th} study in a meta-regression with S studies is given by,

$$y_s \sim N(\mathrm{x}_s\beta, \tau^2 + \sigma_s^2)$$

or equivalently

$$y_s = \mathrm{x}_s\beta + \epsilon_s + u_s, \text{ where } \epsilon_s \sim N(0, \sigma_s^2) \text{ and } u_s \sim N(0, \tau^2),$$

where,

y_s: observed effect for the s^{th} study

β: a k x 1 vector of coefficients which may include a constant if fitted

x_s: a 1 x k vector of covariate values in the s^{th} study which includes a 1 if a constant is fitted

τ^2: amount of heterogeneity

σ_s^2: sampling variance for the s^{th} study

ϵ_s: error contribution of the sampling variance to the observed effect in the s^{th} study

u_s: error contribution of the between-study variance to the observed effect in the s^{th} study.

In contrast to sub-group analysis, which explores the contribution of each covariate to heterogeneity separately, meta-regression allows one to simultaneously compare the contributions of different covariates to heterogeneity. Ideally, study characteristics to be explored in a meta-regression should be decided from the outset on the basis of study design features and biological factors likely to induce variability in the observed study effects, this reduces the burden of multiple testing and the potential for false positives arising from data dredging (Thompson and Higgins, 2002). Similar to simple regression, meta-regression tests for a linear relationship between the outcome variable of a meta-regression, that is the effect measure and included explanatory variables (e.g. study characteristics). The coefficients from a meta-regression describe the change in the outcome variable for every unit increase in the explanatory variables (i.e. predictors). To determine the proportion of heterogeneity in a meta-regression explained by the study characteristics of interest, τ^2 estimates in the empty model (i.e. no covariates) are compared with those in models where covariates have been included. The relative decrease in heterogeneity owing to the explanatory variables is then given by the adjusted-R^2 term,

$$R_{adj}^2 = \frac{\hat{\tau}^2 - \hat{\tau}_0^2}{\hat{\tau}_0^2} ,$$

where,

τ^2: estimate of the amount of heterogeneity in a model with covariates

τ_0^2: estimate of the amount of heterogeneity in model without covariates

R^2_{adj}: proportion of variability in the observed study effects owing to heterogeneity that is explained by the predictors.

When the predictors in a meta-regression model explain less heterogeneity than would be expected by chance adjusted-R^2 values can be negative, especially if the meta-regression only comprises a few studies (Harbord and Higgins, 2008).

The power of a random-effects meta-regression to detect statistically significant relationships between multiple explanatory variables and the outcome variable depends on the number of studies included in a meta-regression rather than the within-study sample size (Thompson and Higgins, 2002). A useful informal guideline for the ideal number of subjects in a multiple regression is to have at least ten subjects per covariate. Borenstein and colleagues suggest extending this guideline to meta-regression to have at least ten studies per covariate (Borenstein et al., 2009). To assess the statistical significance of a model with multiple covariates (i.e. a joint model), a multi-parameter Wald test is used to evaluate the null hypothesis that none of the covariates are associated with the outcome, that is their coefficients are all zero. A significant p-value indicates that at least one of the covariates is associated with the outcome (Harbord and Higgins, 2008).

Limitations: Meta-regression is typically applied to elucidate the independent contributions of study characteristics to heterogeneity. However, study characteristics under investigation in a meta-regression may not only be correlated with the size of the summary effect estimate, but they may also be substantially correlated with each other. Confounding can distort relationships between predictors and the outcome variable, such that an explanatory variable that might not

have been statistically significant on its own shows a significant relationship with the outcome variable in the presence of correlated explanatory variables. Extreme versions of confounding e.g. multicollinearity where explanatory variables may be highly correlated, can lead to inaccurate estimates of meta-regression coefficients and inflated standard errors; as a result, confidence intervals can be wider than expected (compared with if the explanatory variables were independent) and covariates that would otherwise be statistically significant might no longer show a relationship with the outcome variable. Model selection can be employed to reduce the chances of false conclusions due to confounding (Thompson and Higgins, 2002).

1.3 Summary

To summarise, GWAS meta-analysis affords complex trait geneticists the opportunity to maximize sample size and thereby optimize power for the validation and discovery of small effect loci. Dissecting heterogeneity in GWAS meta-analysis is an important consideration both from a quality control perspective and from the vantage point of gaining better insight into the genetic architecture of complex traits. Excessive heterogeneity arising from quality control issues including: inconsistent genotype call rates, variable imputation quality, incorrect minor allele frequencies and flipped alleles could potentially reduce the power of a meta-analysis to detect genetic associations. However, moderate levels of heterogeneity can signal biologically relevant differences among studies in a meta-analysis such as: sub-phenotype variability, epistasis, sexual dimorphism, and gene-environment interactions. Heterogeneity in genetic association meta-analyses is primarily examined on an individual variant basis using single-variant methods (e.g. Q, I^2 and τ^2). The ability to investigate the relative impact of heterogeneity sources across multiple variants (i.e. systematic heterogeneity patterns) has been restricted by the absence of appropriate methodology.

1.4 General aims and outline

General Aims

The overarching goal of this thesis was to explore methods for dissecting heterogeneity in genetic association meta-analyses, in particular to test three hypotheses. First, given that genetic association meta-analyses typically examine millions of variants at a time, I hypothesised that beyond the heterogeneity patterns that occur at individual variants (i.e. marker-specific heterogeneity) there might be heterogeneity patterns that arise across multiple genetic signals in a meta-analysis (i.e. genome-wide or systematic heterogeneity patterns). To test this hypothesis, the first aim of this thesis was to develop a novel multi-variant statistic, *M* that could be used to search for systematic heterogeneity patterns in meta-analyses and identify differences in the design features or ascertainment criteria of participating studies that might explain them. Second, given that meta-analyses comprising smaller numbers of studies (e.g. $\leq$ 10 studies) generally provide imprecise estimates (Thompson and Sharp, 1999), I hypothesised that Bayesian genetic association meta-analyses informed with GWAS-derived empirical heterogeneity priors might provide estimates of average genetic effects and heterogeneity that were more precise than those afforded by frequentist methods, particularly in meta-analyses with few studies. Accordingly, the second aim of this thesis was to develop GWAS-derived empirical heterogeneity priors as an alternative method for refining estimates of average genetic effects and heterogeneity at variants of interest in genetic association meta-analyses based on smaller numbers of studies (e.g. $\leq$ 10 studies). Third, with the availability of contemporary random-effects methods (e.g. Han-Eskin random-effects model) that potentially have greater power than traditional approaches to

identify genetic associations in the presence of heterogeneity, I hypothesised that the Han-Eskin random-effects model might be able to identify small effect heterogeneous loci overlooked by traditional fixed and random-effects meta-analysis methods. Thus the third aim of this thesis was to conduct a critical evaluation of the Han-Eskin random-effects model in relation to traditional meta-analysis methods and apply it to a recent meta-analysis of CAD risk to identify small effect heterogeneous loci that might have eluded detection by traditional fixed and random-effects methods. Below is an outline of the topics addressed in the thesis.

Outline

The topics presented in this thesis are organised as follows. Following a brief historical note and introduction to the exploration of heterogeneity in genetic association meta-analyses (chapter 1), a novel multi-variant heterogeneity statistic, *M* is presented and discussed in chapter 2. Next the focus turns towards the development of methods to improve precision in estimation of average genetic effects and heterogeneity in smaller meta-analyses (chapter 3). Chapter 4, provides a critical evaluation of the Han-Eskin random-effects method with a view to identify small effect heterogeneous loci missed by conventional fixed and random-effects meta-analysis methods. To round out this thesis' dissection of heterogeneity in GWAS meta-analysis, chapter 5 provides a unifying discussion that reflects on the topics presented in chapters 1 – 4.

Approximate number of words in thesis: 45,000 words.

This page is intentionally left blank.

Chapter 2

This chapter proposes a novel multi-variant heterogeneity statistic, *M,* to identify systematic heterogeneity patterns and their direction of effect in genetic association meta-analyses.

2.1 Hypothesis / Research needs

Efforts to explore heterogeneity in GWAS meta-analysis have mainly concentrated on the random dispersion of genetic effects at individual variants (i.e. marker-specific heterogeneity), leaving the contribution of genome-wide or systematic heterogeneity patterns largely unexplored. Commenting on measures of marker-specific heterogeneity (Q, H^2, I^2) in a paper discussing statistical considerations for GWAS meta-analysis Begum and colleagues (Begum et al., 2012) note:

> These statistics are calculated separately for each SNP, which leads to the interesting and unsolved question of whether or how one should make a genome-wide determination of heterogeneity. (Begum et al, 2012: 3779)

As complex trait geneticists endeavour to identify variants conferring increasingly smaller effects with the help of larger international meta-analyses, my hypothesis is that the diverse ancestries, design features and recruitment criteria of participating studies might induce genome-wide (systematic) heterogeneity patterns. Systematic heterogeneity patterns in a GWAS meta-analysis may manifest as outlier studies showing consistently stronger or weaker genetic effects than average across the array of variants examined. The practical advantages of investigating

systematic heterogeneity patterns in GWAS meta-analyses are two-fold. First, the identification of systematic heterogeneity patterns provides an opportunity to do across-study quality control assessments in large-scale GWAS meta-analyses based on summary data, such as those from the CARDIoGRAMplusC4D (CARDIoGRAMplusC4D Consortium, 2015), METASTROKE (Traylor et al., 2012) and GIANT (Wood et al., 2014) consortia. This is important because meta-analyses based on summary data typically rely on participating research groups to conduct quality control checks (e.g. genotype call rate, Hardy-Weinberg equilibrium cut-offs, imputation quality, genomic control) on their own datasets in accordance with a pre-agreed analysis plan. However, participating research groups generally follow the pre-agreed analysis plan to varying degrees; therefore systematic differences arising from variable levels of stringency in the quality control checks of individual studies could potentially lower the power of a meta-analysis to detect genetic signals. Besides quality control issues, systematic differences among studies in a meta-analysis that may affect a large proportion of the variants examined include: ancestry, LD levels, participant recruitment criteria (e.g. family-history, age-of-disease onset), lifestyle and environmental factors.

Furthermore, systematic heterogeneity patterns can reveal "noisy" studies in a meta-analysis, which might dampen the power to detect genetic signals. Noisy studies are typically outliers that show no apparent association with known susceptibility variants of the phenotype of interest, owing to faulty genotype data or incorrect phenotypes. Studies with faulty genotype data may have incorrect minor allele frequencies, flipped alleles or strands. The inclusion of studies with incorrect phenotypes is usually a consequence of phenotype mix-ups, where research groups erroneously submit study results for a different phenotype from the one under investigation.

Second, systematic heterogeneity patterns can elucidate important biological differences among studies in a meta-analysis by revealing sexual dimorphism, differential association by ancestry and sub-phenotype variability. Systematic heterogeneity patterns can also highlight genetically enriched studies in a meta-analysis, which show systematically stronger effects than average; these factors can then be included in the ascertainment criteria when planning new meta-analyses so as to optimize power to detect genetic signals.

Heterogeneity measures originally designed to examine marker-specific heterogeneity at individual variants such as the Q-statistic, can be exploited to explore systematic heterogeneity patterns in meta-analyses. However, my hypothesis is that the burden of multiple testing for multiple variants might render this approach sub-optimal (in terms of power) for investigating systematic heterogeneity; especially since it has been noted that the Q-statistic's power to detect heterogeneity degrades when adjusted for multiple variants (Zeggini and Ioannidis, 2009). As such, to facilitate the efficient identification of systematic heterogeneity patterns in GWAS meta-analyses, an easily interpretable multi-variant heterogeneity statistic, *M*, is proposed which consolidates heterogeneity information from multiple independently associated variants on the extent and direction in which their study effect-size estimates deviate from the average genetic effect. Once systematic heterogeneity patterns have been identified, a meta-regression of the *M* statistics can be performed to investigate the underlying sources of systematic heterogeneity. An important advantage and unique feature of the new multi-variant heterogeneity statistic is that it diagnoses systematic heterogeneity patterns simultaneously across all participating studies in a meta-analysis and reveals their direction of effect. Thus for an array of lead variants selected for heterogeneity analysis, the *M* statistic can be used to delineate between study characteristics in a

meta-analysis that contribute to systematically stronger genetic effects from those that might lower the power of the meta-analysis to replicate genetic associations and discover new genetic signals. This is in contrast to the λ_{meta} heterogeneity statistic (Chen et al., 2016), which examines heterogeneity on a pairwise study basis and does not convey the direction of systematic heterogeneity patterns. To assess the validity and power of the *M* statistic, and to demonstrate its application, this chapter focuses on three specific aims:

Aim 1: Examine the *M* statistic's null distribution, type-1 and 2 error rates using Monte-Carlo simulation studies.

Aim 2: Apply the *M* statistic to the CARDIoGRAMplusC4D GWAS meta-analysis of CAD risk to demonstrate its utility in identifying systematic heterogeneity patterns.

Aim3: Investigate factors underlying systematic heterogeneity in the CARDIoGRAMplusC4D GWAS meta-analysis.

2.2 Methods

2.2.1 Identifying systematic heterogeneity patterns in GWAS meta-analysis

Computing standardized predicted random effects: Standardized predicted random effects (*SPRE*s) are precision weighted residuals, which convey the direction and extent to which observed study effects in a meta-analysis deviate from the estimated summary effect. *SPRE*s are useful for identifying outlying studies at individual variants in a meta-analysis.

To calculate *SPRE*s, an inverse-variance weighted random-effects meta-analysis of the genetic variant of interest is performed. This is done to obtain estimates for the summary effect and the amount of marker-specific heterogeneity in the meta-analysis. Consider for example a meta-analysis of CAD risk comprising *S* studies, performed to identify outlier studies contributing to heterogeneity at the rs2891168 locus on chromosome 9. Where the observed study effects are log-transformed regression coefficients from a logistic regression (e.g. per-allele log odds ratios and their accompanying standard errors). An alternative scenario could be a meta-analysis to investigate heterogeneity at a locus (e.g. *LYPLAL1*) associated with waist-hip ratio (Heid et al., 2010), where the study effect-size estimates are regression coefficients from a linear regression. Similar to the random-effects model outlined in chapter 1 (section 1.2.3), the random-effects meta-analysis is performed under the assumption that the observed study effects are normally distributed. Additionally, the sampling variances of the participating studies in the meta-analysis are treated as known when in fact they have been estimated from the data in each

study (Harbord and Higgins, 2008). As per a standard inverse-variance weighted random-effects meta-analysis, the weight assigned to each study is the reciprocal of its total observed variance such that the weight allocated to the S^{th} study is

$$w_s = \frac{1}{(\sigma_s^2 + \hat{\tau}^2)} ,$$

where,

w_s: weight for the s^{th} study

σ_s^2: sampling variance for the s^{th} study

$\hat{\tau}^2$: estimate of the amount of marker-specific heterogeneity in the meta-analysis.

The amount of heterogeneity at the variant of interest, τ^2 can be estimated using a variety of methods including the method-of-moments (DerSimonian and Laird, 1986) and restricted maximum likelihood (Harville, 1977) algorithms mentioned in chapter 1 (section 1.2.4.3). Subsequently, the average genetic effect for the variant of interest is estimated using an inverse variance-weighted least squares regression of the observed study effects and is denoted by,

$$\hat{\theta} = \frac{\sum_{s=1}^{S} w_s y_s}{\sum_{s=1}^{S} w_s}.$$

Then *SPRE*s can be calculated for each participating study in the meta-analysis such that the *SPRE* statistic for the S^{th} study is given by,

$$SPRE_s = \frac{(y_s - \hat{\theta})}{\sqrt{Var(y_s - \hat{\theta})}} ,$$

or equivalently,

$$SPRE_s = \frac{(y_s - \hat{\theta})}{\sqrt{\sigma_s^2 + \hat{\tau}^2 - s_{p_s}^2}} \ ,$$

where,

y_s : observed genetic effect in the s^{th} study

$\hat{\theta}$: average effect estimate (i.e. prediction excluding random effects)

σ_s^2 : sampling variance in the s^{th} study

$\hat{\tau}^2$: estimate of the extent of marker-specific heterogeneity in the meta-analysis

s_{p_s}: standard error of prediction (fitted values) in the s^{th} study excluding random effects

such that,

$(y_s - \hat{\theta})$ represents the raw residual and $\sqrt{\sigma_s^2 + \hat{\tau}^2 - s_{p_s}^2}$ the unconditional standard error of the predicted random effect (Harbord and Higgins, 2008).

SPRE statistics are essentially *Z*-scores or standard normally distributed random variables, $SPRE \sim N(0,1)$. Once calculated, *SPRE* statistics of the contributing studies in a meta-analysis can be plotted against the quantiles of a standard normal distribution in a normal probability plot to identify the outlying studies (Figure 2.1). Notably, in the absence of marker-specific heterogeneity ($\tau^2 = 0$), *SPREs* are equivalent to the standardized predicted fixed effects that would be obtained from a fixed-effect meta-analysis.

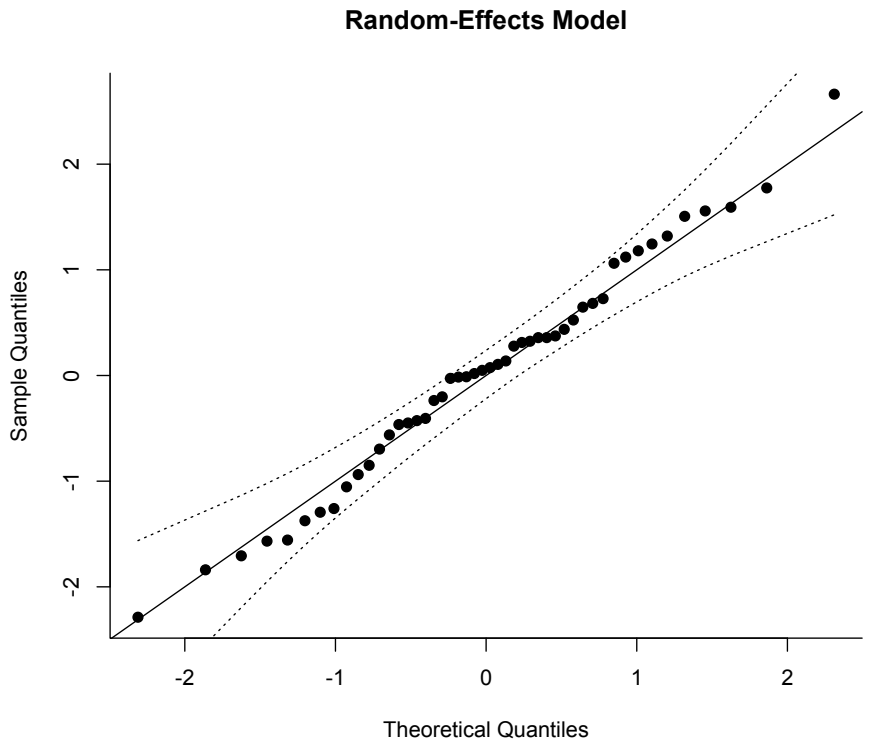


Figure 2.1: A normal probability plot of *SPRE* statistics at the 9p21 coronary artery disease locus (rs2891168) on chromosome 9. Source: Lerato Magosi (Farrall and Hopewell labs).

A novel multi-variant heterogeneity statistic, M: Now consider a meta-analysis project, P comprising a series of S genetic association studies where the objective is to identify systematic heterogeneity patterns across a panel of V variants. The panel of variants may be a list of variants known to be associated with the trait of interest or the GWAS-significant lead variants from a meta-analysis. The variants in the panel are assumed to be independently associated with the trait under study (i.e. in linkage equilibrium with each other). To facilitate the search for systematic heterogeneity patterns effect-size estimates of variants selected for heterogeneity analysis are judiciously "flipped" to align the effect alleles, and orient the average effect estimates in the same direction such that the same allele is compared across the participating studies. Then an inverse-variance weighted random-effects meta-analysis can be performed for each of the V variants. The resulting average genetic effect and heterogeneity estimates can then be used to calculate *SPRE* statistics for every variant in each study affording a regular array of *SPREs*

$$P_{S,V} = \begin{bmatrix} SPRE_{1,1} & SPRE_{1,2} & \cdots & SPRE_{1,v} \\ SPRE_{2,1} & SPRE_{2,2} & \cdots & SPRE_{2,v} \\ \vdots & \vdots & \ddots & \vdots \\ SPRE_{s,1} & SPRE_{s,2} & \cdots & SPRE_{s,v} \end{bmatrix},$$

where the *SPRE* for the v^{th} variant in the s^{th} study is given by,

$$SPRE_{sv} = \frac{y_{sv} - \hat{\theta}_v}{\sqrt{\sigma_{sv}^2 + \tau_v^2 - s_{p_{sv}}^2}},$$

and,

y_{sv} : observed genetic effect of the v^{th} variant in the s^{th} study

$\hat{\theta}_v$: average effect estimate of the v^{th} variant (i.e. prediction excluding random effects)

σ_{sv}^2 : sampling variance for the v^{th} variant in the s^{th} study

$\hat{\tau}_v^2$: estimate of the extent of heterogeneity at the v^{th} variant

$s_{p_{sv}}$: standard error of prediction (fitted values) for the v^{th} variant in the s^{th} study excluding random effects.

To identify outlier studies showing systematic heterogeneity patterns in the meta-analysis, a mean aggregate statistic (*M*) can be calculated that summarises heterogeneity information across the multiple variants independently associated with the trait of interest. *M* statistics are calculated for each contributing study in the meta-analysis by taking the arithmetic mean of *SPREs* statistics in each study, such that the *M* statistic of the s^{th} study is defined as,

$$M_s = \frac{1}{V}\sum_{v=1}^{V} SPRE_{sv} \ .$$

Assuming that *M* statistics are linear combinations of precision-weighted mutually independent standard normal random variables, $SPRE \sim \Phi(0,1)$ then *M* statistics will be normally distributed under the theorem of sums of independent normals, $M_s \sim \Phi\left(0, \frac{1}{V}\right)$ with expectation $E(M_s) = V \times \left(\frac{1}{V}\right) \times \mu = 0$ and variance $Var(M_s) = V \times (\frac{1}{V})^2 \times \sigma^2 = (\frac{1}{V})1^2 = \frac{1}{V}$. Studies showing systematically larger than average genetic effects are denoted by positive *M* values whilst those showing consistently smaller effects are assigned negative *M* values. Studies showing statistically significant systematic heterogeneity patterns are ones where the absolute

value of the *M* statistic, $|M_s|$ exceeds a critical *M* value threshold, corrected for multiple testing of *S* studies at the 5% significant level (Bonferroni corrected family-wise error rate < 0.05).

Treating the *V* variants selected for heterogeneity analysis as a sample of the total population of variants associated with the trait of interest, then the estimate of the population mean is given by $\overline{M}_s$ and standard error,

$$SE_{\overline{M}_s} = \left(\frac{Var(SPRE_s)}{V}\right)^{1/2}.$$

In scenarios where participating studies in the meta-analysis may have missing variants, the *SPREs* for those variants can be substituted by assigning each missing variant the expected mean $E(M_s) = V \times \left(\frac{1}{V}\right) \times 0 = 0$, of the theoretical distribution of the *M* statistic for *V* variants under the null (Ho).

Software to calculate *M* statistics has been made available as a Stata® command (**getmstatistic**) accessible through the SSC (Boston college statistical software components) archive and an R® package (**getmstatistic**) which can be downloaded from CRAN (comprehensive R archive network). The getmstatistic command is compatible with Stata® ≥ 10.1 and R® ≥ 3.1.0.

2.2.2 Empirically evaluating the *M* statistic's null distribution

Monte-Carlo simulation experiments comprising 10,000 replicates each were conducted to generate empirical distributions of the *M* statistic, for meta-analyses scenarios with different numbers of variants (2, 10, 25 or 50). Proceeding with the assumption that *M* statistics are linear combinations of precision-weighted standard normal random variates i.e. *SPRE* statistics where $SPRE \sim N(0,1)$; *p*-values were sampled from a uniform distribution, $U[0,1)$ and subsequently converted to z-scores (*SPRE* statistics) by taking the inverse of the standard normal cumulative distribution function. Z-scores for different numbers of variants (2, 10, 25 or 50) were then aggregated, by calculating their arithmetic mean to yield *M* statistics.

The requisite number of replicates for the simulation experiments was determined using a back-of-the-envelope calculation, which evaluates the number of trials (replicates) needed to obtain a desired level of precision for Binomial confidence intervals at the 5% significance level. For example, the width of the 95% confidence intervals for the Binomial probability of success over a 1,000 replicates would be, [CI: 0.0373, 0.0654] with se: 0.00689 and mean: 0.05. Comparatively, the 95% Binomial confidence interval widths at 10,000 and 100,000 replicates would be [CI: 0.0458, 0.0545] with se: 0.00218 and mean: 0.05; and [CI: 0.0487, 0.0514] with se: 0.000689 and mean: 0.05 respectively. Here, mean and se refer to the expected fraction of successes and the corresponding standard error. Striking a balance between the intensity of the computational requirements and the precision of the 95% Binomial confidence intervals, 10,000 replicates was chosen as a reasonable number of replicates. Calculations were performed using the cii command in Stata® version 10.1.

2.2.3 A comparative analytical power analysis of *M* and the Q-statistic to identify outlier studies showing systematic heterogeneity in GWAS meta-analysis

Analytic power calculations were performed comparing the power of the *M* and Q statistics to diagnose systematic heterogeneity in GWAS meta-analysis scenarios with varying numbers of studies (10, 15 or 30) and variants (10, 25 or 50). In the absence of systematic heterogeneity, Q-statistics approximate a chi-squared distribution with *S-1* degrees of freedom (where *S* is the number of studies), when the number and size of studies in a meta-analysis are large, whilst *M* statistics are normally distributed with an expectation of 0 and variance $1/V$. Therefore, *M* statistics were expressed in a chi-squared form to simplify the comparison of the two statistics. In particular, *M* statistics were standardized to have a variance of 1 and squared to yield Wald statistics of the form, $M^2 = (\frac{M}{SE_M})^2 = \chi_1^2$, where $SE_M = (\frac{1}{V})^{1/2}$ (Sham and Purcell, 2014). In each meta-analysis scenario, Q-statistics were calculated for each variant and *M* statistics computed for each study.

Q-statistic analytic power calculations: The analytic power of Q to detect heterogeneity at a single variant (e.g. the v^{th} variant) is given by standard chi-squared power calculations as, $\omega_v = 1 - nchi2(S-1, Q_{ncp}, Q_{critical})$ (Agresti, 1996). Here, $nchi2$ denotes the cumulative non-central chi-squared distribution function (Stata® 10.1) and *S-1* represents the degrees of freedom (where *S* is the number of studies). Treating Q as a log likelihood ratio statistic (Agresti, 1996; Liu, 1997), allows the observed Q-statistic value for the v^{th} variant to be used as a substitute for the non-centrality parameter, Q_{ncp}. $Q_{critical}$ is the critical Q-statistic value from a

central chi-square distribution with *S -1* degrees of freedom, adjusted for multiple testing of *V* variants at the 5% significance level ($\alpha = 0.05$), $invchi2(S-1, 1-(0.05/V))$. Then, the overall power of Q to detect heterogeneity across multiple variants, ω is given by the probability that at least one variant shows heterogeneity. This can be equivalently stated as 1 – probability that none of the variants show heterogeneity and is given by,

$$\omega = 1 - \prod_{1}^{V} 1 - \omega_v \ .$$

***M*-statistic analytic power calculations:** The analytical power of *M* to detect heterogeneity shown by the s^{th} study in a meta-analysis comprising *S* studies is given by standard chi-squared power calculations of the form, $\omega_s = 1 - nchi2\left(df, M^2_{ncp}, M^2_{critical}\right)$ (Agresti, 1996). Here, df denotes 1 degree of freedom since squared-standardized *M* statistics (i.e. M^2), are chi-squared statistics with a single degree of freedom. M^2_{ncp} is the observed M^2 statistic value for s^{th} study used as a substitute for the non-centrality parameter. $M^2_{critical,}$ is the critical M^2 value from a central chi-squared distribution with a single degree of freedom, corrected for multiple testing of *S* studies at the 5% significance level $(\alpha = 0.05)$ and is given by, $invchi2\left(1, 1-(0.05/S)\right)$. Then, the overall power of *M* to detect heterogeneity at one or more variants, ω is given by,

$$\omega = 1 - \prod_{1}^{S} 1 - \omega_s.$$

The analytical power calculations were performed using Stata® version 10.1.

2.2.4 Assessing *M*'s power to identify outlier studies showing systematic heterogeneity patterns over a background of substantial marker-specific (random) heterogeneity

M's power to diagnose systematic heterogeneity across multiple variants over substantial background random heterogeneity at individual variants was evaluated using GWAS meta-analysis simulations with varying numbers of studies (10, 15 or 30) and variants (10, 25 or 50), each based on 1,000 replicates. The type-1 and 2 error rates of the *M* statistic were given by empirical probabilities, $(r + 1/n + 1)$ as the fraction of Monte-Carlo replicates $(r + 1)$, which had at least one study with an *M* statistic value as or more extreme than the critical *M* value, where n denotes the total number of replicates in each Monte-Carlo simulation experiment (i.e. 1,000 replicates) (Davison and Hinkley, 1997; North et al., 2003). The critical *M* value for a meta-analysis scenario with *V* variants, allowing for multiple testing of *S* studies at the 5% significance level is, $M_{critical} = M_{zscore}\ x\ M_{expected\ SD} + M_{expected\ Mean}$, and can be computed in **R®** (≥ 3.1.0) as,

$$M_{critical} = \mathrm{abs}\big(\mathrm{qnorm}(\,0.05/S/2)\big)\ x\ \mathrm{sqrt}\left(\frac{1}{V}\right) + V\ x\left(\frac{1}{V}\right) x\ 0\ \ .$$

2.2.5 Explaining heterogeneity in the CARDIoGRAMplusC4D meta-analysis

Investigating sources of systematic heterogeneity: Sources of systematic heterogeneity in the CARDIoGRAMplusC4D meta-analysis were explored by an inverse-variance weighted random-effects meta-regression of the *M* statistics employing design features (e.g. family-history, ancestry and genotype call rate) that were likely to vary from study-to-study as covariates. The meta-regression analysis of *M* statistics was performed using Harbord and Higgin's metareg command in Stata® version 10.1 (Harbord and Higgins, 2008), which firstly computes a restricted maximum likelihood (default algorithm) estimate of the amount of heterogeneity, τ^2 and subsequently estimates the regression coefficients, β by a weighted least squares regression, where the weights are given by $1/(SE_{M_s}{}^2 + \hat{\tau}^2)$. Notably, in this extended meta-regression framework, $SE_{M_s}{}^2$ denotes the sampling variance of the *M* statistic for the s^{th} study.

Investigating sources of random heterogeneity: A classical meta-regression of observed study effects (i.e. per allele odds ratios) was performed to explore study characteristics contributing to variable effect-size estimates among participating studies in the CARDIoGRAMplusC4D meta-analysis, at variants selected for heterogeneity analysis. The meta-regression analysis was performed using Harbord and Higgin's metareg command in Stata® version 10.1 (Harbord and Higgins, 2008).

2.3 Results

2.3.1 Aim 1: Simulation studies to examine the properties of the *M* statistic

Three sets of Monte-Carlo simulation studies were performed to assess the *M* statistic's validity (type-1 error) and power to diagnose systematic heterogeneity in GWAS meta-analysis.

Choice of numbers of studies and variants: The numbers of studies and variants used in the simulation experiments described below were chosen to: 1) reflect typical sizes of GWAS meta-analyses encountered in the literature, and 2) to consider genetic association meta-analyses of complex traits where variants may range from a few candidate genes (e.g. ≤10 variants) to a panel of independently associated loci in a genome-wide analysis (e.g. ≥ 50 variants). Variants were assumed to be in linkage equilibrium with each other in order to maximize the power of detecting systematic heterogeneity patterns; this dovetails with a strategy typically employed to optimize power when computing genetic risk scores for Mendelian randomization studies or risk prediction. However, it does not discount the fact that variants in linkage disequilibrium can be independently associated with the trait of interest (Chatterjee et al., 2016).

The first set of simulations considers the theoretical distribution of the *M* statistic. The distribution of *M* statistics in the absence of systematic heterogeneity (i.e. under the null) allows us to observe how the *M* statistic's behaviour may change due to the play of chance; and provides a basis of comparison with the distribution of *M* statistics under alternative hypotheses.

Knowing the expectation (mean) and variance parameters of *M's* theoretical distribution allows us to make inferences about systematic heterogeneity in studies with incomplete genotype data (i.e. missing genetic variants).

***M* approximates a normal distribution under the null**: The *M* statistic's theoretical distribution was evaluated empirically for different numbers of variants (2, 10, 25 or 50) through Monte-Carlo simulations, each involving 10,000 replicates. To approximate the null hypothesis (i.e. the absence of systematic heterogeneity) *SPRE* statistics for each set of variants (2, 10, 25 or 50) were randomly sampled from a standard normal distribution $\phi(0,1)$. The simulations demonstrate a close-fit between the empirical and theoretical distributions of *M* regardless of the number of variants evaluated (Figure 2.2 and Table 2.1). Suggesting that *M* will be normally distributed under the null for a variety of meta-analysis scenarios, ranging from a few loci to large panels of genetic variants. The relatively closer fit between the empirical and theoretical distributions at 50 variants versus 2 variants is to be expected, given that at 50 variants there is more information available for the computation of *M* statistics.

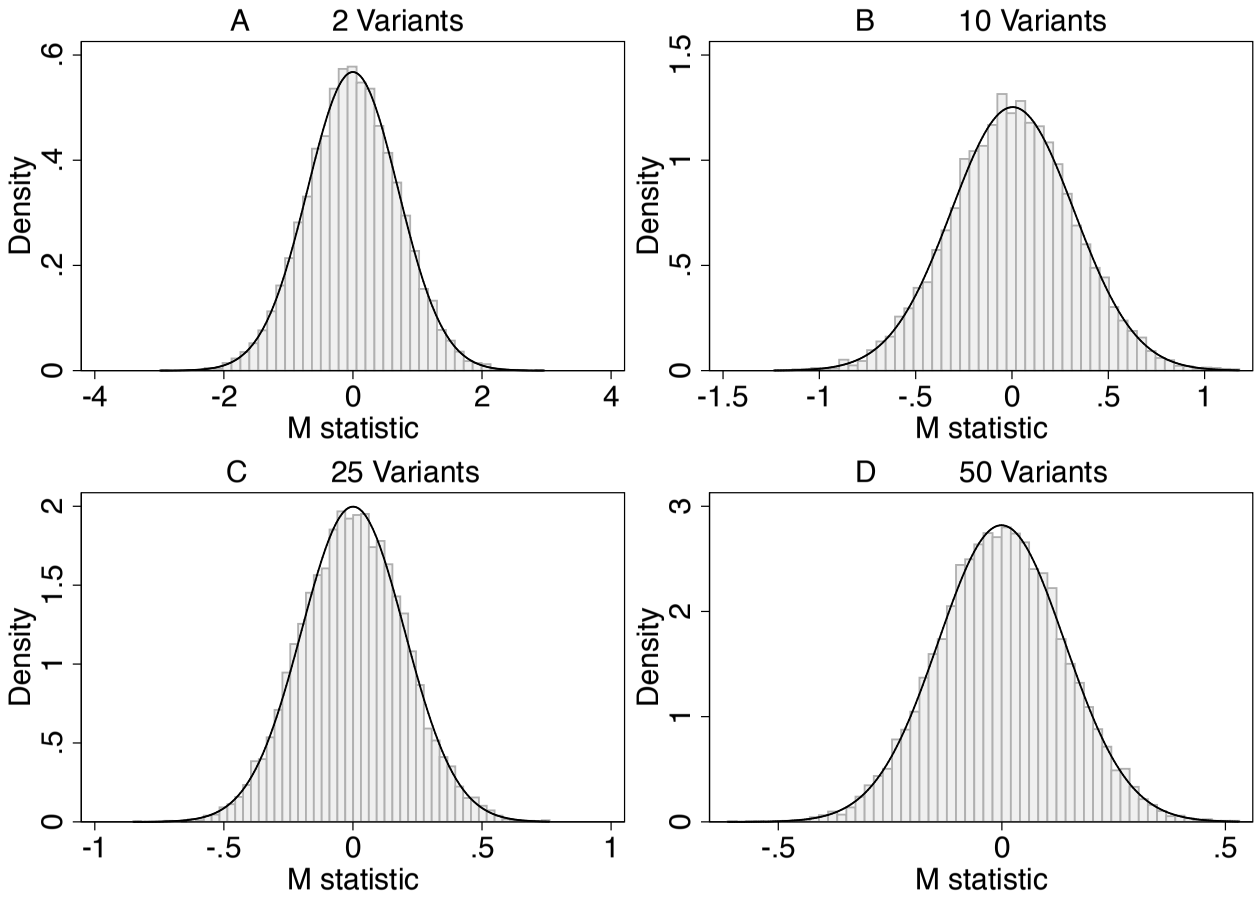


Figure 2.2: Empirical and theoretical distributions of the *M* statistic. Monte–Carlo simulations comparing the empirical (histograms) and theoretical frequency distributions (density plots) of *M* statistics. The four panels show (from A to D) simulations for 2, 10, 25 and 50 variants over 10,000 replicates. The close fit between the empirical and theoretical distributions suggest that the *M* statistic approximates a normal distribution under the null. Source: (Magosi et al., 2017)

Table 2.1: A comparison of the means and standard deviations for the theoretical and empirical null distributions of *M*

	M statistic			
# Variants	Theoretical mean	Theoretical SD	Empirical mean	Empirical SD
2	0.000	0.707	-0.001	0.703
10	0.000	0.316	0.004	0.319
25	0.000	0.200	0.001	0.200
50	0.000	0.141	-0.001	0.141

Table rows represent Monte-Carlo GWAS meta-analysis simulations each based on 10,000 replicates. Source: (Magosi et al., 2017)

The second set of simulations compares the analytic power of *M* to diagnose outlier studies showing systematic heterogeneity to that of the Q-statistic, a well-known heterogeneity metric. The Q-statistic was deemed appropriate for such a comparison as it forms the foundation for several principal measures of heterogeneity (I^2, H^2 and τ^2) in GWAS meta-analysis. A comparative power analysis affords the opportunity to judge whether there is added value in having a dedicated multi-variant statistic to diagnose systematic heterogeneity in meta-analysis or whether traditional single-variant heterogeneity statistics would suffice.

To compare *M*'s analytic power with that of the Q-statistic, GWAS meta-analysis scenarios with varying numbers of studies (10, 15 or 30) and variants (10, 25 or 50) were explored. In order to facilitate a fair comparison between *M* and the Q-statistic, family-wise error rates (FWER) for both statistics were controlled at the 5% significance level through a Bonferroni correction of multiple testing of *V* variants (for the Q-statistic) and *S* studies (for *M*). On the one hand, Q-statistics were adjusted for the number of variants since they (Q-statistics) are computed on a variant-by-variant basis across all participating studies in a meta-analysis. On the other hand, *M*-statistics were adjusted for the number of studies as they (*M*-statistics) are calculated on a study-by-study basis across all the independently associated variants in a meta-analysis. Given that the Q-statistic is better powered to detect heterogeneity when participating studies in a meta-analysis are equally precise (Hardy and Thompson, 1998), all studies were equally weighted (standard error of $\log_e$(odds ratio) = 0.08). Notably, the value for the within-study standard error was inspired by the median standard error for variants in a recent GWAS meta-analysis of CAD risk (CARDIoGRAMplusC4D Consortium, 2015). Moreover, for each of the nine meta-analysis scenarios, variants in the *S*-1 "non-outlier" studies were assigned identical effect-sizes to model homogenous effects. To induce systematic heterogeneity, effect-

sizes for variants in the outlier study were assigned the product of the "non-outlier" effect-size and a parameter (fold-change) ranging from 0 to 2.

***M* outperforms the Q-statistic in detecting systematic heterogeneity**: Figure 2.3 shows that in the absence of systematic heterogeneity (i.e. parameter fold-change = 1) or under the null, the power of both *M* and Q to diagnose systematic heterogeneity was 5% as expected. Meaning that, the proportion of outliers identified by either *M* or Q in the absence of systematic heterogeneity was consistent with that expected by chance. As the parameter fold-change differed from 1, *M* showed greater power than Q to detect systematic heterogeneity. The power to diagnose systematic heterogeneity substantially improved as the number of variants increased but weakened slightly with larger numbers of studies, potentially due to the multiple testing burden. This result is consistent with the hypothesis made in section 2.1 that Q would have lower power than *M* to diagnose systematic heterogeneity patterns in GWAS meta-analysis owing to a greater multiple testing burden. Even in scenarios where both statistics have a relatively similar multiple testing burden e.g. a meta-analysis with 10 variants and 10 studies, Q shows less power. This is consistent with the properties of the Q-statistic outlined in chapter 1 (section 1.2.4.1), where the Q-statistic is generally underpowered to detect heterogeneity when studies are few.

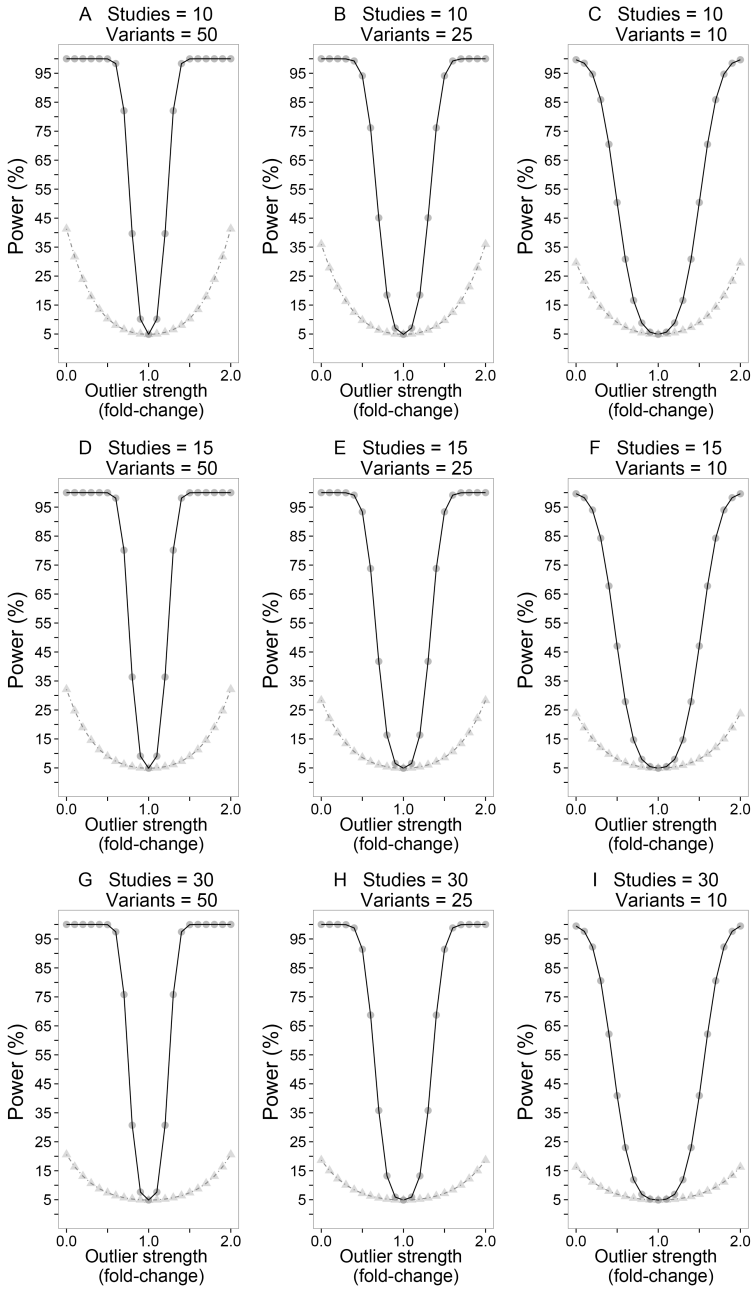


Figure 2.3: A comparative power analysis of *M* and Cochran's Q to detect systematic heterogeneity. The nine panels show (from left to right) simulations for 10, 15 and 30 studies, examined at 50, 25 and 10 variants; Data points for the *M* statistic are represented by filled circles whilst those for Cochran's Q are denoted by filled triangles. Each data point represents a meta-analysis scenario where effect-sizes for the non-outlier studies were held constant (log_e(odds ratio) = 0.182 i.e. odds ratio = 1.2) to model homogeneous effects. The effect-sizes of variants in the outlier study were the product of the non-outlier effect-size (i.e. log_e(odds ratio) = 0.182) and a parameter (fold-change) to model a continuous series of systematic heterogeneity patterns. All studies were equally weighted (standard error of log_e(odds ratio) = 0.1). In all the meta-analysis scenarios explored, *M* displayed greater power to diagnose systematic heterogeneity than Q. Source: (Magosi et al., 2017)

In "real world" meta-analysis scenarios effect-size estimates for contributing studies in a meta-analysis are seldom identical; rather, the typical distribution of common variants conferring effects for complex traits is such that the majority of variants display small effects whilst few exhibit large effects (Park et al., 2011). This "L-shape" distribution of effect-sizes can be attributed to the inverse relationship between allele frequency and effect-size, earlier postulated by Sewall Wright (Wright and Hastie, 2001; Wright, 1968) and later confirmed by Park and colleagues (Park et al., 2011). Furthermore, real world meta-analyses often have random background heterogeneity i.e. marker-specific heterogeneity. As such, the third set of simulations evaluates *M* in "real world" meta-analysis scenarios, where a background of marker-specific heterogeneity has been induced and the panel of variants being examined confers effects of varying magnitudes. Consequently, there would be two types of heterogeneity at play, marker-specific heterogeneity acting on individual variants and systematic heterogeneity acting across multiple variants. The focus of this set of simulations was to evaluate *M* on the basis of three criteria.

First, whether *M* affords the expected levels of statistical significance under the null hypothesis of no systematic heterogeneity. The expectation here is that the proportion of outliers identified by *M* in the absence of systematic heterogeneity should be similar to that expected per chance. If the proportion of outliers is considerably greater than that expected by chance, that suggests *M* is anti-conservative and vulnerable to false positives. Alternatively, if the proportion of outliers is much less than that expected per chance, that suggests *M* may be too conservative and likely to miss outlier studies showing systematic heterogeneity. Second, to assess *M*'s power to identify "noisy" outlier studies in a meta-analysis, that is, outliers showing no apparent association with known susceptibility variants for the phenotype of interest. In addition to evaluate *M*'s power to identify outlier studies showing systematically stronger or

weaker effects than other participating studies in the meta-analysis. Third, to evaluate the impact of different levels of marker-specific heterogeneity on *M*'s power to identify systematic heterogeneity patterns in a meta-analysis.

To empirically assess *M*'s type 1 and 2 error rates at the 5% significance level, GWAS meta-analysis simulations with varying numbers of studies (10, 15 or 30) and variants (10, 25 or 50) were employed, each based on 1,000 replicates (Table 2.2). In each meta-analysis scenario, variant effect-sizes (i.e. β coefficients scaled as $\log_e$(odds ratio)) were sampled from a $\Phi\ (\beta, \sigma = 0.10)$ distribution, where Φ denotes the cumulative probability distribution function of a normal distribution (Table 2.3). This was done to model effect-size distributions that would be typically found in GWAS of complex traits (Park et al., 2011), and to induce marker-specific heterogeneity. Furthermore, studies were equally weighted. In particular, the median value of standard errors for variants in a recent GWAS meta-analysis of CAD risk (CARDIoGRAMplusC4D Consortium, 2015) was used as the accompanying standard error for each simulated variant effect-size (i.e. $SE(\beta) = 0.08$). Equal weighting among the contributing studies in each meta-analysis scenario simplified computations of marker-specific heterogeneity at individual variants, such that each variant was modelled with a background random heterogeneity inconsistency index, $I^2 = 100 \times 0.10^2/(0.10^2 + 0.08^2) = 60.5\%$. In addition, a Bonferroni adjustment for multiple testing (at the 5% significance level) was applied in each meta-analysis scenario to correct for the number of studies.

Table 2.2: Empirical type-1 error rates and power of *M* to detect an outlier study showing no association with the phenotype of interest across the variants examined, at threshold α = 0.05.

# Studies	# Variants	Type- 1 error	Power (%)
30	50	0.050	98.9
	25	0.054	98.1
	10	0.056	39.9
15	50	0.033	99.2
	25	0.034	97.6
	10	0.039	40.3
10	50	0.024	98.7
	25	0.033	96.7
	10	0.025	35.7

The table shows nine GWAS meta-analysis simulation experiments each based on 1,000 Monte-Carlo replicates. Studies were equally weighted (i.e. SE(β coefficients) = 0.1) in each meta-analysis simulation. Furthermore, variant effect-sizes for studies in the type-1 error analysis were sourced from an L-shaped distribution in Table 2.3. In the power analysis, variant effect-sizes for studies showing typical effects were sourced from Table 2.3, whilst effect-sizes for variants in the outlier study were sampled from a $\Phi(\beta = 0, \sigma = 0.10)$ distribution (i.e. fold-change = zero) to model a random pattern of association. *M's* type-1 error rate is generally conservatively controlled. *M's* power to detect the outlier study increases with the number of independently associated variants. Source: (Magosi et al., 2017).

Table 2.3: Number of variants included in each effect-size bin for 3 meta-analysis simulation scenarios comprising 50, 25 and 10 variants.

	Scenario (No. Variants)		
Effect size bin[a]	50	25	10
0.04	35	13	6
0.12	8	6	2
0.2	4	3	1
0.28	2	2	1
0.4	1	1	0

Effect-sizes (β coefficients) are scaled as log_e(odds ratios)^a, β coefficients were sampled from a normal distribution with random error (SE = 0.1). Each simulation scenario generates an L -shaped distribution of effect-sizes. Source: (Magosi et al., 2017).

***M's* empirical type-1 error rate (α = 0.05) is conservatively controlled:** *M*'s type-1 error rate was generally conservatively controlled, more so when the number of studies was small (Table 2.2). The conservative behavior of the *M* statistic in meta-analysis scenarios with few studies could be a reflection of the relatively small total sample size, and in turn low power to detect systematic heterogeneity patterns over a considerable background of marker-specific heterogeneity i.e. ($I^2 = 60.5\%$) per variant. That said, the 95% binomial confidence bound for the type-1 error rate of a meta-analysis scenario comprising 10 studies and 10 variants was, 1.6% - 3.7%.

***M's* power to diagnose "noisy" outlier studies in meta-analyses depends on the number of independently associated variants.** In addition to *M*'s type-1 error rate, Table 2.2 displays results from Monte-Carlo simulations where the S^{th} study in each GWAS meta-analysis scenario was a "noisy" outlier study. Typically, "noisy" outliers would be studies that show no apparent association with the trait of interest even at known susceptibility variants. To model a "noisy" outlier study, that is a study showing a random pattern of association, the β coefficients for the V variants in the S^{th} study were sampled from a $\Phi(\beta = 0, \sigma = 0.10)$ distribution (i.e. fold-change = zero). *M*'s power to identify the outlier study increased with the number of variants examined and was relatively independent of the number of studies; suggesting that *M*'s power to diagnose "noisy" systematic outlier studies depends on the number of independently associated variants examined.

***M's* power to identify systematic outlier studies increases with the number of independently associated variants.** Figure 2.4 displays results from Monte-Carlo simulations where the S^{th} study in each GWAS meta-analysis scenario showed systematically stronger genetic

effects than the rest of the contributing studies in the meta-analysis (i.e. 1 < parameter fold-change ≤ 2.25). Analogous to *M*'s power to diagnose "noisy" outlier studies, the power to identify an outlier study showing either systematically stronger or weaker than average effects increased with the number of variants interrogated, and was relatively independent of the number of studies in the meta-analysis. *M*'s dependence on the number of variants rather than studies is consistent with the manner in which *M* statistics are computed; in that, *M* statistics are calculated by aggregating *SPREs*, where *SPRE* statistics are calculated for each independently associated variant in the meta-analysis.

***M*'s power to identify systematic outlier studies is inversely related to the level of marker-specific heterogeneity in a meta-analysis.** The previous set of "real world" meta-analyses scenarios evaluated *M*'s power to identify systematic heterogeneity patterns in challenging environments, where all the susceptibility variants showed substantial marker-specific heterogeneity, i.e. $(I^2 = 60.5\%)$ per variant. In reality, risk variants in a meta-analysis will generally have varying levels of marker-specific heterogeneity. For example, risk variants in a recent GWAS meta-analysis of CAD had marker-specific heterogeneity levels ranging from $(0\% \leq I^2 \leq 60\%)$ (CARDIoGRAMplusC4D Consortium, 2015). As such, Monte-Carlo simulations were conducted to assess the impact of marker-specific heterogeneity on *M*'s power to identify systematic outliers. *M*'s power to diagnose systematic outliers was inversely related to the level of background random heterogeneity (i.e. marker-specific heterogeneity), Figure 2.5. Notably, *M* was better powered $(\geq 80\%)$ to detect systematic outliers when there were moderate levels of marker-specific heterogeneity $(I^2 \leq 50\%)$.

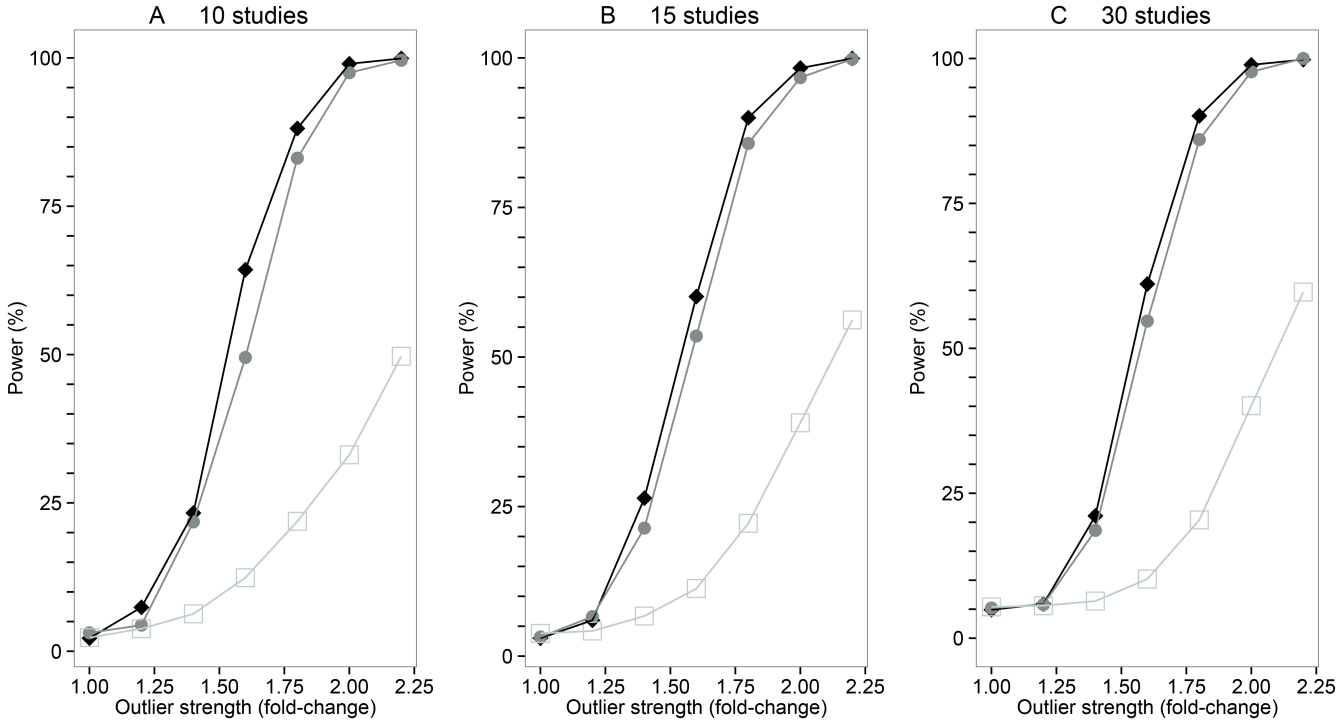


Figure 2.4: The power of the *M* statistic to detect systematic outlier studies. A power analysis of the *M* statistics in meta-analysis scenarios with varying numbers of studies and variants. The three panels show (from left to right) simulations for 10, 15 and 30 studies; within each panel 50, 25 and 10 variant simulations are shown by filled diamonds, filled circles, or open squares respectively. Furthermore, each data point represents a meta-analysis simulation with 1,000 Monte-Carlo replicates, where an outlier study was assigned genetic effects that are x-fold stronger than the effects assigned to the remaining studies showing typical effects. Effect-sizes for variants in the studies showing typical effects were allocated from an L – shaped distribution (Table 2.3) whilst effect-sizes for variants in the outlier study were calculated as a multiple of the typical effect-size. For example, effect-sizes for variants in an outlier study 2-fold-stronger than studies showing typical effects would be computed as (2 x (0.04, 0.12, 0.2, 0.28, 0.4, σ = 0.10). *M's* power improves as the number of independently associated variants increases. And is relatively independent of the number of studies.
Source: (Magosi et al., 2017)

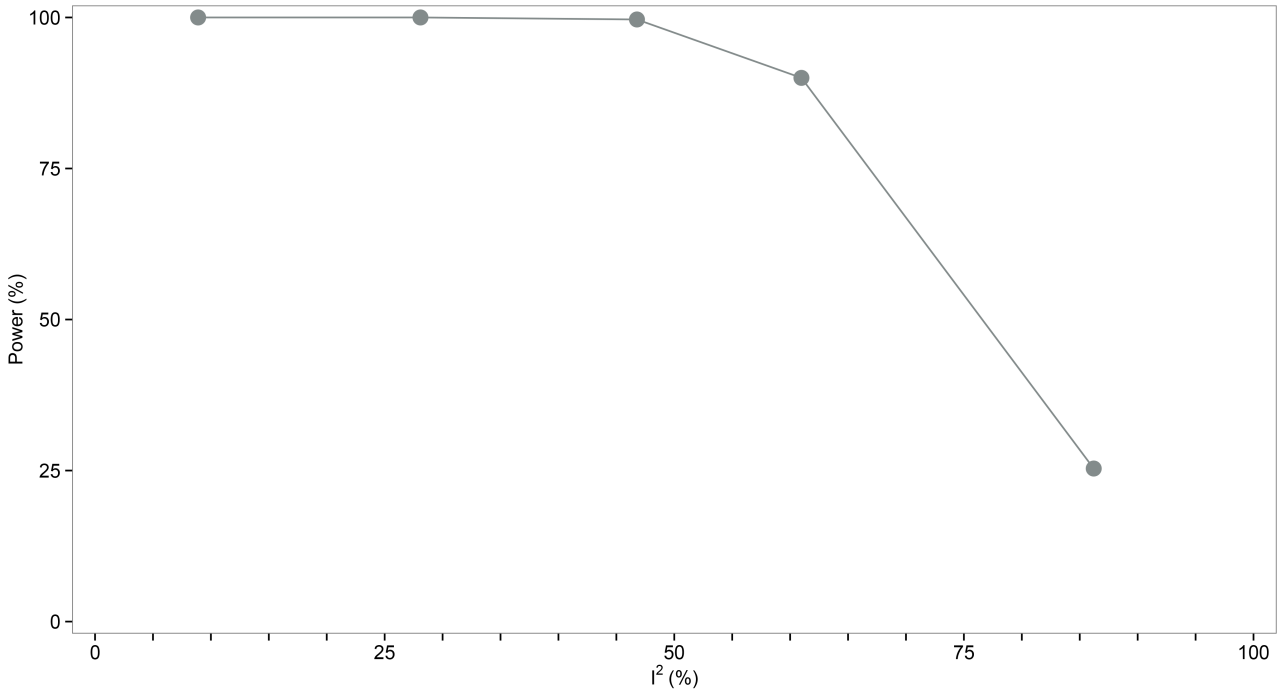


Figure 2.5: Power to detect systematic outlier studies in the presence of variable background heterogeneity. A power analysis of *M* involving Monte-Carlo meta-analysis scenarios varying the level of background heterogeneity (I^2 from 8.89% to 86.21%). Each data point represents a simulation based on 15 studies and 50 variants. Furthermore, all studies were equally weighted (i.e. SE = 0.08). Effect-sizes for variants in the studies showing typical effects were allocated from an L – shaped distribution (Table 2.3) whilst effect-sizes for variants in the outlier study were calculated as a multiple of the typical effect-size (i.e. 1.80 x (0.04, 0.12, 0.2, 0.28, 0.4, σ = 0.10) to model a 1.8-fold stronger-than-typical outlier study. *M's* power to identify systematic outliers is inversely related to the level of background heterogeneity at individual variants. Source: (Magosi et al., 2017)

2.3.2 Aim 2: Identifying systematic heterogeneity patterns in a GWAS meta-analysis of coronary artery disease

The *M* statistic was applied to a GWAS meta-analysis of CAD risk to demonstrate its utility in identifying outlier studies showing systematic heterogeneity patterns.

The CARDIoGRAMplusC4D consortium's 1,000 genomes based GWAS meta-analysis was an international project commissioned to identify common and low frequency risk loci for CAD and MI (CARDIoGRAMplusC4D Consortium, 2015). The meta-analysis was conducted using summary data (i.e. logistic regression coefficients scaled as log odds ratios) from 48 studies, based on individuals (60,801 cases and 123,504 controls) from six ancestral populations including: African American, Hispanic American, East Asian, South Asian, Middle Eastern and European. In addition, the meta-analysis had a broad CAD case definition that included individuals with clinical diagnoses of MI (with or without ST-elevation), acute coronary syndrome, chronic stable angina or coronary stenosis (>50%) in at least one coronary vessel. Case recruitment was predominantly retrospective (i.e. prevalent cases), for the prospective studies cases included a combination of incident and prevalent disease. The controls included three sets of population samples: 1) unscreened individuals for CAD (e.g. the UK 1958 Birth Cohort and National Blood Service controls genotyped for the WTCCC (Wellcome Trust Case Control Consortium et al., 2010); 2) individuals without a personal history of CAD; and 3) individuals that had not shown radiological evidence of vessel stenosis during coronary angiography.

The CARDIoGRAMplusC4D meta-analysis imputed variants genotyped in the individual studies to the 1000 genomes phase 1 version 3 haplotype training set, to avail a common set of ~9 million variants for analysis across the participating studies. Variants in the common set had minor allele frequencies > 0.005 and were well imputed in 60% of the 48 studies (i.e. imputation quality metrics, rsq > 0.3 for Minimac or info_proper > 0.4 for IMPUTE2). An additive-effects-only meta-analysis of the ~9 million variants revealed 2,213 genome-wide significant variants (i.e. $P < 5 \times 10^{-8}$) (CARDIoGRAMplusC4D Consortium, 2015).

Generally, *M* analyses can be performed with the GWAS significant lead variants in a meta-analysis or with a known panel of risk variants for the trait of interest. Two sets of *M* analyses were performed to dissect systematic heterogeneity in the CARDIoGRAMplusC4D meta-analysis dataset. The first set involved an *M* analysis of GWAS significant lead variants in the CARDIoGRAMplusC4D meta-analysis. The motivation here was to optimize *M*'s power to diagnose systematic heterogeneity patterns by using as many independently and strongly associated variants as available. The second set of analyses concerned an *M* analysis of the CARDIoGRAMplusC4D dataset using a panel of variants comprising both genome wide significant lead variants (i.e. $P < 5 \times 10^{-8}$) and suggestively associated lead variants (i.e. $P < 5 \times 10^{-5}$, FDR < 5%). The motivation for a polygenic analysis was to establish whether the inclusion of suggestively associated variants could reveal additional systematic heterogeneity patterns; especially since simulation studies had shown that *M*'s power to identify outlier studies displaying systematic heterogeneity patterns increases with the number of independently associated variants examined.

***M* analysis of the 46 GWAS significant lead variants**: To select as many strongly and independently associated genetic signals as available for heterogeneity analysis, the 2,213 variants reaching genome-wide significance in the CARDIoGRAMplusC4D meta-analysis were sorted by estimated effect-size and collapsed into 46 discrete loci. Notably, effect-size estimates (i.e. β coefficients scaled as $\log_e$(odds ratio)) were standardized and expressed in their chi-squared form to facilitate ranking of variants, such that the lead variant at each locus was the variant with the largest Wald statistic value, $\frac{\beta^2}{se^2}$. In order to select variants that were most likely to be in linkage equilibrium and thus independently associated with CAD risk, each lead variant was surrounded by a genetic distance window of ±0.5 centiMorgans. However, it is important to acknowledge the possibility of long range LD existing among the lead variants selected for heterogeneity analysis. Most of the lead variants showed some level of marker-specific heterogeneity, with 35 of the 46 lead variants having $(I^2 > 0\%)$ (Table 2.4). Given that the I^2 values in Table 2.4 are not accompanied by confidence intervals, marker-specific heterogeneity cannot be ruled out for the variants reporting $(I^2 = 0\%)$; especially since I^2 values tend to have wide confidence intervals, as noted in chapter 1 (section 1.2.4.2).

Furthermore, to minimize the occurrence of false positives owing to genomic inflation, a study-level genomic correction was applied to down-weight the influence (i.e. "expand" the standard errors) of studies that might have inflated association statistics. This was an important step that also ensured that systematic outlier studies identified by the *M* statistics were not a mere artefact of the winner's curse phenomenon. The fact that *M* statistics are precision weighted will also help to mitigate the impact of the winner's curse phenomenon, where studies with smaller sample sizes tend to have upwardly biased effect-size estimates. In addition, to

facilitate consistent comparisons of the studies in the CARDIoGRAMplusC4D meta-analysis, effect-size estimates (i.e. regression coefficients) for the 46 lead variants were judiciously "flipped" to: 1) align the effect alleles (and their corresponding summary allele frequencies), and 2) to ensure that the average effect-estimates of the variants examined were all in the same direction, such that the same allele was compared across the contributing studies of the meta-analysis.

Table 2.4: Lead variants at 46 genome-wide significant loci (p-value $\leq$ 5e-08) selected for computation of *M* statistics in the CARDIoGRAMplusC4D GWAS meta-analysis.

Chromosome	Variant identifier	Physical position (bp)	Genetic location (cM)	p-value	I^2 (%)
1	rs11206510	55,496,039	80.92	2.34E-08	0.00
	rs9970807	56,965,664	82.73	5.00E-14	21.64
	rs7528419	109,817,192	138.71	1.97E-23	15.49
	rs6689306	154,395,946	168.56	2.60E-09	40.59
	rs67180937	222,823,743	246.47	1.01E-12	38.28
2	rs16986953	19,942,473	39.63	1.45E-08	32.77
	chr2:21378433:D	21,378,433	42.04	2.89E-08	0.00
	chr2:44074126:D	44,074,126	68.66	2.60E-08	10.42
	rs7568458	85,788,175	115.47	3.62E-10	0.27
	rs17678683	145,286,559	166.34	3.00E-09	5.59
	chr2:203828796:I	203,828,796	215.74	2.15E-18	0.02
3	chr3:138099161:I	138,099,161	150.52	2.89E-09	0.00
4	rs17087335	57,838,583	75.65	4.59E-08	20.26
	rs4593108	148,281,001	155.95	8.82E-10	5.96
	rs72689147	156,639,888	164.90	6.07E-09	0.00
6	rs9349379	12,903,957	28.89	1.81E-42	27.04
	rs56336142	39,134,099	59.69	1.85E-08	0.00
	rs12202017	134,173,151	140.74	1.98E-11	31.13
	chr6:160434842:I	160,434,842	176.99	4.02E-11	0.06
	rs55730499	161,005,610	177.56	5.39E-39	39.28
7	rs2107595	19,049,388	33.84	8.05E-11	0.00
	rs11556924	129,663,496	139.51	5.34E-11	23.17
	rs3918226	150,690,176	168.01	1.69E-09	15.19
9	rs2891168	22,098,619	43.63	2.29E-98	57.68
	rs2519093	136,141,870	166.21	1.19E-11	31.07
10	rs2487928	30,323,892	57.93	4.41E-11	15.04
	rs1870634	44,480,811	67.07	5.55E-15	6.80
	rs1412444	91,002,927	117.68	5.15E-12	32.87
	rs11191416	104,604,916	130.61	4.65E-09	21.14
11	rs10840293	9,751,196	19.75	1.28E-08	17.19
	rs2128739	103,673,277	112.07	7.05E-11	16.10
12	rs2681472	90,008,959	103.39	6.17E-11	14.31
	rs11065979	112,059,557	130.27	1.93E-10	24.46
13	rs11617955	110,818,102	121.01	3.55E-08	0.00
	rs4773141	110,954,353	121.56	2.14E-09	31.55
	rs11838776	111,040,681	122.19	1.83E-10	0.00

Table continued on next page.

Table 2.4 – continued from previous page

Chromosome	Variant identifier	Physical position (bp)	Genetic location (cM)	p-value	I^2 (%)
14	rs10139550	100,145,710	106.34	1.38E-08	34.88
15	rs56062135	67,455,630	91.27	4.52E-09	0.00
	rs4468572	79,124,475	102.35	4.44E-16	31.44
	rs8042271	89,574,218	115.64	3.68E-08	15.50
17	rs7212798	59,013,488	86.12	1.88E-08	13.77
18	rs663129	57,838,401	82.16	3.20E-08	0.00
19	rs56289821	11,188,247	31.87	4.44E-15	29.36
	rs4420638	45,422,946	70.73	7.07E-11	46.18
21	rs28451064	35,593,827	36.28	1.33E-15	30.52
22	rs180803	24,658,858	22.67	1.64E-10	0.00

Genome-wide significant variants (p-value $\leq$ 5e-08) in the CARDIoGRAMplusC4D GWAS meta-analysis were grouped by genetic distance ($\pm$ 0.5 centiMorgans from lead variant) yielding 46 loci. The lead variant at each locus was defined as the variant with the smallest p-value. Source: (Magosi et al., 2017).

***M* identifies substantial systematic heterogeneity in the CARDIoGRAMplusC4D GWAS meta-analysis**: To search for systematic heterogeneity patterns in the CARDIoGRAMplusC4D meta-analysis, *M* statistics and their accompanying standard errors were calculated for each of the 48 participating studies, giving 48 *M* statistics in all. *M* statistic values for several studies deviated from the average genetic effect (*M* statistic value = 0) signalling the presence of systematic heterogeneity (Figure 2.6 and Table 2.5). In particular, studies 9, 38 and 48, showed significantly stronger than average effects (Bonferroni corrected p-values < 0.05) while 10, 19, 24 and 28 reported noticeably weaker effects than average (Bonferroni corrected p-values < 0.05). An inverse-variance weighted meta-analysis of the *M* statistics exposed considerable dispersion in the study-effect estimates across the variants examined, indicating substantial systematic heterogeneity in the CARDIoGRAMplusC4D dataset $(I^2 = 85.9\%)$ (Figure 2.6).

Explaining systematic heterogeneity in the CARDIoGRAMplusC4D meta-analysis: Next, it was important to explore differences in the characteristics of the participating studies that could offer some explanation for the observed systematic heterogeneity in the study effects. Since the primary interest was in identifying sources of systematic heterogeneity across genetic signals rather than those for marker-specific heterogeneity at individual variants, the classical meta-regression framework was extended by conducting an inverse-variance weighted (i.e. $1/({SE_{M_s}}^2 + \hat{\tau}^2)$) random-effects meta-regression of the *M* statistics, where ${SE_{M_s}}^2$ denotes the sampling variance in the s^{th} study. Meta-regressions involving multiple explanatory variables are subject to a multiple comparison ("statistical testing") burden. One advantage that a meta-regression of *M* statistics has over single variant meta-regressions is that, meta-regressing *M* statistics lowers the dimensionality of the multiple testing burden *V*-fold, where *V* denotes the

number of variants; this is because *M* statistics aggregate heterogeneity information across multiple variants.

Study characteristics explored in the meta-regression of *M* statistics could be loosely grouped into two categories, those relating to quality control checks (i.e. imputation quality, genotype call rate, Hardy Weinberg equilibrium thresholds, systematic allele frequency differences) and those concerning ascertainment criteria of the participating studies (i.e. family-history of CAD, age-of-CAD onset, case-control ratio, percentage of cases with myocardial infarction and ancestry). The family-history covariate was coded into two groups, those that had recruited cases with a positive family-history of CAD and those that had not. Ancestry was defined by six study groupings representing: African and Hispanic American, East and South Asian, Middle Eastern and European ancestries. Studies in the CARDIoGRAMplusC4D meta-analysis had overlapping eligibility criteria for cases; as such the age-of-CAD onset variable was defined by broadly grouping studies into earlier-onset $(\leq 55\ years)$ and later-onset CAD groups (Table 2.6). Notably, the above study characteristics were selected for exploration as potential sources of heterogeneity in the CARDIoGRAMplusC4D meta-analysis for the following reasons: 1) participating studies had been independently commissioned and were based on individuals of diverse ancestries, as such there was a possibility of heterogeneity arising from variable: case-control ratios, quality control thresholds and ancestry; 2) studies that recruited cases with a family-history of CAD and earlier-onset CAD would likely be genetically enriched and better powered to detect genetic signals than other studies; 3) to explore whether sub-phenotype variability (i.e. percentage of cases with MI) contributed to systematic heterogeneity in the CARDIoGRAMplusC4D meta-analysis.

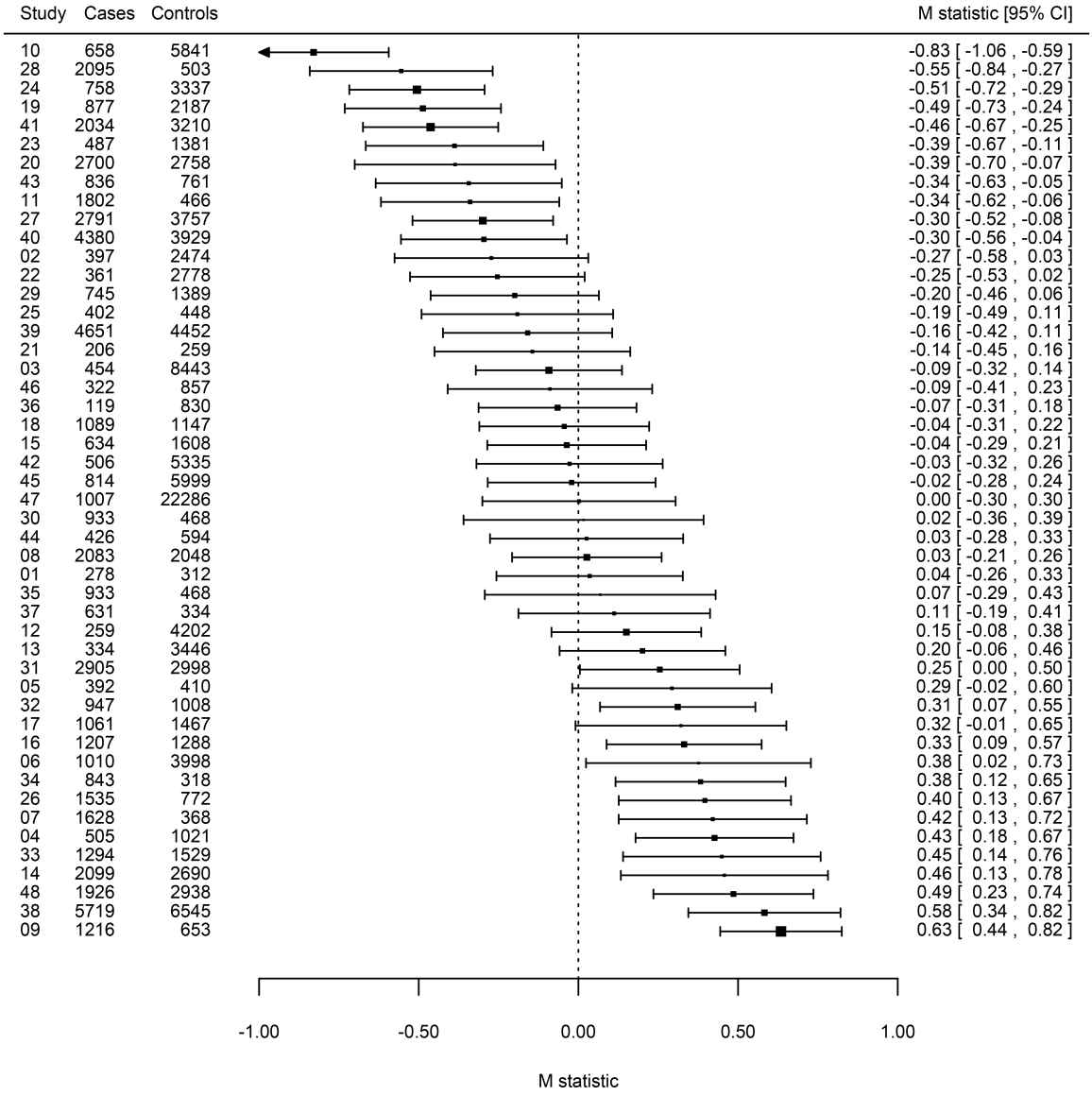


Figure 2.6: Forest plot of *M* statistics summarizing systematic patterns of heterogeneity among studies in the CARDIOGRAMplusC4D GWAS meta-analysis. Sorted *M* statistics are presented for individual studies represented by filled squares with their 95% confidence intervals shown by horizontal lines; the sizes of the squares are proportional to each studies' inverse-variance weighting. Studies showing weaker (M $<$ 0) than average genetic effects can be distinguished from those showing stronger (M $>$ 0) than average effects. The dispersion of *M* statistics indicates substantial systematic heterogeneity in the CARDIOGRAMplusC4D meta-analysis ($I^2 = 85.9\%$). Source: (Magosi et al., 2017).

Table 2.5: Studies showing substantial systematic heterogeneity patterns in the CARDIoGRAMplusC4D GWAS meta-analysis at threshold, $\alpha = 0.05$

Study identifier	# Variants	Bonferroni corrected p-value	M	Direction of effect
9	45	8.09E-04	0.634	stronger than average
38	44	3.73E-03	0.583	
48	46	4.76E-02	0.486	
10	46	8.91E-07	-0.829	weaker than average
19	46	4.59E-02	-0.487	
24	46	2.90E-02	-0.506	
28	46	8.07E-03	-0.555	

The table shows results from an *M* analysis of lead variants at GWAS significant loci(46) in the CARDIoGRAMplusC4D meta-analysis. Bonferroni correction applied for multiple testing of 48 studies. ***M*** (α = 0.05): $\pm$0.4835. Source: (Magosi et al., 2017).

Table 2.6: Ancestry, age-of-disease onset and family-history of disease grouping for studies in the CARDIoGRAMplusC4D GWAS meta-analysis.

Study identifier	Ancestry	Family-history	Earlier-onset $\leq$ 55 years	Age-of-disease onset, cases (mean $\pm$ SD)	References	Comments
ADVANCE	European	NO	YES	45.8$\pm$6.2	PMID:21378990	
AGES	European	NO	NO	79.1$\pm$5.5[a]	PMID:17351290; PMID:26950853	Cohort study. Examined genetic susceptibility to disease in old age.
ARIC	European	NO	NO	65.17$\pm$6.9[a]	PMID:2646917; PMID:26950853	Cohort study.
BAS	East Asian	NO	YES	52.13$\pm$8.62	PMID:22751097	
CARDIOGENICS	European	YES	YES	male(49.0$\pm$7.4)/female(52.8$\pm$8.2)	PMID:17634449	Ascertainment critera included having at least one first degree relative with premature coronary artery disease.
CAS	East Asian	YES	YES	49.18$\pm$8.03	PMID:22751097	83.8% of cases had a family history of CAD.
CCGB_2	European	YES	YES	49.8$\pm$7.7	PMID:22319020	Featured early-onset CAD cases with a family history of CAD. Age of onset of CAD was required to be: men $\leq$ 55 years, women $\leq$ 65 years.
COROGENE	European	NO	NO	66.0$\pm$11.8	PMID:21642350; PMID:23202125	
DUKE_2	European	YES	NO	56.7$\pm$9.7	PMID:22319020	Featured early-onset CAD cases with a family history of CAD. Ascertainment criteria required cases to be: men $\leq$ 55 years, women $\leq$ 65 years.
EGCUT	European	NO	NO	see comments	PMID:24518929; PMID:23202125	Study participants were aged: $\geq$18 years. Mean age of CAD onset for EGCUT GWAS (69.8$\pm$13.8); EGCUT metabochip (61.7$\pm$10.8).
FGENTCARD	Middle Eastern	NO	NO	61.0$\pm$11.1	PMID:23202125; PMID:22745674	
FHS	European	NO	NO	75.2$\pm$12.2[a]	PMID:26950853	Prospective study.

Table continued on next page.

Table 2.6 – continued from previous page

Study identifier	Ancestry	Family-history	Earlier-onset $\leq$ 55 years	Age-of-disease onset, cases (mean $\pm$ SD)	References	Comments
FamHS	European	YES	YES	see comments	PMID:8651220	Population based study. Featured participants with: family history of coronary heart disease, early-onset coronary heart disease. Probands were participants aged 45 to 69 years. Mean age of CAD onset not indicated.
GenRIC	East Asian	NO	YES	51.7$\pm$7.5	PMID:23364394; PMID:23202125	
GerMIFSI	European	YES	YES	50.2$\pm$7.9	PMID:21088011	Study population featured participants having: first degree sibling(s) with severe CAD, early-onset disease. Myocardial infarction $\leq$ 65 years.
GerMIFSII	European	YES	YES	51.3$\pm$7.6	PMID:21088011	Study population featured participants having: a family history of CAD, early-onset disease, myocardial infarction $\leq$ 60 years.
GerMIFSIII	European	NO	NO	58.4$\pm$9.8	PMID:21088011	Age at myocardial infarction event $\leq$ 60 years.
GerMIFSIV	European	NO	NO	see comments	PMID:26343387	Age of study participants: men $<$ 65 years, women $<$ 70 years. Mean age of CAD onset not indicated.
GoDARTS	European	NO	NO	61.5$\pm$10.5	PMID:23202125; PMCID: PMC2127363	
HPS	European	NO	NO	58.8$\pm$8.4	PMID:21378988	Average age at first myocardial infarction event: $\leq$ 60 years.
HSDS	European	NO	NO	see comments	PMID:18615340	Prospective case-control autopsy study of men aged 35 to 69 years. Mean age of CAD onset not indicated.
IPM-AfricanAmerican	African American	NO	NO	see comments	PMID:23743551	Age of study participants: $\geq$ 18 years. Mean age of CAD onset not indicated.

Table continued on next page.

Table 2.6 – continued from previous page

Study identifier	Ancestry	Family-history	Earlier-onset $\leq$ 55 years	Age-of-disease onset, cases (mean $\pm$ SD)	References	Comments
IPM-EuropeanAmerican	European	NO	NO	see comments	PMID:23743551	Age of study participants: $\geq$ 18 years. Mean age of CAD onset not indicated.
IPM-HispanicAmerican	Hispanic	NO	NO	see comments	PMID:23743551	Age of study participants: $\geq$ 18 years. Mean age of CAD onset not indicated.
ITH_2	European	NO	NO	63.3$\pm$10.6	PMID:15364185; PMID:23202125	
LIFE-Heart	European	NO	NO	see comments	PMID:22216169	Study participants were aged: 18 to 85 years. Mean age of CAD onset not indicated.
LOLIPOP	South Asian	NO	NO	59.3$\pm$9.7	PMID:23202125; PMID:21378988	Average age at first myocardial infarction event: $\leq$ 60 years.
LURIC	European	NO	NO	58.9$\pm$13.0	PMID:23202125; PMID:11258203	Mean age of study participants was 62.11$\pm$11 years in men and 65$\pm$10 years in women.
MAYOVDB	European	NO	NO	see comments	PMID:24689004	Mean age of study partcipants was 70.4$\pm$11.2. Mean age of CAD onset not indicated.
MEDSTAR	European	NO	YES	48.9$\pm$6.4	PMID:21239051	Cases were selected to be young men $\leq$ 55 years, women $\leq$ 60 years.
MIGEN	European	NO	YES	male(41)/female(47)	PMID: 19198609	Featured participants with: early-onset myocardial infarction, men $\leq$ 50 years, women $\leq$ 60 years.
OHGS_A2	European	YES	YES	48.2$\pm$7.0	PMID: 22319020	Featured early-onset CAD cases with a family history of CAD, men $\leq$ 55 years, women $\leq$ 65 years.
OHGS_B2	European	YES	YES	49.8$\pm$7.7	PMID: 22319020	Featured early-onset CAD cases with a family history of CAD, men $\leq$ 55 years, women $\leq$ 65 years.
OHGS_C2	European	YES	YES	54.4$\pm$8.6	PMID: 22319020	Featured early-onset CAD cases with a family history of CAD, men $\leq$ 55 years, women $\leq$ 65 years.

Table continued on next page.

Table 2.6 – continued from previous page

Study identifier	Ancestry	Family-history	Earlier-onset $\leq$ 55 years	Age-of-disease onset, cases (mean $\pm$ SD)	References	Comments
PENNCATH	European	NO	YES	52.7$\pm$7.6	PMID:21239051	Cases were selected to be young, men $\leq$ 60 years, women $\leq$ 65 years.
PIVUS	European	NO	NO	65.0$\pm$7.2	PMID:23202125; PMID:16141402	Elderly population
PREDICTCVD	European	NO	NO	see comments	PMID:26343387; PMID: 25422363	Age of study participants was: 30 to 59 years. Mean age of CAD onset not indicated.
PROCARDIS	European	YES	YES	53.2$\pm$7.2	PMID:21378988	Precocious; 80% had a sibling in whom CAD was diagnosed before age 66.
PROMIS1	South Asian	NO	YES	53.8$\pm$10.6	PMID:21378988; PMID:19404752	Age of study participants was: 30 to 80 years.
PROMIS2	South Asian	NO	YES	53.8$\pm$10.6	PMID:21378988	Average age at first myocardial infarction event: $\leq$ 60 years.
PROSPER	European	NO	NO	see comments	PMID:21977987	Prospective study. Age of study participants was 70 to 82 years. Mean age of participants was 75.3$\pm$3.4 years. Mean age of CAD onset not indicated.
RS	European	NO	NO	80.6$\pm$10.1^a	PMID:24258680; PMID:26950853	Prospective study. Examined determinants of chronic disease in older adults. Age of study participants was $\geq$ 55 years.
SDS	South Asian	NO	NO	see comments	PMID:23209189; PMID:24795349	Mean age of study particpants was 53.1$\pm$12.2. Mean age of CAD onset not indicated.
THISEAS	European	NO	NO	61.1$\pm$10.4	PMID:20167083; PMID:23202125	
TWINGENE	European	NO	NO	see comments	PMID: 8981957; PMID: 26343387	Study population included participants aged 52 to 86 years. Mean age of participants was 65.5 years. Mean age of CAD onset not indicated.

Table continued on next page.

Table 2.6 – continued from previous page

Study identifier	Ancestry	Family-history	Earlier-onset $\leq$ 55 years	Age-of-disease onset, cases (mean $\pm$ SD)	References	Comments
ULSAM	European	NO	NO	72.5$\pm$8.7	PMID:16030278; PMID:23202125	Elderly men. $\geq$ 70 years.
WGHS	European	NO	NO	see comments	PMID:18070814	Study population comprised women $\geq$ 45 years. Mean age of CAD onset not indicated.
WTCCC	European	YES	YES	49.8$\pm$7.7	PMID: 22319020	Featured early-onset CAD cases with a family history of CAD, men $\leq$ 55 years, women $\leq$ 65 years.

a Incident myocardial infarction age (mean$\pm$SD). CAD: Coronary Artery Disease. Source: (Magosi et al., 2017).

Differences in ancestry, family-history and age-of-disease onset contribute to systematic heterogeneity in the CARDIoGRAMplusC4D GWAS meta-analysis: A simple forward model selection approach was adopted to identify the most informative covariates in the CARDIoGRAMplusC4D meta-analysis. Given that critical levels of significance customarily employed for inclusion of variants in forward selection models (e.g. $0.15 \leq P \leq 0.20$) (Hosmer and Lemeshow, 2005) do not reflect the actual type-1 error rate for each explanatory variable added, it is important to note that the use of forward selection as a model selection tool is exploratory rather than definitive. Differences in ancestry ($F_{5,42}$ = 2.52, P = 0.044), CAD family-history ($F_{1,46}$ = 29.49, P = 2.0 $\times 10^{-6}$) and age-of-CAD onset ($F_{1,46}$ = 20.65, P = 0.00004) explained a substantial proportion of the systematic heterogeneity in the CARDIoGRAMplusC4D meta-analysis (Figure 2.7). In particular, studies that included cases with a positive family-history of CAD and/or earlier-onset CAD showed systematically stronger effects than average.

To dissect the influence of individual ancestry groups on the variability in the average effect-size, further meta-regressions of the *M* statistics were conducted comparing each ancestry to all other ancestry groups. The 3 East Asian studies showed genetic effects that were systematically stronger than those for other ancestry groups ($F_{1,46}$ = 4.75, P = 0.034). Comparatively, there was no evidence to suggest that genetic effects reported by the 4 South Asian studies ($F_{1,46}$ = 2.99, P = 0.090) or the 38 European studies ($F_{1,46}$ = 1.24, P = 0.271) systematically deviated from the average genetic effect. Furthermore, a multiple meta-regression model of the *M* statistics was performed employing East Asian ancestry, earlier-onset and family-history of CAD as covariates, to assess their joint contribution to observed systematic heterogeneity in the CARDIoGRAMplusC4D meta-analysis. Together, the three covariates

explained a considerable fraction of systematic heterogeneity in the CARDIoGRAMplusC4D meta-analysis ($F_{3,44}$ = 13.91, P = 1.6 ×10^{-6}; adjusted R^2 = 53.2%) (Table 2.7).

Conversely, the contribution of variable quality control criteria (i.e. imputation quality, genotype call rate, Hardy Weinberg equilibrium thresholds), to systematic heterogeneity in the CARDIoGRAMplusC4D meta-analysis was negligible (i.e. $P > 0.05$). Similarly, there was no evidence to suggest either case-control imbalance (i.e. variable case-control ratios) or sub-phenotype variability as substantial sources of systematic heterogeneity in the CARDIoGRAMplusC4D meta-analysis (i.e. $P > 0.05$). The negligible contribution of sub-phenotype variability (i.e. percentage of cases with MI) to systematic heterogeneity was consistent with findings by the CARDIoGRAMplusC4D consortium (CARDIoGRAMplusC4D Consortium, 2015), where a sub-group analysis of cases with a reported history of MI did not uncover any additional variants in the CARDIoGRAMplusC4D meta-analysis that preferentially associated with MI.

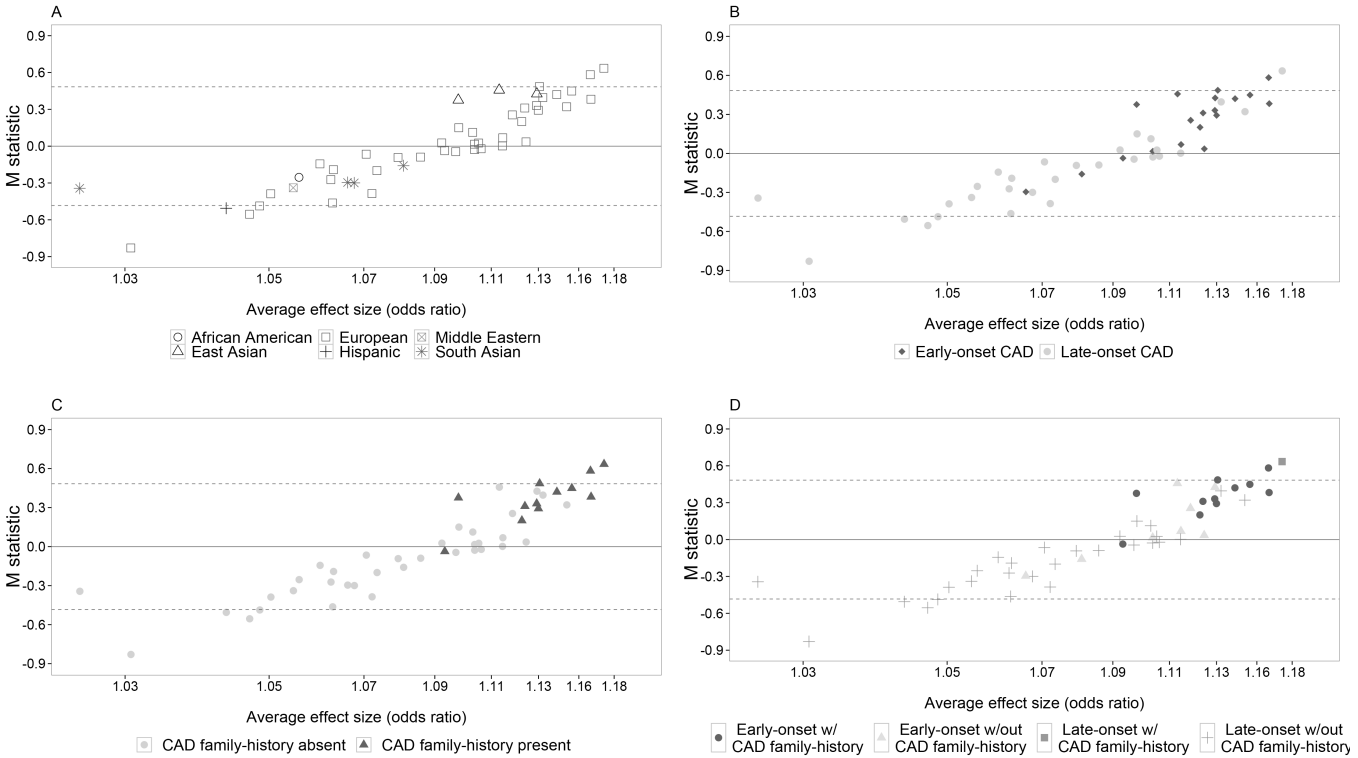


Figure 2.7: Systematic heterogeneity in the CARDIoGRAMplusC4D meta-analysis can be explained by differences in age-of-CAD onset, CAD family-history and ancestry. *M* statistics for each study in the CARDIoGRAMplusC4D meta-analysis (Y- axis) are plotted against the average variant effect-size (expressed as odds ratios) (X-axis) in each study. Panel A shows the ancestry of each study, panel B distinguishes early-onset from late-onset studies and panel C identifies studies ascertained with a positive family-history of coronary artery disease. Panel D is a composite plot showing the degree of genetic enrichment among the studies in the meta-analysis, which ranged from non-enriched (late-onset studies without a positive family-history of coronary artery disease) to doubly enriched (early-onset studies with a positive family-history of coronary artery disease). The dashed lines indicate the Bonferroni corrected 5% significance threshold ($M = \pm 0.483$) to allow for multiple testing of 48 studies. Source: (Magosi et al., 2017)

Table 2.7: Meta-regression of *M* statistics in the CARDIoGRAMplusC4D GWAS meta-analysis employing East Asian ancestry, age-of-CAD onset and CAD family-history as covariates.

Variable	β	SE(β)	p-value
East-Asian ancestry	0.348	0.172	0.049
early-onset	0.117	0.105	0.268
family-history	0.409	0.111	0.001

Family-history was a stronger predictor of systematic heterogeneity in the CARDIoGRAMplusC4D meta-analysis as compared to East Asian ancestry and early-onset CAD. Source: (Magosi et al., 2017).

Investigating systematic allele frequency differences as a source of heterogeneity: To consider whether the remaining variability in the average effect-size could be partly explained by systematic allele frequency differences among studies in the CARDIoGRAMplusC4D meta-analysis. A multiple meta-regression of *M* statistics was performed employing the top-2 coordinates of a pairwise Euclidean genetic distance matrix as covariates. The 48 x 48 pairwise Euclidean genetic distance matrix compared summary allele frequencies of GWAS-significant lead variants (46) in the CARDIoGRAMplusC4D meta-analysis across the 48 participating studies. To better illustrate the construction of the pairwise Euclidean genetic distance matrix, consider for example a genetic association meta-analysis with *S* studies and *V* GWAS-significant lead variants, where pairwise Euclidean distances of summary allele frequencies at the *V* GWAS-significant lead variants are calculated between a reference and all participating studies, wherein studies in the meta-analysis successively act as the reference study yielding a square array. Then the pairwise Euclidean genetic distance matrix would be given by,

$$\begin{bmatrix} 0 & D_{1,2} & D_{1,3} & \cdots & D_1, s \\ D_{2,1} & 0 & D_{2,3} & \cdots & D_2, s \\ D_{3,1} & D_{3,2} & 0 & \cdots & D_3, s \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ D_s, 1 & D_s, 2 & D_s, 3 & \cdots & 0 \end{bmatrix}.$$

The pairwise Euclidean genetic distance between summary allele frequencies at the *V* GWAS-significant lead variants in the s^{th} study and the reference study, r would be given by $D_{s,r} = \sqrt{\sum_{v=1}^{V}(p_{rv} - p_{sv})^2}$ such that $D_{s,r} = D_{r,s}$. Where r is anyone of the s participating studies, p_{rv} and p_{sv} are summary allele frequencies at the v^{th} GWAS-significant lead variant for the reference and s^{th} study, respectively.

The 48 x 48 pairwise Euclidean genetic distance matrix of summary allele frequencies at GWAS-significant lead variants (46) in the CARDIoGRAMplusC4D meta-analysis was then subjected to a multi-dimensional scaling analysis (also known as a principal coordinates analysis) using the cmdscale function in the *R*® stats package, to extract the top-2 coordinates that captured the majority of the summary allele frequency variability. A multiple meta-regression of *M* statistics employing the resulting top-2 coordinates as covariates revealed a negligible contribution of systematic allele frequency differences to the observed systematic heterogeneity in the CARDIoGRAMplusC4D meta-analysis ($F_{2,45}$ = 0.17, P = 0.842). This might reflect the fact that the CARDIoGRAMplusC4D consortium had inspected the dataset for systematic flipping of alleles as part of their quality control procedure. Notably, the inspection involved binning allele frequencies for each study in the meta-analysis and comparing them with those of the other participating studies (CARDIoGRAMplusC4D Consortium, 2015).

Inferring ancestry groups of studies in the CARDIoGRAMplusC4D meta-analysis from summary allele frequencies of the GWAS-significant lead variants (46): Further to facilitating an investigation of heterogeneity due to systematic allele frequency differences. The multi-dimensional scaling analysis of summary allele frequencies at GWAS-significant lead variants in the CARDIoGRAMplusC4D meta-analysis provided an opportunity to check concordance between demographic information from participating studies and the ancestry information captured by genetic data in the meta-analysis. Notably, a multi-dimensional scaling plot of the top-2 coordinates largely showed agreement between the demographic information furnished by studies in the CARDIoGRAMplusC4D meta-analysis and ancestral populations reflected in the summary allele frequencies (Figure 2.8 A). Curiously, studies 30 and 35 identified as European in their demographic information but did not cluster with other studies based on

individuals of European ancestry (Figures 2.8 B-D). Multi-dimensional scaling plots of coordinates 1,4 suggested possible admixture with South Asian ancestry in the two studies. Alternatively, the outlier status of studies 30 and 35 could signal possible genotyping errors, especially since the two studies were genotyped at the same site.

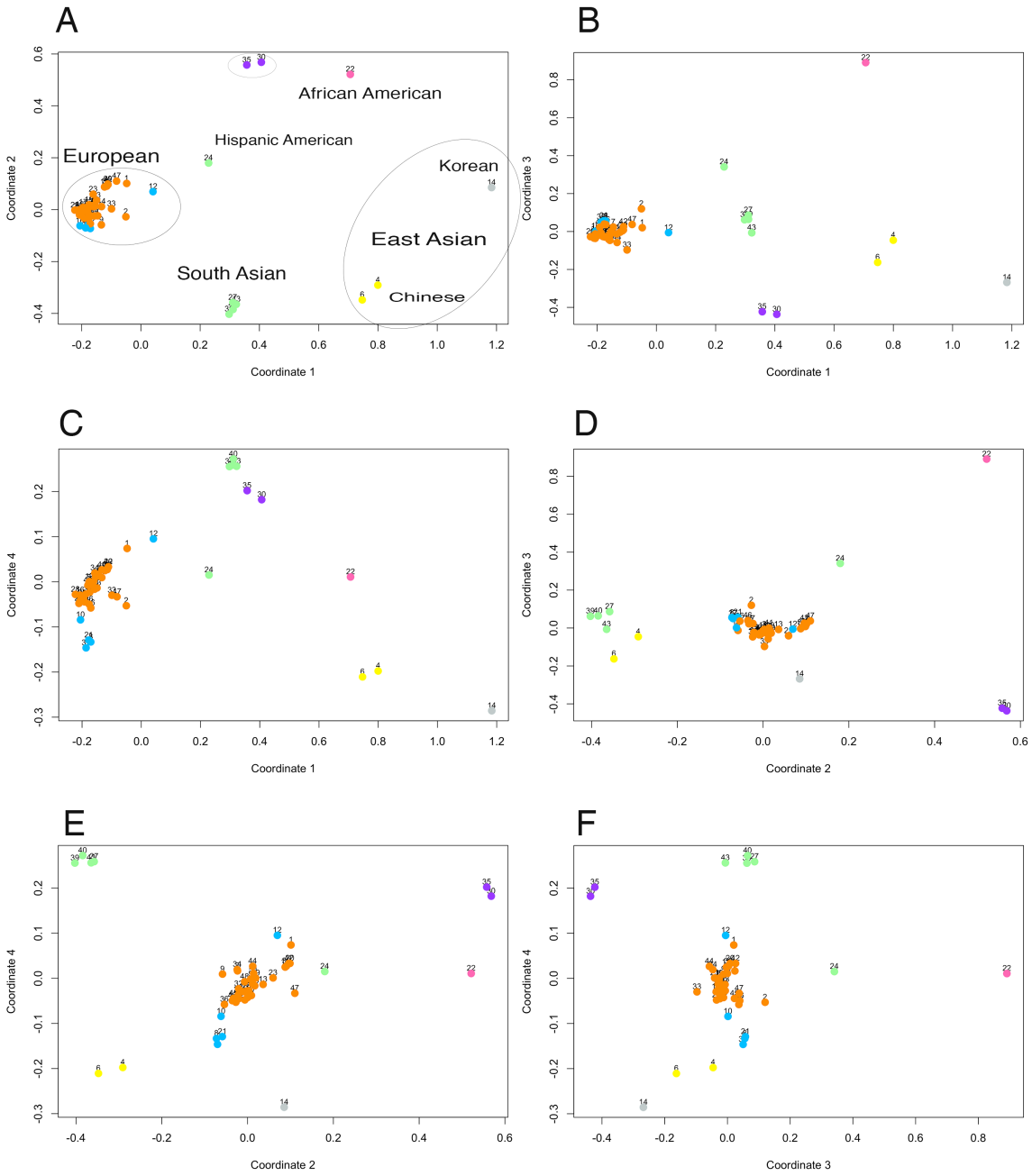


Figure 2.8: Inferring ancestry groups of studies in the CARDIoGRAMplusC4D meta-analysis from summary allele frequencies of lead variants at GWAS-significant loci (46). A multi-dimensional scaling plot generated from pairwise Euclidean distances of summary allele frequencies at, 46 GWAS-significant loci ($P < 5 \times 10^{-8}$) in the CARDIoGRAMplusC4D meta-analysis. Panels A-F show the top-4 coordinates that captured the majority of the allele frequency variability in the CARDIoGRAMplusC4D meta-analysis. The multi-dimensional scaling plot is largely consistent with the demographic information submitted by the participating studies and is able to show diverse ancestry groups (African American, Hispanic American, East Asian, South Asian and European) in the CARDIoGRAMplusC4D meta-analysis. Studies 30 and 35 identified as European in their demographic information and yet were isolated from the European ancestry cluster in the multi-dimensional scaling plot. This may highlight potential admixture with South Asian ancestry (Panel C) or genotyping errors in the two studies (Panels D-F). Source: Lerato Magosi (Farrall and Hopewell labs).

Polygenic *M* analysis employing an extended panel of 214 FDR-defined lead variants: Next an *M* analysis of the CARDIoGRAMplusC4D meta-analysis was conducted, this time using an extended panel of variants that included both genome-wide significant (i.e. $P < 5 \times 10^{-8}$) and suggestively associated variants (i.e. $P < 5 \times 10^{-5}$, FDR < 5%), (Table 2.8). The FDR-defined list comprised 214 independently associated lead variants, each centred on a genetic distance window of 1.0 centiMorgans (CARDIoGRAMplusC4D Consortium, 2015). Notably, 44 of the 214 lead variants were also part of the list of 46 GWAS-significant lead variants (Table 2.8). The systematic heterogeneity patterns identified in the *M* analysis of the 214 FDR-defined lead variants were largely consistent with those found in the *M* analysis of the 46 GWAS-significant lead variants, as evidenced by the identification of studies: 10, 28, 38 and 48 as significant outliers in both analyses (Figures 2.9 and 2.10; Table 2.9). Additionally, four studies (7, 15, 40 and 41) that were previously statistically insignificant in the *M* analysis of the 46 GWAS-significant lead variants were now significant outliers in the *M* analysis of the extended panel of 214 FDR-defined lead variants (Table 2.9). Suggesting that the additional variants in the extended FDR-panel improved power to detect systematic heterogeneity patterns in the CARDIoGRAMplusC4D meta-analysis. Moreover, three studies (9, 19 and 24) that had previously shown significant systematic heterogeneity patterns in the *M* analysis of the 46 GWAS-significant lead variants were now statistically insignificant (Table 2.9). This might reflect the fact that the polygenic *M* analysis only included 44 of the 46 GWAS-significant lead variants; and that the 2 GWAS-significant lead variants that were not part of the polygenic *M* analysis might have been important contributors to systematic heterogeneity patterns in the three studies (9, 19 and 24).

Table 2.8: FDR-defined lead variants (214) selected for a polygenic *M* analysis of the CARDIoGRAMplusC4D GWAS meta-analysis ($FDR \leq 5\%$, $P < 5 \times 10^{-5}$).

No.	Variants	Chromosome	GWAS-significant
1	rs11206510	1	YES
2	rs11810571	1	NO
3	rs2789422	1	NO
4	rs2820315	1	NO
5	rs2843152	1	NO
6	rs28470722	1	NO
7	rs35465346	1	NO
8	rs3930524	1	NO
9	rs6689306	1	YES
10	rs6701118	1	NO
11	rs72676528	1	NO
12	rs72702224	1	NO
13	rs7528419	1	YES
14	rs9970807	1	YES
15	rs67180937	1	YES
16	rs76233235	1	NO
17	chr2:203828796:I	2	YES
18	chr2:21378433:D	2	YES
19	chr2:227029688:D	2	NO
20	chr2:44074126:D	2	YES
21	rs10168194	2	NO
22	rs114031344	2	NO
23	rs1150500	2	NO
24	rs12619842	2	NO
25	rs13003675	2	NO
26	rs143803699	2	NO
27	rs16986953	2	YES
28	rs17517928	2	NO
29	rs17678683	2	YES
30	rs2552527	2	NO
31	rs2712172	2	NO
32	rs2715879	2	NO
33	rs33998987	2	NO
34	rs3771719	2	NO
35	rs57759964	2	NO
36	rs7568458	2	YES
37	rs78490412	2	NO
38	rs925229	2	NO
39	chr3:138099161:I	3	YES
40	chr3:172117455:D	3	NO
41	rs114423570	3	NO
42	rs12632550	3	NO
43	rs142695226	3	NO
44	rs433903	3	NO
45	rs73222236	3	NO
46	rs748431	3	NO
47	rs7622285	3	NO
48	rs7623687	3	NO
49	chr4:82625720:D	4	NO
50	rs10857147	4	NO
51	rs11099401	4	NO
52	rs11723436	4	NO

Table continued on next page.

Table 2.8 – continued from previous page

No.	Variants	Chromosome	GWAS-significant
53	rs13132376	4	NO
54	rs147339903	4	NO
55	rs17087335	4	YES
56	rs1873197	4	NO
57	rs2280003	4	NO
58	rs2522518	4	NO
59	rs35879803	4	NO
60	rs4593108	4	YES
61	rs72689147	4	YES
62	rs869396	4	NO
63	rs10071096	5	NO
64	rs112941079	5	NO
65	rs1963861	5	NO
66	rs2081914	5	NO
67	rs288187	5	NO
68	rs4074793	5	NO
69	rs4835911	5	NO
70	rs58811396	5	NO
71	chr6:2274487:D	6	NO
72	rs12202017	6	YES
73	rs1591805	6	NO
74	rs194937	6	NO
75	rs2146324	6	NO
76	rs2153219	6	NO
77	rs2876643	6	NO
78	rs421329	6	NO
79	rs4472337	6	NO
80	rs55730499	6	YES
81	rs56336142	6	YES
82	rs57349798	6	NO
83	rs742115	6	NO
84	rs742493	6	NO
85	rs9306	6	NO
86	rs9349379	6	YES
87	rs9379774	6	NO
88	rs9486719	6	NO
89	rs115696548	6	NO
90	rs11556924	7	YES
91	rs2107595	7	YES
92	rs3735352	7	NO
93	rs3918226	7	YES
94	rs4719608	7	NO
95	rs68170813	7	NO
96	rs73184021	7	NO
97	rs140683786	7	NO
98	rs189046492	7	NO
99	chr8:142230002:D	8	NO
100	rs113756303	8	NO
101	rs12675257	8	NO
102	rs144014513	8	NO
103	rs17411031	8	NO
104	rs2001846	8	NO
105	rs2575716	8	NO
106	rs28463395	8	NO
107	rs56356271	8	NO

Table continued on next page.

Table 2.8 – continued from previous page

No.	Variants	Chromosome	GWAS-significant
108	rs59334183	8	NO
109	rs181489382	8	NO
110	rs10818576	9	NO
111	rs10868271	9	NO
112	rs111915373	9	NO
113	rs1154673	9	NO
114	rs1411357	9	NO
115	rs2519093	9	YES
116	rs2891168	9	YES
117	rs75657982	9	NO
118	rs781622	9	NO
119	chr10:75595440:D	10	NO
120	rs11191416	10	YES
121	rs1412444	10	YES
122	rs1870634	10	YES
123	rs1888657	10	NO
124	rs2487928	10	YES
125	rs55702147	10	NO
126	rs59898454	10	NO
127	rs7082705	10	NO
128	rs2257129	10	NO
129	rs148498907	10	NO
130	chr11:100634736:D	11	NO
131	chr11:77185247:I	11	NO
132	rs10840293	11	YES
133	rs11601507	11	NO
134	rs12801636	11	NO
135	rs1351525	11	NO
136	rs2128739	11	YES
137	rs73562014	11	NO
138	rs9630156	11	NO
139	rs184875903	11	NO
140	chr12:76386403:I	12	NO
141	rs10785143	12	NO
142	rs10841443	12	NO
143	rs11057789	12	NO
144	rs11065979	12	YES
145	rs11172113	12	NO
146	rs11174220	12	NO
147	rs12826942	12	NO
148	rs2244608	12	NO
149	rs2681472	12	YES
150	rs3782774	12	NO
151	rs6538177	12	NO
152	rs7965800	12	NO
153	rs11170820	12	NO
154	rs11617955	13	YES
155	rs11838776	13	YES
156	rs138668271	13	NO
157	rs1924981	13	NO
158	rs7994130	13	NO
159	rs9532984	13	NO
160	rs9591012	13	NO
161	chr14:75614504:I	14	NO
162	rs10139550	14	YES

Table continued on next page.

Table 2.8 – continued from previous page

No.	Variants	Chromosome	GWAS-significant
163	rs11850624	14	NO
164	rs7158691	14	NO
165	rs76465615	14	NO
166	rs11632899	15	NO
167	rs11636914	15	NO
168	rs12899265	15	NO
169	rs17581137	15	NO
170	rs2521501	15	NO
171	rs4468572	15	YES
172	rs56062135	15	YES
173	rs62017965	15	NO
174	rs7174745	15	NO
175	rs8042271	15	YES
176	chr16:75308440:D	16	NO
177	rs34434037	16	NO
178	rs4783977	16	NO
179	rs7500448	16	NO
180	rs8045120	16	NO
181	chr16:70153814:D	16	NO
182	rs11079536	17	NO
183	rs1638446	17	NO
184	rs35895680	17	NO
185	rs3760128	17	NO
186	rs61182333	17	NO
187	rs7212798	17	YES
188	rs72823056	17	NO
189	rs741782	17	NO
190	rs9914266	17	NO
191	rs12922	18	NO
192	rs3813127	18	NO
193	rs663129	18	YES
194	rs833509	18	NO
195	chr19:17857761:I	19	NO
196	chr19:41790086:D	19	NO
197	rs1964272	19	NO
198	rs4420638	19	YES
199	rs56289821	19	YES
200	rs7260187	19	NO
201	rs75437342	19	NO
202	rs8104311	19	NO
203	chr20:33591107:I	20	NO
204	chr20:47456892:I	20	NO
205	rs13734	20	NO
206	rs17000775	20	NO
207	rs6129767	20	NO
208	rs763475	20	NO
209	rs11911017	21	NO
210	rs2836633	21	NO
211	rs28451064	21	YES
212	rs133441	22	NO
213	rs9608859	22	NO
214	rs180803	22	YES

Source: Lerato Magosi (Farrall and Hopewell labs).

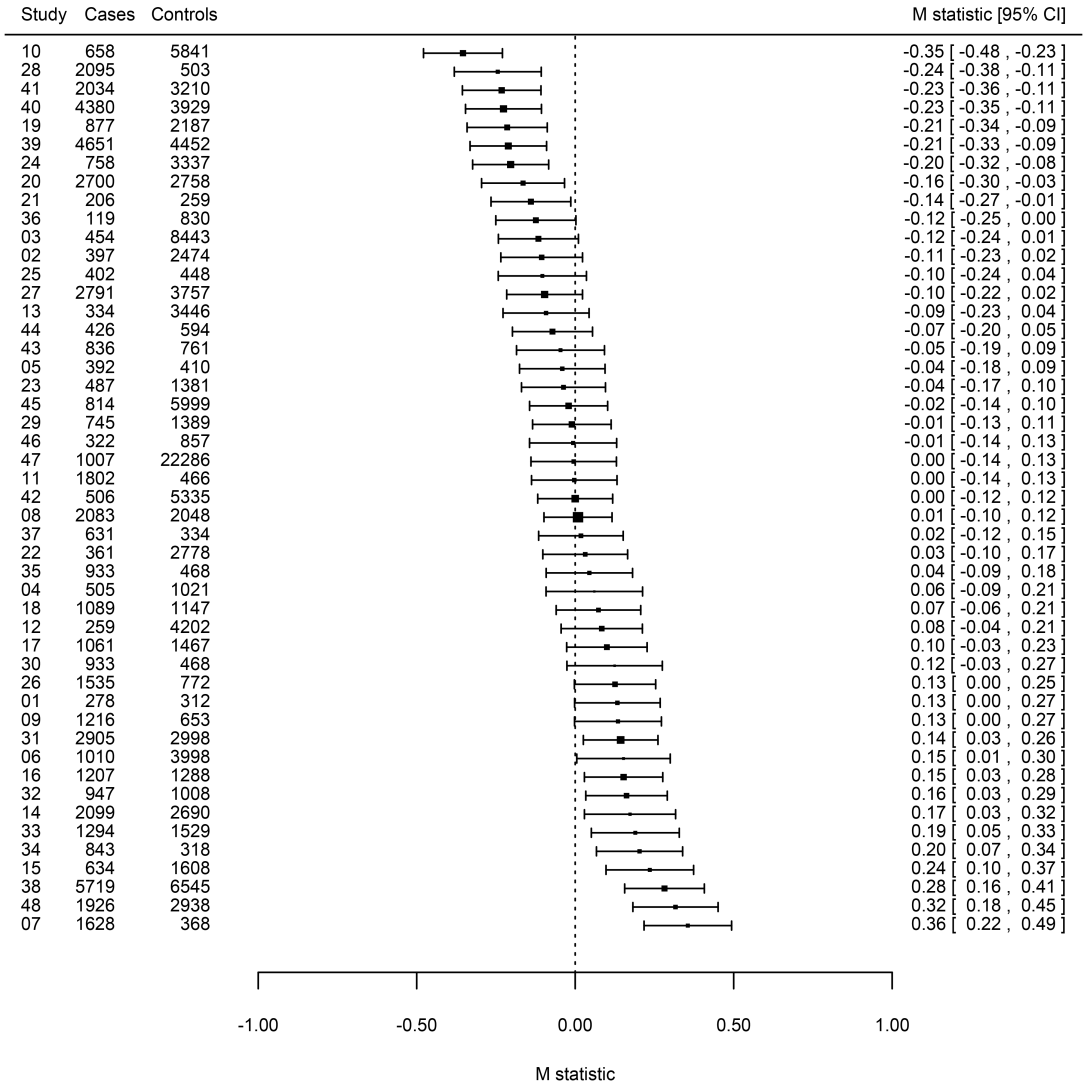


Figure 2.9: Forest plot of *M* statistics computed across 214 loci $(FDR < 5\%)$, summarizing systematic patterns of heterogeneity among studies in the CARDIOGRAMplusC4D GWAS meta-analysis. Sorted *M* statistics are presented for individual studies represented by filled squares with their 95% confidence intervals shown by horizontal lines; the sizes of the squares are proportional to each studies' inverse-variance weighting. Studies showing weaker (M $<$ 0) than average genetic effects can be distinguished from those showing stronger (M $>$ 0) than average effects. There was considerable systematic heterogeneity among studies in the CARDIoGRAMPlusC4D meta-analysis ($I^2 = 82.84\%$). Source: (Magosi et al., 2017).

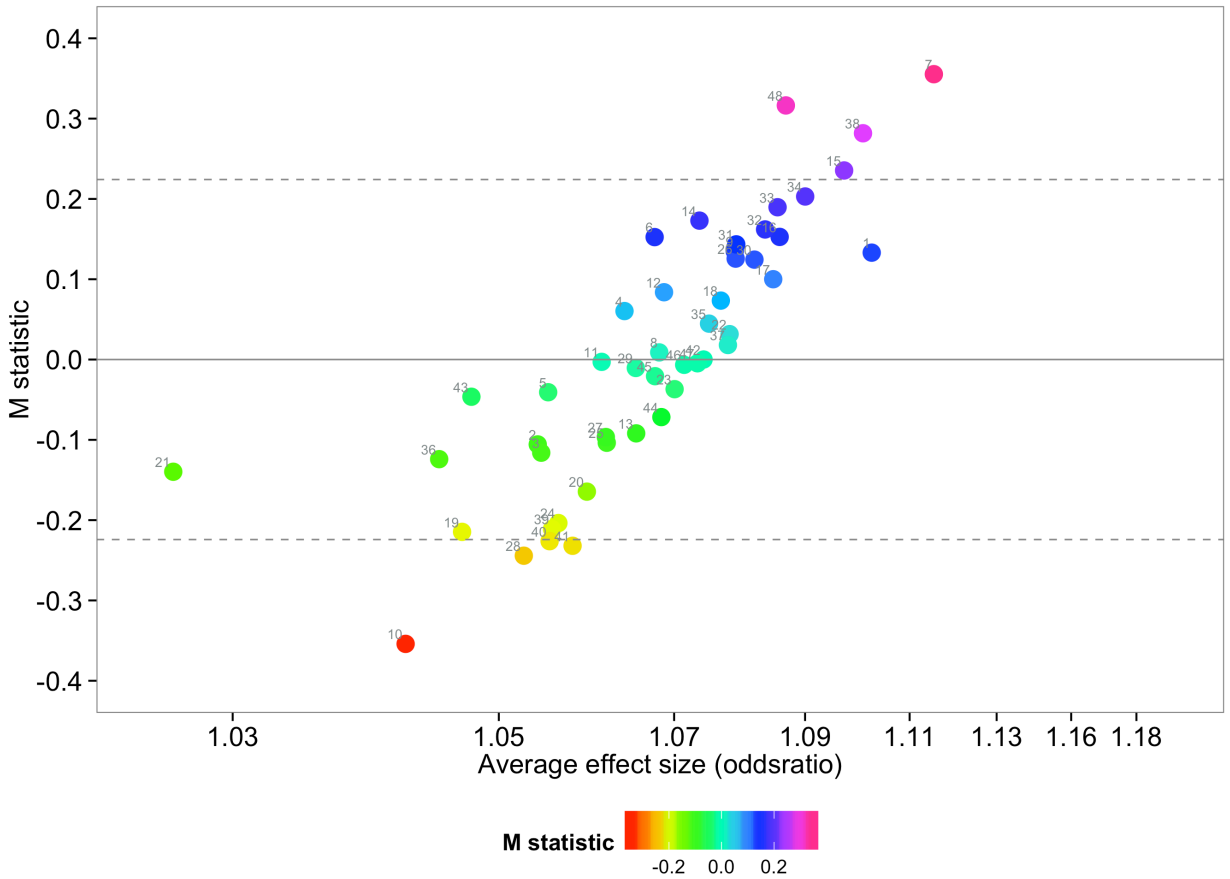


Figure 2.10: A scatterplot of *M* statistics computed across 214 FDR-defined loci $(FDR < 5\%)$ in the CARDIoGRAMplusC4D meta-analysis. *M* statistics for each study in the CARDIoGRAMplusC4D meta-analysis (Y- axis) are plotted against the average variant effect-size (expressed as odds ratios) (X-axis) in each study. A colour gradient was employed to highlight the distribution of *M* statistics. The dashed lines indicate the Bonferroni corrected 5% significance threshold (M = $\pm$0.224) to allow for multiple testing of 48 studies. Studies 7, 15, 38 and 48 show systematically stronger than average effects, whilst 10 and 28 consistently exhibit weaker than average effects across the loci examined. Source: (Magosi et al., 2017)

Table 2.9: Studies showing systematic heterogeneity patterns across 214 FDR-defined lead variants ($FDR \leq 5\%$) in the CARDIoGRAMplusC4D GWAS meta-analysis at, α = 0.05

Study identifier	# Variants	Bonferroni corrected p-value	M	Direction of effect
7	186	9.67E-06	0.355	stronger than average
15	214	2.75E-02	0.235	
38[a]	212	1.80E-03	0.282	
48[a]	214	1.77E-04	0.316	
9[b]	203	1.00E+00	0.135	
10[a]	214	1.07E-05	-0.354	weaker than average
28[a]	213	1.69E-02	-0.244	
40	211	4.49E-02	-0.226	
41	213	3.32E-02	-0.232	
19[b]	211	8.13E-02	-0.215	
24[b]	214	1.39E-01	-0.204	

A Bonferroni correction was applied for multiple testing of 48 studies. ***M*** (α = 0.05): $\pm$0.2241. Studies showing substantial systematic heterogeneity patterns both across the 214 FDR-defined lead variants ($FDR \leq 5\%$, $P < 5 \times 10^{-5}$) and the 46 GWAS significant loci ($P < 5 \times 10^{-8}$) in the CARDIoGRAMplusC4D meta-analysis[a]. Studies that showed statistically significant systematic heterogeneity patterns across the 46 GWAS significant loci but were statistically insignificant across 214 FDR-defined lead variants[b]. Source: (Magosi et al., 2017).

East Asian ancestry, earlier-onset CAD and positive CAD family-history confirmed as sources of systematic heterogeneity in the CARDIoGRAMplusC4D meta-analysis: A meta-regression of *M* statistics derived from the analysis of the 214 FDR-defined lead variants, confirmed the substantial contribution of variable ancestry, age-of-CAD onset and CAD family-history to systematic heterogeneity in the CARDIoGRAMplusC4D meta-analysis (Table 2.10); wherein East Asian ancestry, earlier-onset CAD and a positive family-history of CAD contributed to studies showing systematically stronger effects than average ($F_{3,44}$ = 9.47, P = 0.0001; adjusted R^2 = 44.8%). Given that family-history is an important risk factor for CAD and that studies which have a case ascertainment criteria including CAD family-history tend to be genetically enriched, it was unsurprising that family-history was a strong predictor of systematic heterogeneity in the CARDIoGRAMplusC4D meta-analysis.

Exploring the contribution of East Asian ancestry, earlier-onset CAD and positive CAD family-history to marker-specific heterogeneity at individual variants: Having identified East Asian ancestry, earlier-onset CAD and positive CAD family-history as the most informative predictors of systematic heterogeneity in the CARDIoGRAMplusC4D meta-analysis through a meta-regression of *M* statistics, I then explored their contribution to marker-specific heterogeneity at individual variants within the context of the classical (single-variant) meta-regression framework. Single variant meta-regressions of the observed effects (i.e. β coefficients scaled as $\log_e$(odds ratio)) at each of the 46 GWAS-significant lead variants were carried out employing East Asian ancestry, earlier-onset CAD and CAD family-history as joint predictors.

After a Bonferroni adjustment for multiple testing of 46 variants, evidence of stronger associations in studies with a positive CAD family-history and earlier-onset CAD was found at rs2891168 on chromosome 9 ($F_{3,44}$ = 6.71, P = 8 $\times 10^{-4}$; adjusted R^2 = 44.33%; I^2 = 42.47%).

Additionally, there was evidence of stronger associations in studies based on individuals of East Asian ancestry at rs6689306 on chromosome 1 ($F_{3,44}$ = 7.69, P = 3 ×10^{-4}; adjusted R^2 = 71.48%; I^2 = 7.68%), (Figures 2.11 and 2.12, Table 2.11).

Table 2.10: Meta-regression of *M* statistics computed across 214 FDR-defined lead variants in the CARDIoGRAMplusC4D GWAS meta-analysis (FDR $\leq 5\%$).

Variable	β	SE(β)	p-value
East Asian ancestry	0.08556	0.08361	0.312
early-onset	0.05546	0.05563	0.324
family-history	0.19565	0.05946	0.002

Family-history was the strongest predictor of systematically stronger genetic effects across the 214 FDR-defined loci examined in the CARDIoGRAMplusC4D meta-analysis. Source: (Magosi et al., 2017).

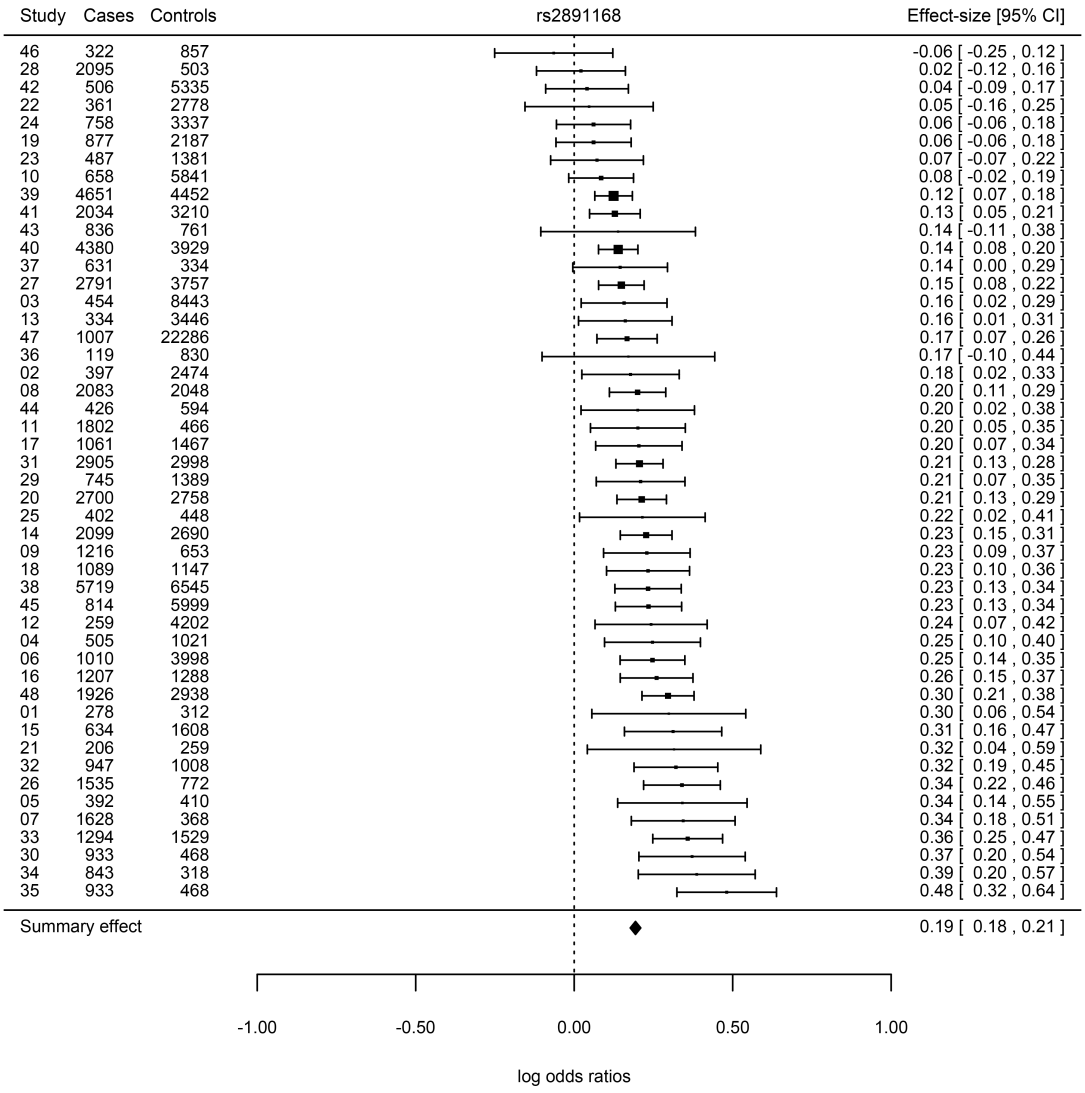


Figure 2.11: Forest plot of effect-size estimates (log odds ratios) at rs2891168 (chromosome 9) highlighting locus specific heterogeneity among studies in the CARDIOGRAMplusC4D GWAS meta-analysis ($I^2 = 57.69\%$). Sorted effect-size estimates are presented for individual studies represented by filled squares with their 95% confidence intervals shown by horizontal lines; the sizes of the squares are proportional to each studies' inverse-variance weighting. A filled diamond represents the summary effect-size estimate. Source: (Magosi et al., 2017).

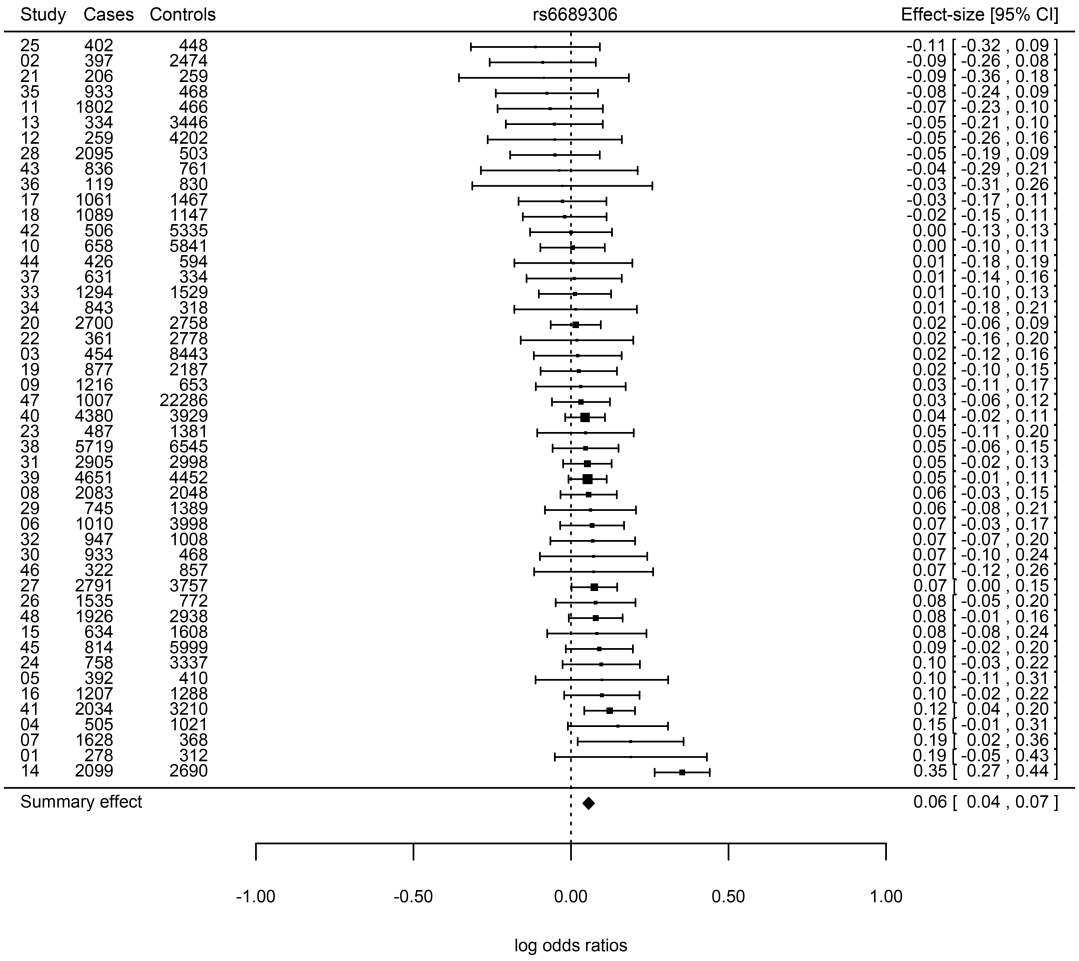


Figure 2.12: Forest plot of effect-size estimates (log odds ratios) at rs6689306 (chromosome 1) highlighting locus specific heterogeneity among studies in the CARDIOGRAMplusC4D GWAS meta-analysis ($I^2 = 40.57\%$). Sorted effect-size estimates are presented for individual studies represented by filled squares with their 95% confidence intervals shown by horizontal lines; the sizes of the squares are proportional to each studies' inverse-variance weighting. A filled diamond represents the summary effect-size estimate. Source: (Magosi et al., 2017).

Table 2.11: Meta-regression of effect-sizes in rs2891168 and rs6689306 employing East Asian ancestry, early age-of-CAD onset and family-history of CAD as covariates

Variable	β	SE(β)	p-value
rs2891168			
East Asian ancestry	-0.0092	0.04971	0.85374
early-onset	0.06785	0.03235	0.04175
family-history	0.06802	0.03517	0.05958
rs6689306			
East Asian ancestry	0.15679	0.03962	0.00027
early-onset	0.03238	0.02504	0.20281
family-history	-0.0259	0.02821	0.36421

A Bonferroni correction was applied for multiple testing of 46 lead variants. Earlier-onset CAD and a positive CAD family-history contributed to stronger genetic effects at rs2891168 (chromosome 9), whilst East Asian ancestry contributed to genetic enrichment at rs6689306 (chromosome 1). Source: (Magosi et al., 2017).

2.4 Discussion

2.4.1 Rationale for investigating systematic heterogeneity in GWAS meta-analysis

Heterogeneity investigations in GWAS meta-analyses have mostly been approached from a single-variant perspective; where Q, I^2 and τ^2 are commonly employed to measure random variation in the size and direction of genetic effects at individual variants (i.e. marker-specific heterogeneity). The search for common and low frequency variants underlying complex traits is increasingly being supported by large-scale international GWAS meta-analyses, based on studies conducted by independent research groups under varying conditions (e.g. variable study design features and ascertainment criteria). The principal design of GWAS meta-analysis is such that each component study contributes effect-size estimates for an array of variants to be examined; with this in mind, I hypothesised that diverse study characteristics among participating studies in a GWAS meta-analysis could induce genome-wide (systematic) heterogeneity patterns across genetic signals. Moreover, I hypothesised that current single-variant heterogeneity measures (e.g. the Q-statistic) would have lower power to diagnose systematic heterogeneity. Considering that excessive heterogeneity of genetic effects can compromise the power of meta-analyses to detect genetic signals, and the exploration of moderate heterogeneity can provide useful insight into the biology of complex traits; the primary focus of this chapter was to describe and evaluate a novel aggregate heterogeneity statistic, *M* that could be employed to identify outlier studies showing systematic heterogeneity patterns in GWAS meta-analysis and reveal their direction of effect.

2.4.2 Assumptions and properties of the *M* statistic

M statistics are formulated by aggregating standardized predicted random effects or *SPRE*s. These precision-weighted standardized residuals capture the direction and extent to which observed study effects at individual variants differ from the estimated average genetic effect. *M* statistics are computed within an inverse-variance weighted random-effects meta-analysis framework under the following assumptions: 1) each participating study submits study-level summary data for a common set of lead variants selected for heterogeneity analysis, that is effect-size estimates and their corresponding standard errors; 2) the lead variants examined are independently associated (i.e. in linkage equilibrium) with the trait of interest; 3) the observed study effects, random effects (sampling and between-study variance) and the mean underlying effect are normally distributed; and 4) the amount of between-study variance (heterogeneity) is estimated from the data. *M* analyses can be conducted using genetic association statistics from either binary or quantitative trait meta-analyses. For example, a meta-analysis based on case-control studies may use regression coefficients from a logistic regression as effect-size estimates (i.e. beta coefficients = log odds ratios of disease per-allele); and a meta-analysis of quantitative trait associations may use regression coefficients from a linear regression as the observed study effects (e.g. beta coefficients = mmHg per-allele in a blood pressure study).

Monte-Carlo simulation studies were employed to empirically evaluate the *M* statistic's null distribution, validity (type-1 error rate) and power (type-2 error rate) to identify outlier studies showing systematic heterogeneity patterns in genetic association meta-analyses. The results presented in aim 1 (i.e. section 2.3.1) demonstrate that *M* approximates a normal distribution under the null hypothesis of no systematic heterogeneity, for meta-analyses with 2

or more variants. GWAS meta-analyses typically examine an array of variants at a time, suggesting that the null distribution of *M* can be expected to be normal under most practical meta-analysis scenarios. *M*'s type-1 error rate was generally well controlled and depended on the number of participating studies in a meta-analysis. For meta-analysis scenarios with many studies (e.g. 30 studies), there was no evidence to suggest that *M*'s false positive rate differed from that expected at the 5% significance level. For meta-analyses with fewer contributing studies (e.g. 10 studies), *M*'s false positive rate was conservative i.e. smaller than that expected at the 5% significance level. Though conservative tests can suffer from a loss in power it is important to note that the type-1 error rate simulations considered a "challenging/worse-case" scenario, where *M* attempted to diagnose systematic heterogeneity patterns over a substantial background of marker-specific heterogeneity. In practice, the level of marker-specific heterogeneity in GWAS meta-analyses is seldom as high as (I^2 = 60.5%) for all variants examined, rather as shown in Table 2.4 the level of marker-specific heterogeneity tends to vary among genetic variants (0% < I^2 < 60%). *M*'s conservative false positive rate demonstrates its reliability as a measure to identify systematic heterogeneity patterns in genetic association meta-analyses.

M's power to identify systematic heterogeneity patterns was analytically assessed against that of the Q-statistic, a popular measure for detecting heterogeneity in meta-analyses. In all meta-analysis scenarios explored, *M* showed greater power than the Q-statistic to detect outlier studies showing systematically stronger or weaker effects than average. In line with the hypothesis made in section 2.1, the comparative power analysis between *M* and Q demonstrated that the Q-statistic has relatively low power to detect systematic heterogeneity patterns in genetic association meta-analyses. Aside from the fact that conventional heterogeneity metrics such as the Q-statistic are primarily designed to measure marker-specific heterogeneity patterns, the Q-

statistic's low power to diagnose systematic heterogeneity patterns across genetic signals might reflect the multiple testing burden associated with adjusting the Q-statistic for the number of variants examined (Zeggini and Ioannidis, 2009).

Furthermore, simulation studies revealed that *M*'s power to diagnose studies showing either systematically stronger or weaker effects than average increases with the number of variants examined, and is relatively independent of the number of contributing studies in a meta-analysis. Suggesting that the power to identify outliers in an *M* analysis could be optimized by including as many independently and strongly associated variants as available. Moreover, simulation studies showed that *M* could be optimally applied to meta-analyses with as few as 10 participating studies. *M*'s power to diagnose systematic heterogeneity patterns across genetic signals was also shown to be dependent on the level of marker-specific heterogeneity at individual variants; practically, this means that it will be easier to identify systematic heterogeneity patterns in meta-analyses with low-to-moderate marker-specific heterogeneity (e.g. $0\% < I^2 < 50\%$), (Figure 2.5).

A key difference between the properties of the *M* and Q statistics is that *M* depends on the number of independently associated variants examined whilst Q depends on the number of studies in a meta-analysis. As highlighted in the introduction (section 1.2.4.2), magnitudes of the Q, I^2 and H^2 metrics increase with the precision of studies in a meta-analysis; meaning that, two different meta-analyses which might have similar amounts of heterogeneity (τ^2), but varying within-study sample sizes can be expected to yield different values for Q, I^2 and H^2 (Rücker et al., 2008; Thompson and Higgins, 2002). Given that the magnitude of *M* values also depends on

the precision of studies included in a meta-analysis it should be noted that *M* quantitatively measures the impact of systematic heterogeneity in a meta-analysis rather than its extent.

2.4.3 Applying the *M* statistic to a meta-analysis of CAD risk

M analyses can be conducted using either the GWAS-significant lead variants in a meta-analysis or a known panel of variants independently associated with the trait of interest. To demonstrate the *M* statistic's utility in identifying systematic heterogeneity patterns, the *M* statistic was applied to the CARDIoGRAMplusC4D consortium's meta-analysis of CAD risk (CARDIoGRAMplusC4D Consortium, 2015); first in an *M* analysis of 46 GWAS-significant lead variants; and second in a polygenic *M* analysis of 214 FDR-defined lead variants comprising both genome-wide significant (i.e. $P < 5 \times 10^{-8}$) and suggestively associated variants (i.e. $P < 5 \times 10^{-5}$, FDR < 5%). Consistent with the hypothesis made in section 2.1, both *M* analyses identified substantial systematic heterogeneity in the CARDIoGRAMplusC4D meta-analysis (Figures 2.6 and 2.9); a large proportion of which could be explained by studies based on individuals with a positive CAD family-history, earlier-onset CAD and East Asian ancestry showing systematically stronger effects than average (Figures 2.8 and 2.11). The additional systematic outlier studies identified in the polygenic *M* analysis underscored observations made in the simulation studies where *M*'s power increased with the number of independently associated variants examined. The contribution of differences in ancestry, age-of-CAD onset and CAD family-history to variability of study effect-size estimates in the CARDIoGRAMplusC4D meta-analysis was also confirmed by single-variant meta-regressions of the 46 GWAS-significant lead variants.

The genetic enrichment in studies where individuals had a positive CAD family-history and/or earlier-onset CAD was in accordance with the multi-locus threshold liability model's predictions; wherein individuals with a family-history of disease and/or an early age-of-disease

onset are likely to have accumulated more risk loci than the general population (Risch, 2001; Watkins and Farrall, 2006). For example, taking into account background heritability, Risch demonstrated in a paper evaluating the impact of multiple affected siblings on genotype relative risks (GRR) (Risch, 2001), that GRR in case-control association studies employing unrelated controls were more inflated in multiplex sibships (represented by dizygotic twins concordant at a locus of interest) compared to sporadic disease cases (represented by singletons).
Background heritability refers to the polygenic continuum or additive variance from untested loci that contribute to disease susceptibility in addition to the locus of interest that is genotyped (tested) in a genetic association study.

As an exemplar, Risch considered an additive multi-locus liability threshold model of a single bi-allelic locus, where expected frequencies of the disease allele in the affected (singletons, concordant monozygote and dizygote twins) were compared with population allele frequencies. Assuming low background heritability (25%), disease prevalence of 0.14% (corresponding liability threshold, T = 3), and disease allele population frequency of 0.1 the increase in frequency of the disease allele in sporadic cases was 0.1 (corresponding disease allele frequency: 0.2) compared with 0.2 in concordant monozygote twins (corresponding disease allele frequency: 0.3) and 0.17 in concordant dizygote twins (corresponding disease allele frequency: 0.27). Notably, the degree of inflation of the disease allele frequency in multiplex sibships increased with decreasing levels of background heritability. The work done by Risch, suggested that genotype relative risks of individuals with a positive family-history of the trait of interest would be genetically enriched compared to those in sporadic disease cases.

Notably, the magnitude to which the average genetic effect size was enriched in the CARDIoGRAMplusC4D meta-analysis by the three predictors was: 14% for East Asian

ancestry, 15% for CAD family-history and 11% for earlier-onset CAD; though modest, it is anticipated that the observed genetic enrichment could lower the numbers of cases and controls required to detect an associated locus by up to 50%. One might speculate that the enriched genetic signal in the East Asian studies could be attributable to several factors including: 1) stronger LD between markers selected for heterogeneity analysis in the CARDIoGRAMplusC4D meta-analysis and CAD causal variants in East Asian populations; and 2) differential CAD rates owing to comparatively lower use of preventative therapies in East Asia versus other regions (e.g. Europe) (Chen et al., 2014; Yang et al., 2013).

Additionally, this work has demonstrated how pairwise Euclidean distances of summary allele frequencies at GWAS-significant lead variants can be easily used to infer ancestry information of participating studies in meta-analyses based on summary data; thereby affording an opportunity to crosscheck demographic information provided by contributing studies in genetic association meta-analyses. Pairwise Euclidean distances were preferred as a simple method to generate an abstract "map" of the genetic (summary allele frequency) data as they make no biological assumptions regarding mutation rates and genetic drift in contrast to Reynold's (Reynolds et al., 1983) and Nei's distances (Nei, 1978). Chen and colleagues have recently developed methods that employ F_{ST} genetic distances (F_{pc}) and principal components analysis (Meta-PC) to infer ancestry information from summary allele frequencies of independently associated variants (Chen et al., 2016). A key difference between the F_{pc} approach by Chen and colleagues and the manner in which pairwise Euclidean distances were used to infer genetic origin in the CARDIoGRAMplusC4D meta-analysis, was the creation of a square matrix where participating studies consecutively acted as the reference population. Whilst Chen and colleague's Meta-PC method of performing a principal component analysis of summary allele

frequencies gets around F_{pc}'s dependence on the choice of reference; Meta-PC shares a limitation with our method of pairwise Euclidean distances in that they both depend on the set of studies analysed, where substituting a few studies could alter genetic distances between studies and in turn their relative positioning on the inferred ancestry "map". Though sophisticated algorithms such as Meta-PC are available the pairwise Euclidean distance approach was favoured for its simplicity.

Other possible explanations for the residual heterogeneity in the CARDIoGRAMplusC4D meta-analysis might include: varying LD patterns, epistasis and gene-environment interactions. It would have been interesting to explore whether there were systematic effect-size differences attributable to gender in the meta-analysis. However, gender information was only available for a few participating studies in the CARDIoGRAMplusC4D meta-analysis.

2.4.4 Advantages and limitations of the *M* approach

The main advantage of exploring genome-wide heterogeneity patterns using *M* statistics is that they not only identify outlier studies showing systematic heterogeneity patterns but also reveal their direction of effect, that is whether the outliers show systematically stronger or weaker genetic signals. Practically, this can be useful in exploring: sexual dimorphism, gene-environment interactions (e.g. lifestyle variables contributing to systematically stronger or weaker genetic effects), ancestry, and sub-phenotype variability among genetic variants examined in a meta-analysis. *M* statistics can also be used to explore systematic heterogeneity in large prospective studies. For example, the UK Biobank prospective study comprising 500,000 individuals was conducted at 22 different centres in different regions of the UK (i.e. England, Wales and Scotland) (Littlejohns et al., 2017). As such it stands to reason that there would be systematic differences in: lifestyle variables, ancestries and clinical phenotypes that might result in stronger or weaker genetic effects. Treating genetic association analyses of the different centres as "independent studies", *M* statistics could be applied to identify systematic outliers and highlight variables preferentially associated with the trait under study in different ancestral sub-populations, age groups and genders. With respect to dissecting sub-phenotype variability, *M* statistics could be used to evaluate the power (via non-centrality parameters) gain or loss from performing meta-analyses with broad phenotypes versus separating them into their sub-types.

Another practical benefit of the *M* approach lies in across-study quality control of meta-analyses based on summary data. In particular, *M* analyses can reveal outlier studies showing consistently weaker than average effects across the majority of variants examined; the quality control criteria (e.g. allele frequencies, imputation quality, genotyping call rates, association

analyses models) of these outlier studies can then be reviewed to maximize power to detect genetic signals. In scenarios where the quality control issues of outlier studies might be challenging to address owing to restricted access to the individual-level genotype data; it might be useful to calculate the non-centrality parameters and in turn power of conducting the meta-analysis in the presence or absence of the outlying study.

M statistics are an aggregation of standardized predicted random effects (*SPRE*s) also known as internally studentized residuals. A limitation of basing *M* statistics on internally studentized residuals is that summary effect estimates for each lead variant are calculated using effect-size estimates of all participating studies in a meta-analysis; therefore, studies showing extreme effects can influence or skew summary effect estimates, such that the value of the summary effect-estimate is dragged towards the observed effect of the extreme outlier. This can make it harder to identify outlier studies. Leave-one-out methods such as externally studentized residuals (i.e. deleted residuals) can be used as an alternative approach to calculating *M* statistics. The principle behind deleted residuals is that the suspected outlier study is excluded from the calculation of the average genetic effect estimate so that the deleted (externally studentized) residual for the V^{th} variant in the S^{th} study, in a meta-analysis comprising *S* studies and *V* variants is computed as,

$$ESR_{sv} = \frac{\left(y_{sv} - \hat{\theta}_{v_{(-s)}}\right)}{\sqrt{Var\left(y_{sv} - \hat{\theta}_{v_{(-s)}}\right)}} \ ,$$

or equivalently,

$$ESR_{sv} = \frac{\left(y_{sv} - \hat{\theta}_{v_{(-s)}}\right)}{\sqrt{\sigma_{sv}^2 + \hat{\tau}_{(-s)v}^2 - s_{p_{(-s)v}}^2}} \ ,$$

where,

y_{sv} : genetic effect estimate of the v^{th} variant in the s^{th} study

$\hat{\theta}_{v_{(-s)}}$: average effect estimate of the v^{th} variant excluding the contribution of the s^{th} study

σ_{sv}^2 : sampling variance for the v^{th} variant in the s^{th} study

$\hat{\tau}_{(-s)v}^2$: estimate of between-study variance (heterogeneity) for the v^{th} variant excluding the contribution of the s^{th} study

$s_{p_{(-s)v}}$: standard error of prediction (fitted values) for the v^{th} variant excluding the s^{th} study and random effects.

Such that, $\left(y_{sv} - \hat{\theta}_{v_{(-s)}}\right)$ represents the raw residual, and $\sqrt{\left(\sigma_{sv}^2 + \hat{\tau}_{(-s)v}^2 - s_{p_{(-s)v}}^2\right)}$ the standard error of the raw residual (Viechtbauer and Cheung, 2010).

Notably, the choice of using internally studentized residuals or *SPRE*s instead of deleted residuals was out of convenience since *SPRE*s can be readily obtained using the metareg command in Stata ® (Harbord and Higgins, 2008).

A drawback of computing *M* statistics within a frequentist random-effects meta-analysis framework is that, estimates for τ^2 (the amount of heterogeneity at individual variants) will likely be imprecise in meta-analyses comprising few studies (e.g. $\leq$ 10 studies) and small within-study sample sizes. As such, it would be useful to explore the extent to which Bayesian approaches

with empirically informed heterogeneity priors could improve estimates of marker-specific heterogeneity, τ^2 in GWAS meta-analysis; this could provide an alternative framework in which *M* statistics could be computed for meta-analyses with small within-study sample sizes and few studies.

2.4.5 Alternative approaches

As an alternative to the *M* approach, systematic heterogeneity patterns in a meta-analysis could be explored by summing the squares of *SPRE* statistics of independently associated variants in each study; this would generate a cumulative heterogeneity statistic, say A^2, for each participating study in the meta-analysis. Notably, the squares of the *SPRE* statistics would be chi-squared with 1 degree of freedom since *SPRE*s are standard normally distributed statistics. Following the additive property of independent chi-squares, the null distribution of the cumulative heterogeneity statistic would be approximated by a chi-square distribution with v degrees of freedom (v = number of variants). The caveat with this approach, however, is that it does not indicate the direction of effect of the systematic heterogeneity patterns; rather it just reveals their presence, as such the *M* statistic was preferred instead.

Currently, the *M* approach employs independently associated (i.e. linkage equilibrium) variants to explore systematic heterogeneity patterns in genetic association meta-analyses. Going forward, it would be interesting to assess whether the power of *M* to detect systematic heterogeneity patterns could be improved by broadening the criteria of variants examined to include variants in weak LD. In particular, simulation studies could be employed to evaluate type 1 and 2 error rates for *M*, in meta-analysis scenarios where the panels of variants examined had varying levels of LD. A gradient of r^2 values could be explored to find suitable levels of LD, such that the genotypic correlation between any two variants in the panel of variants examined does not exceed a particular threshold. Ideally, the aim would be to find variants in weak LD, which are independently associated with the trait of interest, such an investigation was beyond the scope of this research program and was thus reserved for future study.

2.4.6 Overlapping approaches

Across-cohort statistics recently developed by Chen and colleagues (Chen et al., 2016) to examine heterogeneity in GWAS meta-analysis overlap with the *M* approach; in that Chen and colleague's pairwise heterogeneity statistic, λ_{meta} can be used to indirectly explore systematic heterogeneity patterns. λ_{meta} assesses the extent to which effect-size estimates for a panel of independently associated variants (i.e. in linkage equilibrium) agree between two participating studies. For each pair of studies in a meta-analysis, λ_{meta} is computed as the ratio between the median T values and the median of a chi-squared distribution with 1 degree of freedom (0.455). Across a pair of participating studies in a meta-analysis, T values are computed for each variant in a panel of independently associated variants by taking the ratio between the squared differences of effect-size estimates and the sum of their sampling variances. For a meta-analysis with *V* variants and *S* studies, the T value for the v^{th} variant between study 1 (s1) and study 2 (s2) is defined as,

$$T_v = \frac{(y_{1v} - y_{2v})^2}{\sigma_{1v}^2 + \sigma_{2v}^2},$$

where,

y_{1v}: genetic effect estimate of the v^{th} variant in study 1

y_{2v}: genetic effect estimate of the v^{th} variant in study 2

σ_{1v}^2 : sampling variance of the v^{th} variant in study 1

σ_{2v}^2 : sampling variance of the v^{th} variant in study 2.

Such that the λ_{meta} value for *s1* and *s2* across a panel of *V* variants is,

$$\lambda_{meta} = \frac{median(T)}{median(\chi_1^2)}.$$

$\lambda_{meta} > 1$ indicates systematic differences in the effect-size estimates reported by a pair of studies for a panel of variants examined (i.e. systematic heterogeneity). $\lambda_{meta} < 1$, points to the presence of correlated effect-size estimates owing to the presence of close relatives or samples that have been used in more than one study (i.e. overlapping samples) in the meta-analysis.

In contrast to the *M* statistic, which explores systematic heterogeneity simultaneously across all participating studies in a meta-analysis, λ_{meta} explores systematic heterogeneity on a pairwise study basis. Even though λ_{meta} is a useful method for detecting systematic heterogeneity patterns between pairs of studies in a meta-analysis, the number of combinations to explore in large meta-analyses can complicate the interpretation of the analysis. Furthermore, λ_{meta} provides little information of the direction of effect of systematic heterogeneity in meta-analyses. For example, applying the λ_{meta} statistic to the CARDIoGRAMplusC4D dataset highlighted larger systematic differences in effect-size estimates between the following pairs of studies (9 and 10: $\lambda_{meta} = 3.27$, $P = 3.38 \times 10^{-8}$), (10 and 14: $\lambda_{meta} = 3.34$, $P = 5.32 \times 10^{-8}$), (7 and 42: $\lambda_{meta} = 3.58$, $P = 1.86 \times 10^{-9}$) and (40 and 48: $\lambda_{meta} = 3.75$, $P = 2.51 \times 10^{-11}$), (Figure 2.13). However, it was challenging to compare systematic heterogeneity patterns across pairs of studies, in order to make overall inferences about outlier studies showing enhanced genetic effects (i.e. stronger than average genetic signals) compared to all other studies. Similarly it was challenging to identify outlier studies showing weaker than average effects compared to the rest of the participating studies. From the *M* analysis of the CARDIoGRAMplusC4D dataset we now know that studies 7, 9 and 48 were more genetically enriched than other studies in the meta-analysis, whilst studies 10 and 40 displayed weaker than average effects.

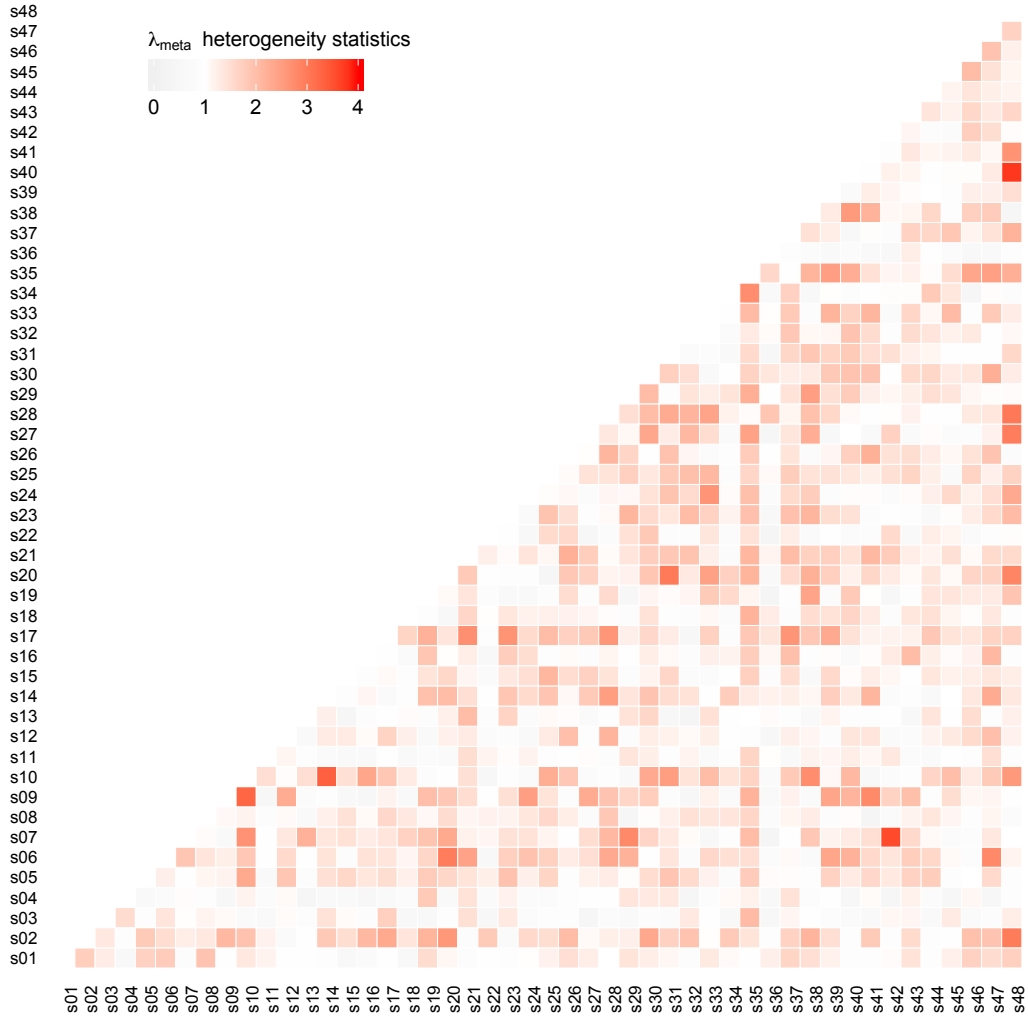


Figure 2.13: Heatmap of λ_{meta} heterogeneity statistics comparing effect-size estimates at 46 GWAS-signficant lead variants, across pairs of studies in the CARDIoGRAMplusC4D meta-analysis. Source: Lerato Magosi (Farrall and Hopewell labs).

2.5 Chapter summary

To summarize, this chapter draws attention to the existence of systematic (i.e. genome-wide) heterogeneity patterns across genetic signals and presents a new aggregate heterogeneity statistic (*M)* to quantitatively describe their impact in genetic association meta-analyses. There were three key findings in this chapter. First, *M*'s power to identify systematic heterogeneity patterns in genetic association meta-analyses outperformed that of conventional heterogeneity statistics, particularly the Q-statistic. Second, *M*'s power to diagnose outlier studies showing systematically stronger or weaker genetic effects than average depended on the number and strength of the independently associated variants examined. Third, systematic heterogeneity patterns were successfully identified in the CARDIoGRAMplusC4D meta-analysis of CAD risk; the systematic heterogeneity patterns could be partly explained by studies with East Asian ancestry, a positive CAD family-history and earlier-onset CAD showing systematically stronger effects than average. As a comprehensive strategy for exploring heterogeneity in genetic association meta-analyses, I would recommend investigating genome-wide (systematic) heterogeneity patterns as well as marker-specific heterogeneity patterns at individual variants.

This page is intentionally left blank.

Chapter 3

This chapter proposes the use of GWAS-derived empirical heterogeneity priors as a method to improve precision in estimation of average genetic effects and heterogeneity, in small genetic association meta-analyses.

3.1 Hypothesis/ Research needs

Estimates of summary effects and heterogeneity in meta-analyses comprising small numbers of studies are often imprecise (Thompson and Sharp, 1999), owing to the low power to detect heterogeneity (Borenstein et al., 2010). Typically, meta-analyses are performed under a frequentist framework (Cantor et al., 2010; Rhodes et al., 2016), where the extent of heterogeneity is commonly estimated using method-of-moments (DerSimonian and Laird, 1986) or restricted maximum likelihood (Harville, 1977) algorithms. However, when participating studies are few, it can be more advantageous to conduct meta-analyses under a Bayesian setting (Pullenayegum, 2011; Turner et al., 2015).

In contrast to frequentist approaches, Bayesian methods allow the analyst to include evidence (through informed priors) on the likely extent of heterogeneity in a meta-analysis; thereby, reducing the parameter space of possible heterogeneity values, and potentially improving the precision with which average genetic effects and between-study variance are estimated (Higgins and Whitehead, 1996). Furthermore, Bayesian random-effects meta-analyses naturally take into account the uncertainty surrounding estimates for both average genetic effects

and heterogeneity (through posterior distributions), whilst frequentist methods seldom acknowledge the variability associated with heterogeneity estimates (Smith et al., 1995; Turner et al., 2012).

Common scenarios where a GWAS meta-analysis might be performed on a small collection of studies include: the early stages of collaborative meta-analysis projects, where GWAS results are pooled as they become available from participating research groups (i.e. a cumulative meta-analysis); as well as niche GWAS investigations aimed at exploring a phenotype of interest in a specific ancestry group. A useful example is that of work done by Band and colleagues, where a meta-analysis of three cohorts was undertaken to identify variants predisposing individuals of Gambian, Kenyan and Malawian descent (i.e. African ancestry) to severe malaria (Band et al., 2013). Another scenario where meta-analyses might be based on smaller numbers of studies is of complex traits that have been interrogated relatively less frequently through GWAS. A notable example is that of endometriosis, where the largest meta-analysis of endometriosis performed to date only comprised 11 studies (Sapkota et al., 2017). This is in contrast to a recent meta-analysis of CAD risk which included 48 studies (CARDIoGRAMplusC4D Consortium, 2015).

Examples where informed heterogeneity priors have been used to improve inference on summary effects and heterogeneity, in meta-analyses conducted on a small number of clinical trials include: Higgins and Whitehead's Bayesian random-effects meta-analysis evaluating the use of beta-blockers versus sclerotherapy for the prevention of first bleeding in cirrhosis (Higgins and Whitehead, 1996). Higgins and Whitehead's meta-analysis was informed with an empirically derived inverse-gamma heterogeneity prior generated from 18 gastroenterology meta-analysis

datasets. Two methods were presented for the construction of the informed heterogeneity prior. The first involved calculating method-of-moments heterogeneity estimates for the 18 gastroenterology meta-analyses to yield an empirical distribution of heterogeneity estimates. Curve fitting was then used to identify the theoretical distribution that most closely matched the empirical distribution of method-of-moments heterogeneity estimates. The parameters of the best fitting theoretical distribution were then used as an empirically informed heterogeneity prior. The second method presented for formulating informed heterogeneity priors was more sophisticated, it entailed performing a hierarchical Bayesian random-effects meta-analysis of the 18 gastroenterology meta-analyses to generate a predictive heterogeneity distribution; the predictive heterogeneity distribution could then be used as an informed heterogeneity prior in the meta-analysis comparing the effects of beta-blockers and sclerotherapy. Predictive heterogeneity distributions are based on the Bayesian idea of exchangeability (Higgins et al., 2009); that is, meta-analyses performed in the "same" subject area are likely to have similar sources and amounts of heterogeneity, such that the level of heterogeneity in historical meta-analyses can be indicative of the extent of heterogeneity in future meta-analyses.

Hierarchical Bayesian meta-analyses differ from the conventional Bayesian meta-analysis framework in the way prior distributions are defined. In a standard Bayesian model the analyst typically provides parameters of a prior distribution as fixed values and is therefore not obliged to acknowledge any uncertainty associated with them. By contrast, parameters of prior distributions in a hierarchical Bayesian model (i.e. hyperparameters) are allowed to vary, and are themselves sourced from prior distributions called hyper-priors; as a result, hierarchical Bayesian meta-analysis models are able to take into account the uncertainty associated with parameters of prior distributions (Beaumont and Rannala, 2004).

Informed prior distributions for heterogeneity were also considered by Eleanor Pullenayegum in a small (i.e. 5 studies) Bayesian random-effects meta-analysis of clinical trials, evaluating antibiotics as a treatment for acute otitis media in children (Pullenayegum, 2011). Instead of declaring an independent prior distribution for heterogeneity that could be specified before conducting a meta-analysis, as Higgins and Whitehead had done (Higgins and Whitehead, 1996); Pullenayegum formulated a joint prior for heterogeneity and the average intervention effect, in which the heterogeneity prior was dependent on the magnitude of the intervention effect (Turner et al., 2012). One could posit that the choice of a joint prior was meant to reflect the relationship between the extent of heterogeneity in a meta-analysis and the summary intervention effect. To formulate the joint prior, Pullenayegum performed a hierarchical Bayesian random-effects meta-regression of 314 meta-analyses published in the Cochrane Database of Systematic Reviews between 2008-2009; on applying the lognormal joint prior distribution to the acute otitis media meta-analysis, Pullenayegum noted an improvement in the precision of credible intervals for the summary intervention effect and heterogeneity, as compared to when using vague (or reference) priors.

Other examples where the use of informed heterogeneity priors has helped to improve precision in the estimation of summary effects and heterogeneity in small meta-analyses include a series of papers by Higgins and colleagues (Rhodes et al., 2015; Turner et al., 2012; Turner et al., 2015); where hierarchical Bayesian random-effects models were used to generate predictive heterogeneity distributions that could be employed as 'off-the-shelf' priors, in meta-analyses of clinical trials with either binary or continuous outcomes.

GWAS meta-analyses interrogate large numbers of genetic variants, often on the order of millions; this provides a rich source of heterogeneity estimates that I hypothesise could be used to improve precision in the estimation of average genetic effects and heterogeneity in small Bayesian genetic association meta-analyses. In particular, empirical heterogeneity priors derived from heterogeneity estimates of variants examined in a GWAS meta-analysis could be used to refine estimates of average genetic effects and heterogeneity at (heterogeneous) variants suggestively associated with a trait of interest in the same dataset. Importantly, accurate estimates of the extent of heterogeneity and in turn average genetic effects are particularly relevant for variants that are likely to be employed in risk prediction.

The influence of prior distributions for heterogeneity tends to be greater on the results (posterior distributions) of smaller meta-analyses than in larger ones (Pullenayegum, 2011; Rhodes et al., 2015); this is because of the sparse amount of data in meta-analyses with few studies. Practically, the sensitivity of posterior distributions in smaller meta-analyses to the choice of prior, means that applying different heterogeneity priors to the same small meta-analysis dataset can yield divergent results (Shoemaker et al., 1999). Notably, empirically informed heterogeneity priors derived from GWAS meta-analyses could have the added advantage of reducing subjectivity in the choice of priors for Bayesian genetic association meta-analyses.

This chapter focuses on two specific aims to evaluate the utility of GWAS-derived empirical heterogeneity priors in improving precision of estimates for average genetic effects and heterogeneity in smaller genetic association meta-analyses:

Aim 1: Compare the precision of Bayesian and frequentist estimates for average genetic effects and heterogeneity using Monte-Carlo simulation studies; where the Bayesian random-effects meta-analyses have been informed with GWAS-derived empirical heterogeneity priors.

Aim 2: Demonstrate the utility of empirically informed heterogeneity priors in improving inference on average genetic effects and heterogeneity, through a re-appraisal of suggestively associated heterogeneous variants in the CARDIoGRAMplusC4D GWAS meta-analysis.

3.2 Methods

3.2.1 Bayesian statistical inference

Statistical inference is a process where random samples drawn from a population of interest are used to make judgements or inferences about the properties of that population (Beaumont and Rannala, 2004); the properties explored are typically parameters, which describe the distribution of the population of interest (e.g. mean and variance). Under a Bayesian setting, deductions about possible values for the parameters of interest are based on posterior probability distributions and credible intervals of those parameters; this is in contrast to the frequentist paradigm where inferences about parameters of interest are made using long-term frequency statements (e.g. p-values and confidence intervals) (Shoemaker et al., 1999).

The Bayesian approach to statistical inference is more direct compared to classical (frequentist) methods. Under a Bayesian setting, one assesses the probability or extent to which they favour certain parameter values. However, for frequentist inference, one evaluates the evidence against the null hypothesis in an imagined long-term frequency experiment (Shoemaker et al., 1999); the evidence being the proportion of replicates in the long-term frequency experiment, where the sample (or observed) value of the test statistic was as (or more) extreme than the critical value of significance set in the distribution of the test statistic under the null.

Bayesian inference is built around Reverend Thomas Bayes' rule, a formula which defines the evidence in favour of particular parameter values (or posterior distribution, $P(\Phi|D)$) for a parameter of interest as, the joint probability, $P(D, \Phi)$ i.e. $P(D|\Phi)\ x\ P(\Phi)$ of the likelihood, $P(D|\Phi)$ and prior, $P(\Phi)$ normalised by the marginal likelihood, $P(D)$. That is,

$$P(\Phi|D) = \frac{P(D|\Phi)\ x\ P(\Phi)}{P(D)}\ ,$$

where, D denotes the data (or observed random variables) and Φ represents the unobserved parameter of interest (Beaumont and Rannala, 2004).

Put simply, Bayes' rule is based on the idea that one starts out with a set of beliefs about plausible values for an unobserved parameter of interest; the set of beliefs represent the level of credibility assigned to each possible parameter value and are summarised in a distribution called the prior. As data is gathered about the parameter of interest through random sampling, the prior beliefs are updated by calculating posterior probabilities for the possible parameter values. Essentially re-allocating a larger share of the credibility to parameter values which are more probable given the observed data (Kruschke, 2014).

Solving Bayes' formula analytically can be challenging in a practical setting where Bayesian random-effects meta-analysis models are defined using several parameters, the difficulty typically lies in integrating over all possible values (i.e. the parameter space) of the parameters of interest to compute the marginal likelihood. Instead of computing posterior distributions analytically which would involve solving the marginal likelihood integral, one

popular approach for performing Bayesian random-effects meta-analyses is to use Markov Chain Monte-Carlo methods (e.g. Gibbs sampling) to approximate posterior distributions of parameters of interest (Higgins and Whitehead, 1996). Notably, Gibbs sampling has been implemented in a variety of software packages including: *WinBUGS* (Lunn et al., 2000) and the *R*® package, *rjags* (Kruschke, 2014). Markov Chain Monte-Carlo (MCMC) methods constitute a set of techniques, where complex distributions can be approximated by randomly collecting large representative samples from them.

Furthermore, MCMC methods are particularly useful for Bayesian inference as they can facilitate random representative sampling from non-normalised target distributions. Notably, the marginal likelihood in Bayes' rule is a normalising factor (Shoemaker et al., 1999) hence non-normalised target distributions are distributions that just include the joint probability of the likelihood and the prior, $P(D, \Phi) = P(D|\Phi)\ x\ P(\Phi)$. Consequently, MCMC methods preclude the need to calculate the marginal likelihood in Bayes' rule, and allow representative samples to be generated from an un-normalised posterior distribution (Kruschke, 2014). To collect samples from an un-normalised posterior distribution, MCMC methods employ a first order Markov process, that is, they generate random walks around the posterior distribution where the next step in the walk is only dependent on the current position and not on the previous ones; the resulting series of steps is commonly referred to as a Markov Chain. The aim is typically to collect a larger number of samples from areas where the bulk of the posterior distribution lies, in order to accurately estimate credible intervals and measures of central tendency (e.g. mean, median and mode). Popular measures for assessing whether Markov chains have produced satisfactory approximations of posterior distributions, in terms of representative sampling and accuracy include: trace and autocorrelation plots, Gelman-Rubin statistics and Monte-Carlo standard errors, Figure 3.1A-D.

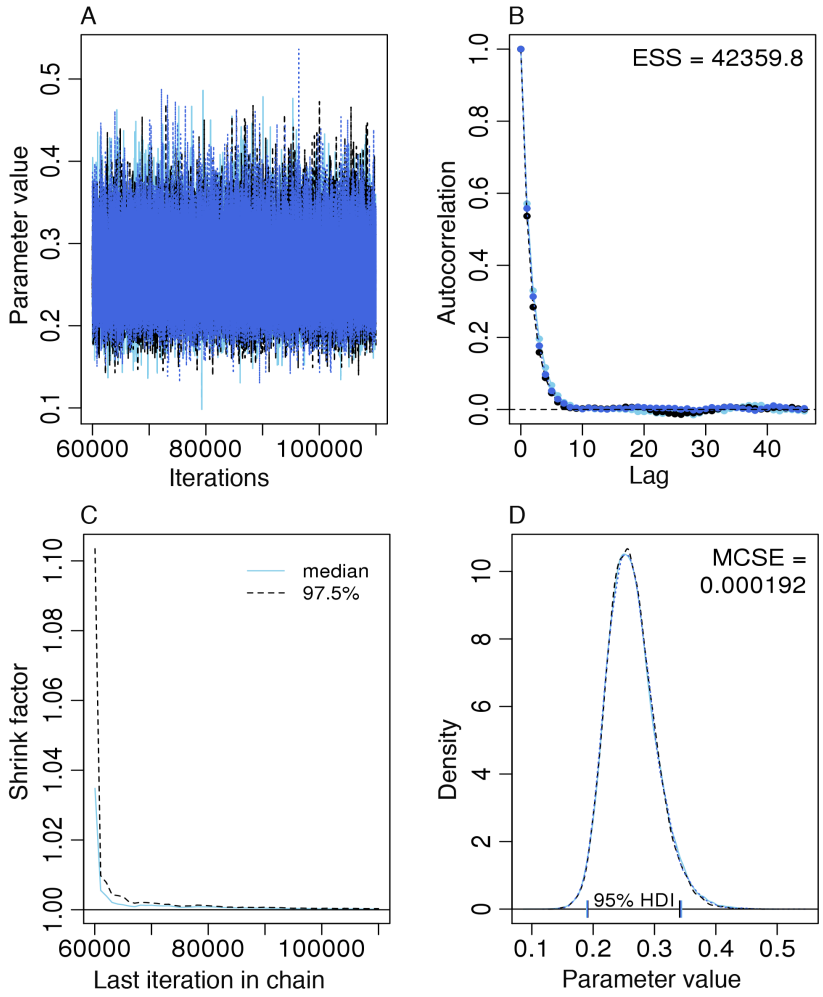


Figure 3.1: An illustration of Markov chain Monte-Carlo diagnostics following a Bayesian random-effects meta-analysis of 5 studies with an inverse gamma heterogeneity prior, $\tau^2 \sim IG(15.83, 1/0.01)$.The four panels show convergence diagnostics of log-transformed heterogeneity estimates at a variant of interest (i.e. $1/-ln(\tau^2)$). The convergence assessment is based on three parallel runs (i.e. 3 Markov chains), where each chain comprises 50,000 iterations of the Gibbs sampler, following a burin-in of 50,000 steps. The three Markov chains are denoted in black, sky blue and royal blue. Panel A shows a trace plot of sampled heterogeneity values (y-axis) as a function of step in the chain (x-axis). The chains are thoroughly overlapping suggesting there is no sign of failure to converge. Panel B shows an autocorrelation plot of sampled chain values with chain values K steps ahead, where K = lag. The amount of autocorrelation in the chains is marginal. Suggesting that successive steps in the chains are providing independent information about the posterior distribution of the between-study variance. Panel C is a plot of the Gelman-Rubin statistic (or potential scale reduction factor), which measures between-chain variance relative to within-chain variance. The shrink factor is near one for the majority of iterations suggesting the three chains have converged adequately. Panel D shows density plots of sampled heterogeneity values for each of the three chains. The substantial overlap among the density plots suggests the chains have converged and are producing representative samples from the posterior distribution for heterogeneity. ESS denotes the effective sample-size whilst MCSE represents the Monte-Carlo standard error. Source: Lerato Magosi (Farrall and Hopewell labs).

Trace plots illustrate sampled parameter values as a function of steps (or iterations) in a Markov chain. Bayesian meta-analyses models are usually run with several chains (e.g. three to ten chains) that have widely dispersed starting values (Sinharay, 2003). When there is representative sampling, Markov chains in a trace plot will mix thoroughly and overlap with each other, signalling that the chains have moved from their initial unrepresentative values and converged in the modal region of the posterior distribution. However, when there is unrepresentative sampling, trace plots will show some chains either being isolated from the rest or lingering for protracted periods at particular values. In practice it is more challenging to confirm representative sampling as chains can also converge in unrepresentative areas of the posterior distribution; thus, trace plots are usually used to diagnose unrepresentative sampling by MCMC methods.

Further to trace plots, Gelman-Rubin statistics are often used as a numerical check for convergence and representative sampling. Gelman-Rubin statistics compare the level of variance between Markov chains to that which occurs (across steps) within chains. A Gelman-Rubin statistic value of 1.0 generally indicates that there has been representative sampling and the chains have converged; meaning that the average difference between chains is the same as that within chains. Gelman-Rubin statistic values that are much larger than, 1.0 can signal a failure of the chains to converge, where some of the chains might have been orphaned or stuck at particular parameter values. In a practical setting, Gelman-Rubin statistic values ranging from 1.0 to 1.1 are deemed acceptable (Kruschke, 2014).

The accuracy of estimates for central tendency measures (i.e. the posterior mean, median or mode) and credible intervals is commonly assessed using autocorrelation plots and Monte-

Carlo standard errors. Ideally, successive steps in a Markov chain should provide independent information about the posterior distribution, so that larger samples yield more accurate estimates of the posterior distribution; in practice there can be some redundancy (i.e. autocorrelation) in the information provided by succeeding steps in a Markov chain. Consequently chains that are highly auto-correlated (or clumpy) require more steps than a non auto-correlated chain would to thoroughly explore the posterior distribution for a parameter of interest. In addition to revealing the level of "clumpiness" in a Markov chain, autocorrelation plots convey the effective sample size, which is a measure of the amount of independent information in the chain. Put differently, the effective sample size measures the number of steps a chain with zero auto-correlation would require to provide the same information as the current auto-correlated chain. Notably, the effective sample size is calculated by dividing the number of steps in a Markov chain by the amount of autocorrelation and is denoted as,

$$ESS = N \Big/ 1 + 2 \sum_{k=1}^{\infty} ACF(k) \ ,$$

where,

ESS: Effective sample size

N : Number of steps in a Markov chain (i.e. total sample size)

$ACF(k)$: Amount of autocorrelation (i.e. correlation between the current chain values with the chain values k steps ahead).

Generally, an effective sample size of at least 10,000 is suggested for accurate estimates of 95% credible intervals and posterior means (or medians) in unimodal distributions (Kruschke, 2014).

When judging the accuracy of estimates for credible intervals and measures of central tendency it is also useful to consider Monte-Carlo standard errors, especially since they are expressed on the same scale as the parameter of interest. Monte-Carlo standard errors are essentially an indication of how noisy MCMC estimates are, thus the lower the Monte-Carlo standard error the higher the accuracy of estimates for credible intervals and posterior means. Monte-Carlo standard errors are denoted by,

$$MCSE = SD/\sqrt{ESS}\,,$$

where SD is the standard deviation of the Markov chain and ESS the effective sample size.

3.2.2 A Bayesian random-effects model for genetic association meta-analysis

The manner in which Bayesian random-effects meta-analysis models are defined bears some similarity to classical (frequentist) random-effects models. In particular, both frameworks commonly assume that the observed effects in each study are drawn from a normal distribution, and that the underlying study effects are normal random variates from a grand population (Turner et al., 2015). Thus allowing inferences about the grand mean or summary effect to be generalizable beyond the collection of studies in the meta-analysis. Also analogous to the classical setting, Bayesian random-effects meta-analysis models generally treat sampling variances of participating studies as known despite the fact that they are estimated in each study from the data (Turner et al., 2015); an assumption that is more likely to hold when contributing studies in a meta-analysis have large within study sample sizes.

Whilst the aim of random-effects models under both frequentist and Bayesian frameworks is to make inferences about unknown parameters of interest, that is, to estimate the average genetic effect and amount of heterogeneity in a meta-analysis; a key difference between the two paradigms is that Bayesian random-effects models are defined to include evidence (in the form of priors) on the likely values for parameters of interest.

For a Bayesian random-effects meta-analysis, comprising *S* genetic association studies with summary data (i.e. genetic effect estimates and standard errors on the log odds ratio scale) for a single variant of interest, the observed genetic effect in the s^{th} study is described as:

$$\ln(OR) \sim N(\theta_s, W_s),$$

where $\theta_s \sim N(\mu, 1/\tau^2)$ and $W_s = 1/\sigma_s^2$.

Such that,

$\ln(OR)$: observed genetic effect in the s^{th} study

θ_s: underlying genetic effect in the s^{th} study

W_s: weight assigned to the s^{th} study

σ_s^2: sampling variance for the s^{th} study

μ: mean of the underlying distribution of study effects i.e. average genetic effect

$1/\tau^2$: precision of the underlying distribution of study effects

τ^2: amount of heterogeneity in the meta-analysis.

Notably, it is common and mathematically convenient for the scale parameters (i.e. σ_s^2 and τ^2) to be expressed as precisions (i.e. $1/\sigma_s^2$ and $1/\tau^2$) when conducting Bayesian meta-analyses (Rhodes et al., 2016). Prior distributions are then placed on the unknown parameters, μ and τ^2. It is customary to declare a flat (or widely diffuse normal) prior for the average genetic effect, e.g. $\mu \sim N(0,0.0001)$ (Lambert et al., 2005), and an informed prior for the extent of

heterogeneity, e.g. $\tau^2 \sim IG(shape = \alpha,\ rate = 1/\beta)$, where α and β denote the shape and scale parameters of an inverse-gamma heterogeneity prior (Higgins and Whitehead, 1996).

MCMC methods can then be employed to generate posterior distributions for the average genetic effect, μ and amount of heterogeneity, τ^2 in the meta-analysis. Notably, it is common to use posterior summaries (e.g. mean or median values) as point estimates for the average genetic effect and extent of heterogeneity.

3.2.3 Constructing a GWAS derived empirically informed heterogeneity prior

Consider a small GWAS meta-analysis project, comprising a series of S studies with summary-level data (e.g. genetic effect-size estimates and corresponding standard errors) for V variants. Let the summary-level data in each study be regression coefficients derived either from logistic or linear regression modelling. In addition consider that the amount of heterogeneity at each of the V variants has been estimated preliminarily by method-of-moments or restricted maximum likelihood to generate an empirical distribution of heterogeneity values. Proceeding under the assumption that genetic variants which confer effects of similar magnitude and are in linkage disequilibrium are also likely to have comparable heterogeneity levels, it stands to reason that some of the information conveyed by the empirical distribution of heterogeneity values might be redundant. Thus the first step in formulating the empirically informed heterogeneity prior is to ensure that the variants are in linkage equilibrium with each other.

The next step is to use maximum likelihood estimation to investigate the distributional form of the method-of-moments (or restricted maximum likelihood) heterogeneity estimates; this involves fitting typical heterogeneity distributions (e.g. lognormal, gamma, inverse-gamma or Weibull) to identify the theoretical distribution that most closely fits the empirical cumulative distribution function (ECDF) of the heterogeneity estimates. Notably, maximum likelihood estimation can be performed in Stata® using a set of curve fitting commands (i.e. lognfit, weibullfit, gammafit and invgammafit) developed by Nicholas J. Cox and Stephen Jenkins (Cox, 1998; Cox and Jenkins, 2011a, b; Jenkins, 2013). Given that the empirical distribution of the method-of-moments heterogeneity estimates is likely to be positively skewed such that most of

the variants show moderate to low heterogeneity levels whilst few exhibit higher amounts of heterogeneity, it can be useful to transform the method-of-moments heterogeneity estimates to the natural log scale prior to investigating their distributional form.

The best (or closest) fitting distribution can then be determined both by visual inspection and comparing Akaike information criteria (AIC) of the different heterogeneity distributions explored (Figure 3.2). Briefly, the Akaike information criterion is a model selection measure, which for a set of models can be used to evaluate the quality of each model relative to all other models. Notably, the Akaike information criterion can be calculated as follows:

$$AIC = -\frac{2(LL - k)}{n},$$

Such that,

LL: log-likelihood of the model

k: number of predictors in the model, where $2k$ is a penalty term adjusting for the number of predictors in the model

n: number of observations in the model.

The parameters of the best fitting distribution can then be used as an empirical heterogeneity prior, to inform Bayesian genetic association meta-analyses. For example, to re-appraise suggestively associated heterogeneous variants in the current GWAS meta-analysis

project, with the aim of refining estimates of both their average genetic effects and extent of heterogeneity.

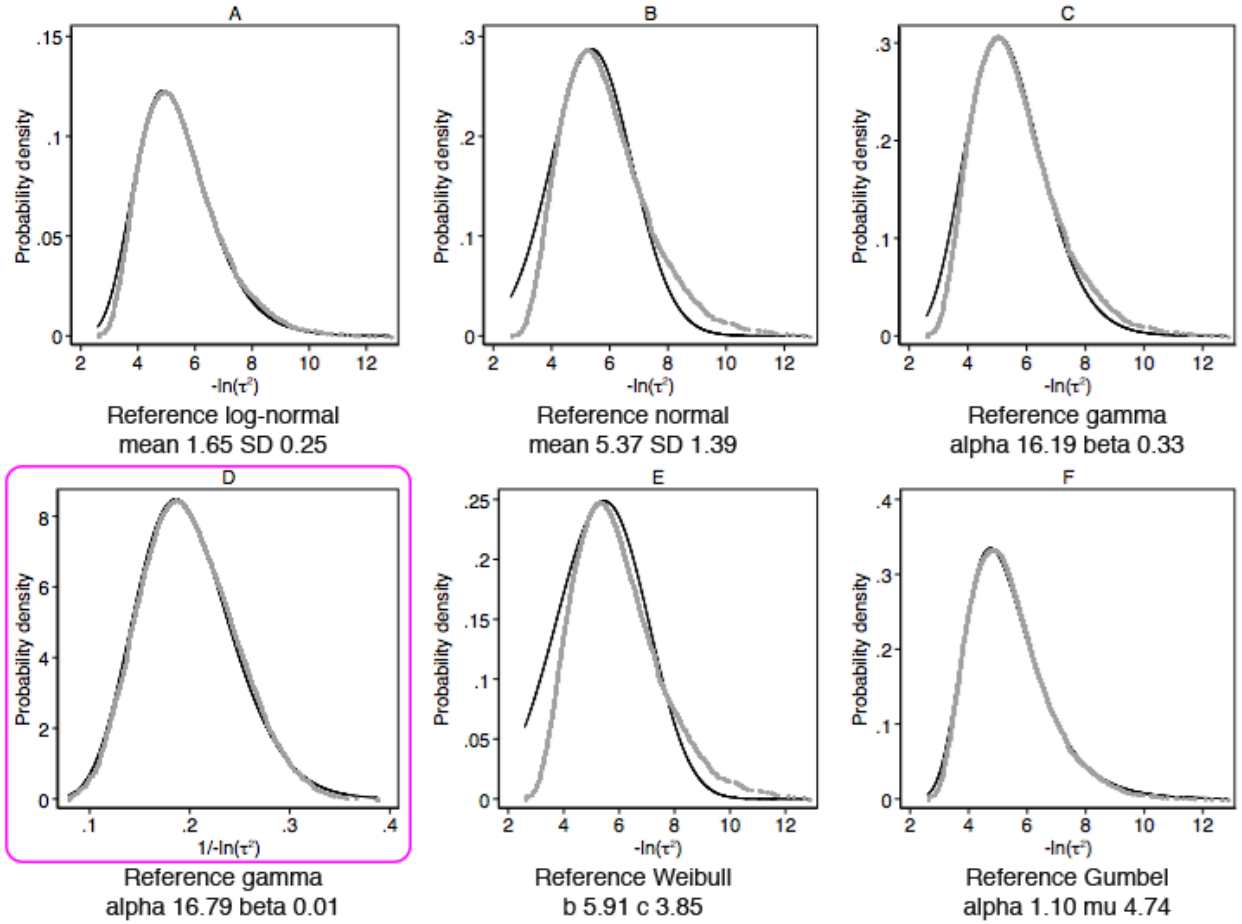


Figure 3.2: An illustration of the use of visual inspection and Akaike information criterion to identify the best fitting theoretical distribution to an empirical distribution of heterogeneity estimates.The six panels (from left to right) show comparisons of different theoretical distributions namely: log-normal (panel A), normal (panel B), gamma (panel C), inverse-gamma (panel D), Weibull (panel E) and Gumbel (panel F) to an empirical distribution of heterogeneity estimates.The heterogeneity estimates were sourced from a Monte-Carlo simulation experiment involving 1,000 replicates, where each replicate simulated a meta-analysis with 10 studies and 5 variants, yielding 50,000 variants altogether. The amount of heterogeneity, τ^2 at each variant was estimated by restricted maximum-likelihood and is scaled as log odds ratios. In each panel the theoretical distribution is denoted by a black curve whilst the empirical distribution is represented by a grey curve. Moreover, the parameters of the theoretical distributions explored were determined through maximum likelihood estimation. The best fitting distribution, that is the closest fitting theoretical distribution with the lowest Akaike information criterion (AIC) value, was the inverse-gamma distribution, $\tau^2 \sim IG(16.79, 1/0.01)$ (panel D). Notably, AIC values for the six panels are shown in Table 3.1. Source: Lerato Magosi (Farrall and Hopewell labs).

Table 3.1: AIC statistics assessing the fit of six theoretical distributions to an empirical distribution of heterogeneity estimates.

Distribution	AIC-unadjusted	AIC
Log-normal	115,516	3.35
Normal	120,646	3.49
Gamma	116,504	3.37
Inverse-gamma	115,211	3.34
Weibull	123,467	3.58
Gumbel	115,253	3.34

AIC-unadjusted denotes AIC statistics where the log-likelihood and penalty terms have not been adjusted for the number of observations in the model. Maximum likelihood estimation models for each of the six distributions included 34,520 observations out of the 50,000 heterogeneity estimates available. The 34,520 observations were variants with $\tau^2 > 0$ i.e. heterogeneous variants. Source: Lerato Magosi (Farrall and Hopewell labs).

3.2.4 Performing Bayesian and frequentist random-effects meta-analyses

Classical (frequentist) random-effects meta-analyses were performed using Harbord and Higgins's metareg command in Stata® versions 10.1 and 13.0 (Harbord and Higgins, 2008), where the extent of heterogeneity was estimated either by method-of-moments or restricted maximum likelihood; and Bayesian random-effects meta-analyses were carried out using Gibbs sampling in R® version 3.2.2 with the rjags package (Plummer, 2016).

3.2.5 Calculating credible intervals for Bayesian average genetic effect estimates

The Bayesian credible interval limits for estimates of average genetic effects and heterogeneity were taken as the 2.5% and 97.5% quantiles of their respective posterior distributions.

3.2.6 Computing coverage probabilities of average genetic effect estimates produced by Bayesian and classical random-effects meta-analyses

The comparative Monte-Carlo simulation studies of Bayesian and classical methods in Aim 1 of the results (section 3.3.1) explored four genetic association meta-analysis scenarios with different numbers of studies (5, 10, 15, 20). Within each of the four scenarios, coverage probabilities of Bayesian and classical average genetic effect estimates at each genetic variant were computed as the proportion of the 1,000 replicates where credible (or confidence) intervals contained the true value of the average genetic effect (i.e. $\log_e$(odds ratio) = 0.0953). Notably, the choice of the true value for average genetic effects in the four meta-analysis scenarios was inspired by the typical variant effect-size in a recent meta-analysis of CAD risk (CARDIoGRAMplusC4D Consortium, 2015).

3.2.7 Calculating the extent of bias in Bayesian and classical heterogeneity estimates

In each of the simulated genetic association meta-analysis scenarios, the extent of bias across the 1,000 replicates at each of the variants examined was calculated as the mean deviation of Bayesian (or classical) heterogeneity estimates from the simulated amount of heterogeneity. That is,

$$\frac{1}{1000}\sum_{i=1}^{1000}(\hat{\tau}_{iv}^2 - \tau_v^2)\,,$$

where,

$\hat{\tau}_{iv}^2$: estimated amount of heterogeneity at the v^{th} variant in the i^{th} replicate

τ_v^2: simulated (or true) amount of heterogeneity at the v^{th} variant.

3.2.8 Calculating mean squared errors associated with Bayesian and classical heterogeneity estimates

Within of each of the simulated genetic association meta-analysis scenarios, the mean squared error across the 1,000 replicates at each variant examined was computed as the average squared difference between Bayesian (or classical) heterogeneity estimates and the actual amount of heterogeneity. That is,

$$\frac{1}{1000}\sum_{i=1}^{1000}(\hat{\tau}_{iv}^2 - \tau_v^2)^2\,,$$

where,

$\hat{\tau}_{iv}^2$: estimated amount of heterogeneity at the v^{th} variant in the i^{th} replicate

τ_v^2: simulated (or true) amount of heterogeneity at the v^{th} variant.

3.3 Results

3.3.1 Aim 1: Simulation studies to compare the precision of Bayesian and frequentist estimates for average genetic effects and heterogeneity

Three criteria were employed to evaluate whether Bayesian random-effects meta-analyses with GWAS-derived empirical heterogeneity priors, could provide estimates for average genetic effects and heterogeneity that were more precise than those afforded by classical methods (e.g. method-of-moments and restricted maximum likelihood). The three criteria included: precision of summary effect intervals, accuracy of heterogeneity estimates and coverage probabilities of average genetic effect estimates.

Design of simulation studies: To evaluate inferences made by Bayesian and classical methods, four genetic association meta-analysis scenarios were simulated with different numbers of studies (5, 10, 15, 20) each involving 1,000 replicates. Moreover, each genetic association meta-analysis scenario comprised a panel of 5 variants with varying levels of heterogeneity; and the panel of variants examined was the same across all four scenarios (τ^2 = 0, 0.001, 0.002, 0.007, 0.020; corresponding I^2 values rounded to the nearest 10% = 0, 10, 20, 50, 70). To simplify the

simulations, effect-sizes (i.e. β coefficients) for all five variants were fixed at $\log_e$(odds ratio) = 0.0953 (i.e. odds ratio = 1.1). Furthermore, studies in each genetic association meta-analysis scenario were assigned equal weights (i.e. standard error of $\log_e$(odds ratio) = 0.08); this simplified calculations of the heterogeneity inconsistency index, I^2 at individual variants and provided a more optimal environment for the method-of-moments algorithm to measure the extent of heterogeneity.

The meta-analysis scenarios described above were chosen to reflect the different sizes of meta-analyses and varying levels of heterogeneity typically encountered in practical settings. In particular, heterogeneity values for variants examined in the simulation studies were based on background random heterogeneity levels $(0\% \leq I^2 \leq 60\%)$ of variants in a recent GWAS meta-analysis of CAD risk (CARDIoGRAMplusC4D Consortium, 2015). Similarly, the genetic effect-size and accompanying standard error assigned to variants in the simulation studies were the typical average genetic effect and median value of standard errors for variants in the CARDIoGRAMplusC4D consortium meta-analysis (CARDIoGRAMplusC4D Consortium, 2015).

Formulating empirically informed heterogeneity priors: Within each of the four genetic association meta-analysis scenarios, the amount of heterogeneity at individual variants, in each of the 1,000 replicates was preliminarily estimated by method-of-moments to yield an empirical distribution of heterogeneity estimates. For example, a meta-analysis scenario with 5 studies afforded an empirical distribution of 25,000 heterogeneity estimates (i.e. 5 studies x 5 variants x 1,000 replicates); the resulting empirical distribution of method-of-moments estimates was then used to construct an informed heterogeneity prior in accordance with section 3.2.3 of the

methods. Inverse-gamma distributions provided the best fit for the empirical heterogeneity distributions in all four scenarios; as such, Bayesian genetic association meta-analyses were informed by empirical inverse-gamma heterogeneity priors as summarised in Table 3.2.

Estimating summary effects and heterogeneity: In each of the four genetic association meta-analysis scenarios, Bayesian and classical random-effects meta-analyses were performed to estimate average genetic effects and heterogeneity at each of the 5 variants. Under the Bayesian setting, means of the posterior distributions were taken as point estimates for average genetic effects and heterogeneity.

Comparing the precision of average genetic effect estimates produced by Bayesian and classical random-effects meta-analyses: Credible and confidence intervals were computed to compare the precision of average genetic effect estimates, afforded by Bayesian and classical random-effects meta-analyses. Intervals for both Bayesian and frequentist average genetic effect estimates narrowed as the number of studies increased, Figure 3.3. Suggesting that the precision of average genetic effect estimates under both methods improves with increasing numbers of studies. This was expected as larger meta-analyses will generally have more information and therefore yield more precise estimates than meta-analyses based on few studies. Moreover Bayesian summary effect estimates were noticeably more precise than their frequentist counterparts with increasing levels of heterogeneity, especially in meta-analysis scenarios with few participating studies (i.e. $\leq$ 10 studies), suggesting that the empirically informed heterogeneity priors had a greater influence on the results of smaller genetic association meta-analyses than larger ones. This was unsurprising, as priors tend to have a greater influence on meta-analysis results when data is sparse (Pullenayegum, 2011). That said, the precision of

Bayesian and frequentist average genetic effect estimates was generally lower at variants with higher amounts of heterogeneity; this can be explained by the fact that highly heterogeneous variants will tend to have widely dispersed underlying study effects leading to imprecise heterogeneity estimates.

Altogether, these findings are generally consistent with simulation studies performed by Rhodes and colleagues, comparing frequentist properties of Bayesian and classical random-effects meta-analyses (Rhodes et al., 2016); where the precision of summary effect estimates was shown to increase with the number of studies and decrease with the extent of heterogeneity. Notable differences between simulation studies by Rhodes and colleagues and this work were the distinct scales of heterogeneity considered as well as the use of unequal weights for within-study variances. In particular the heterogeneity values examined in the simulation studies by Rhodes and colleagues were much larger than what would be expected in a typical complex trait GWAS meta-analysis (τ^2 = 0, 0.029, 0.069, 0.206, 1.302; expressed as log odds ratios). In addition, given that the DerSimonian and Laird's method-of-moments estimator for heterogeneity is based on Cochran's Q statistic (DerSimonian and Laird, 1986), and the Q-statistic is better powered to detect heterogeneity when the contributing studies in a meta-analysis are equally weighted (Hardy and Thompson, 1998), the use of unequal weights in the simulation studies conducted by Rhodes and colleagues would have had an impact on the power to detect heterogeneity and in turn affected the precision of summary effect estimates for the frequentist random-effects meta-analyses, especially when participating studies were few.

Table 3.2. Empirically informed inverse-gamma heterogeneity priors of four genetic association meta-analysis scenarios.

# Studies	# Variants	$\tau^2 \sim \text{IG}(\alpha, 1/\beta)$	
5	5	$\alpha = 15.829$	$\beta = 0.01316$
10		$\alpha = 16.794$	$\beta = 0.01178$
15		$\alpha = 17.924$	$\beta = 0.01086$
20		$\alpha = 17.879$	$\beta = 0.01083$

Four genetic association meta-analysis scenarios were simulated each based on 1,000 Monte-Carlo replicates. The meta-analysis scenarios are presented row-wise. Moreover, the amount of heterogeneity at variants in each scenario was first estimated by method-of-moments to yield an empirical distribution of heterogeneity estimates. For example, in the genetic association meta-analysis scenario based on 5 studies, an empirical distribution of 25,000 heterogeneity estimates (i.e. 5 studies x 5 variants x 1,000 replicates) was obtained. Empirically informed heterogeneity priors were then formulated by fitting inverse-gamma distributions to the empirical cumulative distribution functions of the method-of-moments heterogeneity estimates in each scenario. Source: Lerato Magosi (Farrall and Hopewell labs).

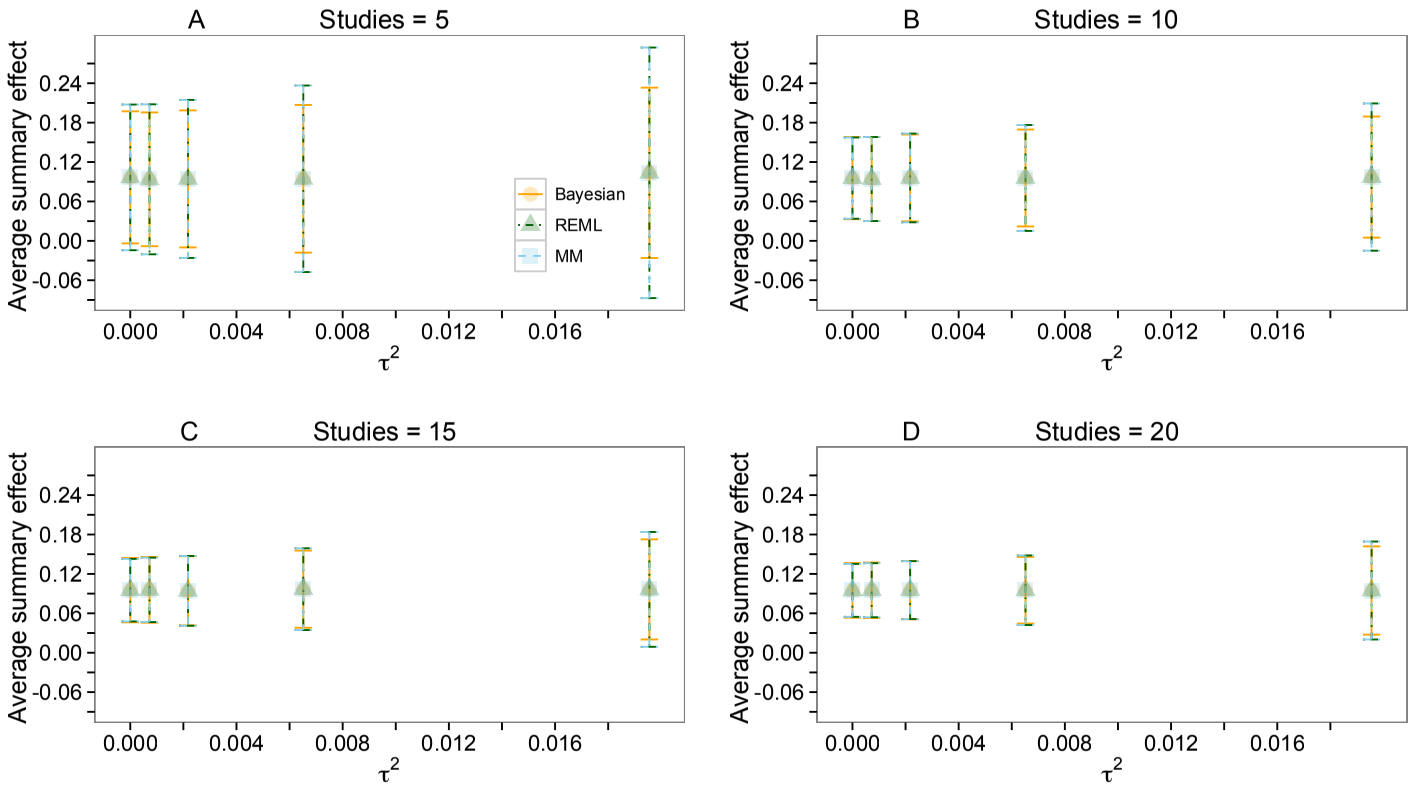


Figure 3.3: A comparison of the precision of Bayesian and frequentist average genetic effect estimates. The four panels (from left to right) show Monte-Carlo simulations of meta-analysis scenarios with 5, 10, 15 and 20 studies, each based on 1,000 replicates. All four scenarios examined the same set of 5 variants, where variant effect-sizes were fixed at log_e(odds ratio) $= 0.0953$ (i.e. odds ratio $= 1.1$). Corresponding standard errors were also held constant (standard error of log_e(odds ratio) $= 0.08$) such that studies were equally weighted in each meta-analysis scenario. However, heterogeneity levels were varied among the 5 variants (τ^2 scaled as log odds ratios $= 0, 0.001, 0.002, 0.007, 0.020$; corresponding I^2 values rounded to nearest $10\% = 0, 10, 20, 50, 70$). Values selected for variant effect-sizes, accompanying standard errors and heterogeneity levels in the simulation studies were based on the typical variant effect-size, extent of marker-specific heterogeneity and median value of standard errors of variants in a recent meta-analysis of coronary artery disease risk (CARDIoGRAMplusC4D, 2015). Furthermore, frequentist heterogeneity estimates were obtained using method-of-moments (MM) and restricted maximum likelihood (REML) algorithms; whilst Bayesian heterogeneity estimates were derived from random-effects models informed with empirical inverse-gamma heterogeneity priors shown in (Table 3.2). Moreover, Bayesian random-effects meta-analyses were performed using Markov chain Monte-Carlo methods, wherein results were based on 50,000 iterations of the Gibbs sampler following a burn-in period of 50,000 steps. Convergence was monitored with trace plots, Gelman-Rubin statistics and autocorrelation plots. Notably, Gelman-Rubin statistic values among replicates of the meta-analysis scenarios explored ranged from 1.00 to 1.06; this was in the conventional acceptable range of 1.00 to 1.10 (Kruschke, 2014), hence no replicates were excluded from analysis for failure to converge. Each data point represents the average summary effect estimate at an individual variant, that is the average of average genetic effect estimates across the 1,000 replicates in the simulation. Bayesian average summary effect estimates and credible intervals are shown as filled circles and solid lines, respectively. Whilst those for classical methods are represented by filled triangles and "dotdash" lines for REML estimates, and filled squares and dashed lines for MM estimates. In all four meta-analysis scenarios explored, Bayesian average effect estimates were generally more precise than their frequentist counterparts. The precision of both Bayesian and frequentist average genetic effect estimates improved with increasing numbers of studies. Source: Lerato Magosi (Farrall and Hopewell labs).

Comparing the accuracy of heterogeneity estimates produced by Bayesian and classical random-effects meta-analyses: In each of the four genetic association meta-analysis scenarios, estimates of heterogeneity afforded by Bayesian and classical random-effects meta-analyses were compared to the actual amount of heterogeneity in the panel of variants examined, Figure 3.4. Both approaches overestimated the extent of heterogeneity when there was little to no dispersion in the true genetic effects (i.e. $\tau^2 = 0$); the positive bias was particularly evident when there were few participating studies (i.e. $\leq$ 10 studies). From a Bayesian perspective, the use of a strictly positive empirical heterogeneity prior (i.e. $\tau^2 > 0$) might have contributed to the upward bias in heterogeneity estimates when the actual amount of heterogeneity was minimal (Thompson and Sharp, 1999); this can be mitigated by including $\tau^2 = 0$ values in the empirical heterogeneity priors. However, with regards to the frequentist setting, it has been suggested that the truncation of negative heterogeneity estimates to zero by method-of-moments and restricted maximum likelihood algorithms, introduces a positive bias in heterogeneity estimates when there are marginal amounts of heterogeneity (i.e. $\tau^2 = 0$) (Viechtbauer, 2005).

The extent of bias for both Bayesian and classical heterogeneity estimates depended on the number of studies, Figure 3.5. Heterogeneity estimates in smaller meta-analyses were substantially more biased than those in larger ones, suggesting that the accuracy of heterogeneity estimates for both Bayesian and classical random-effects meta-analyses improved as the number of studies increased. This was not surprising as larger meta-analyses have more data points (i.e. studies) to estimate the dispersion in true genetic effects compared to meta-analyses comprising few studies. The direction of the bias was positive for both methods when there were moderate to low heterogeneity levels (i.e. $\tau^2 \leq 0.008$, corresponding I^2 value $\leq 55.56\%$), and negative

for the Bayesian approach when there were higher amounts of heterogeneity (i.e. $\tau^2 \geq 0.010$, corresponding I^2 value $\geq 60.98\%$).

Consistent with Bayesian methods accounting for the imprecision in estimates of between-study variance (Rhodes et al., 2016), Bayesian heterogeneity estimates were relatively larger than their frequentist counterparts when there was little heterogeneity, and comparatively smaller when there were higher amounts of heterogeneity, Figure 3.4. Furthermore, Bayesian heterogeneity estimates were more accurate in regions where values of the actual amount of heterogeneity were near the centre of the prior distribution and less so towards the tails, Figure 3.6. This was to be expected, as the empirical heterogeneity priors employed were "symmetric", such that the bulk of the information (i.e. most credible heterogeneity values) was concentrated around the centre of the distribution.

Mean squared errors were calculated in each of the four meta-analysis scenarios to evaluate the efficiency with which Bayesian and classical methods estimate heterogeneity. Mean squared errors are a measurement of the average squared deviation of heterogeneity estimates from the actual (or true) amount of heterogeneity. It is desirable to have mean squared error values near zero, as this reflects minimal bias and/or sample fluctuation (Trikalinos et al., 2013). Notably, sample fluctuation in a meta-analysis simulation can be described as the extent to which between-study variance estimators assume different values in different replicates (Veroniki et al., 2016). In all four meta-analysis scenarios explored, Bayesian heterogeneity estimates had smaller mean squared errors than their frequentist counterparts, Figure 3.7, suggesting that the Bayesian approach was more efficient at estimating heterogeneity in the meta-analysis scenarios explored. Moreover, mean squared errors for both Bayesian and classical methods were dependent on the number of studies; whereby, larger meta-analyses had lower mean squared error values than

meta-analyses with few participating studies. This was consistent with the result in Figure 3.5, where the level of bias decreased with increasing numbers of studies.

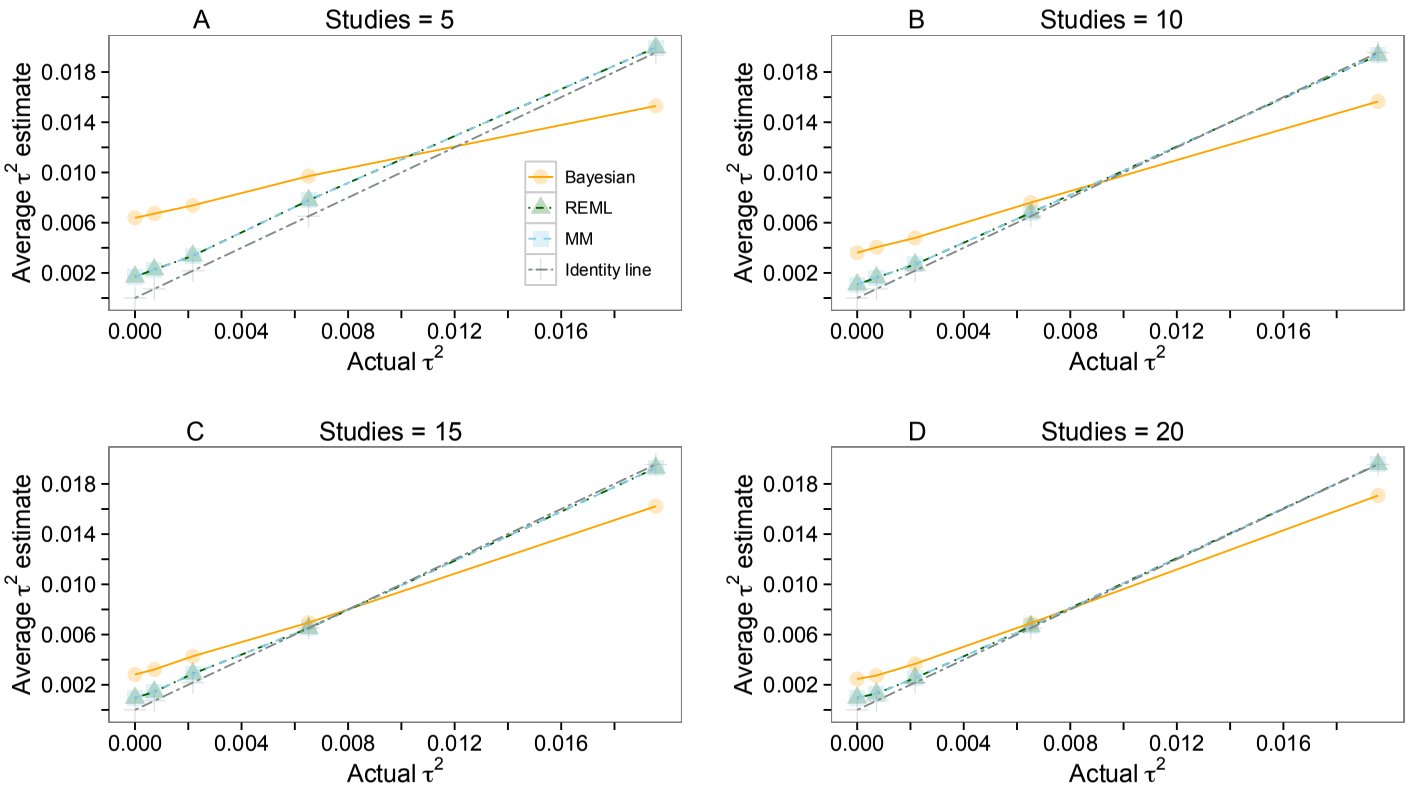


Figure 3.4: A comparison of the accuracy of Bayesian and frequentist heterogeneity estimates. The four panels (from left to right) show Monte-Carlo simulations of genetic association meta-analysis scenarios comprising 5, 10, 15 and 20 studies, each based on 1,000 replicates. The meta-analysis simulations were generated in the same manner as that in Figure 3.3. Each data point represents the average heterogeneity estimate for an individual variant, across the 1,000 replicates in the simulation. The identity line provides a reference for the actual (simulated) amount of heterogeneity in each meta-analysis scenario, and solid crosses denote its data points. Bayesian heterogeneity estimates are represented by filled circles; whilst method-of-moments and restricted maximum likelihood estimates are denoted by filled squares and triangles, respectively. There was a positive bias in both Bayesian and frequentist heterogeneity estimates when the simulated amount of heterogeneity was marginal (i.e. $\tau^2 = 0$). The accuracy of both Bayesian and classical heterogeneity estimates improved with increasing numbers of studies. Source: Lerato Magosi (Farrall and Hopewell labs).

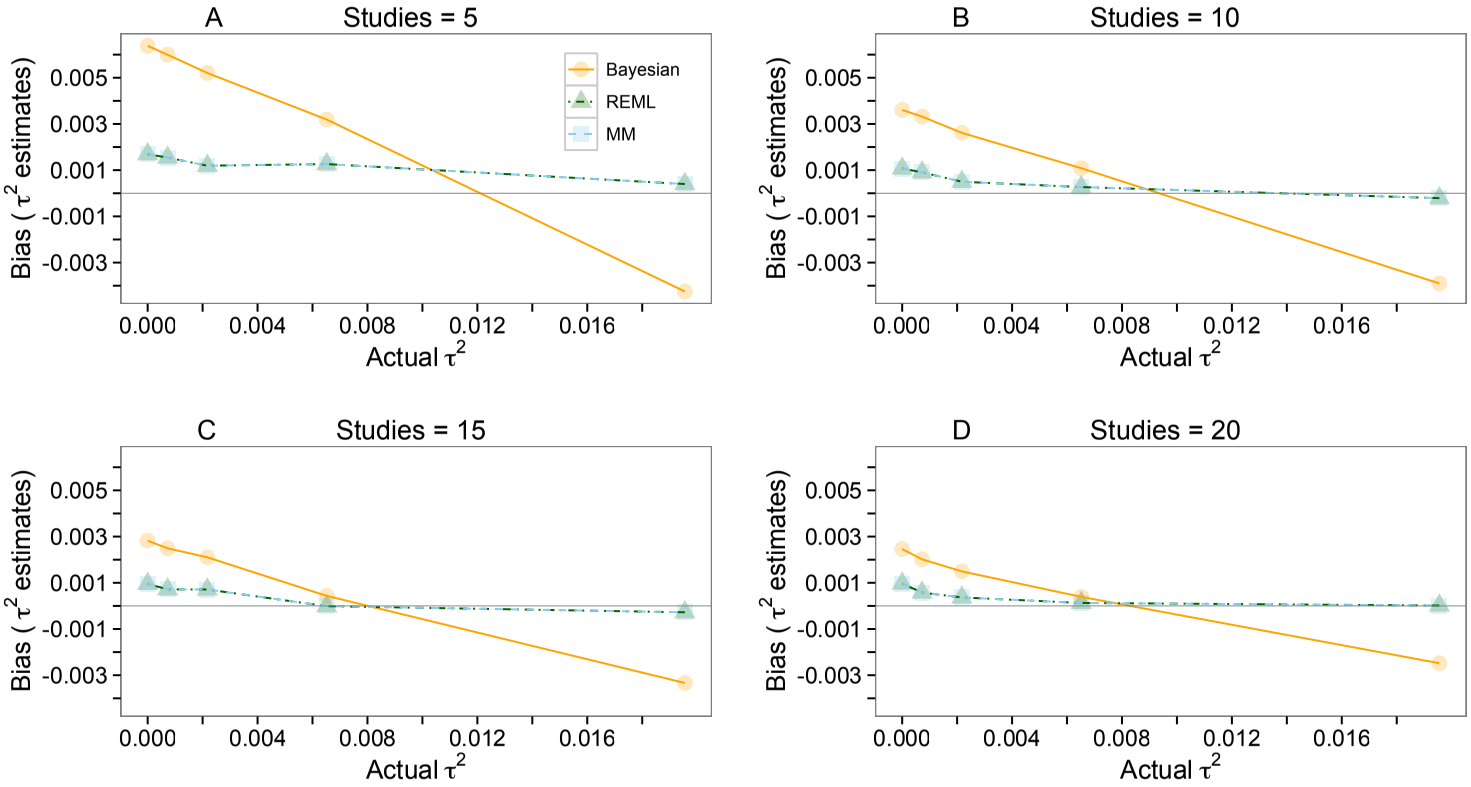


Figure 3.5. A comparison of the extent of bias in heterogeneity estimates produced by Bayesian and classical random-effects meta-analyses. The four panels (from left to right) show Monte-Carlo simulations of genetic association meta-analysis scenarios with 5, 10, 15 and 20 studies, each based on 1,000 replicates. The meta-analysis simulations were generated in the same manner as that in Figure 3.3. Each data point represents the extent of bias at an individual variant for the 1,000 replicates in the meta-analysis simulation. The extent of bias in the Bayesian heterogeneity estimates is denoted by filled circles; whilst the level of bias in restricted maximum likelihood and method-of-moments heterogeneity estimates is represented by filled triangles and squares, respectively. Both approaches (Bayesian and frequentist) generally overestimated heterogeneity when there were low to moderate amounts of heterogeneity (i.e. $\tau^2 \leq 0.008$, corresponding $I^2 \leq 55.56\%$). In addition, the Bayesian approach tended to underestimate heterogeneity when the extent of heterogeneity was relatively high (i.e. $\tau^2 \geq 0.010$, corresponding $I^2 \geq 60.98\%$). The extent of bias in both Bayesian and frequentist heterogeneity estimates generally reduced with increasing numbers of studies. Source: Lerato Magosi (Farrall and Hopewell labs).

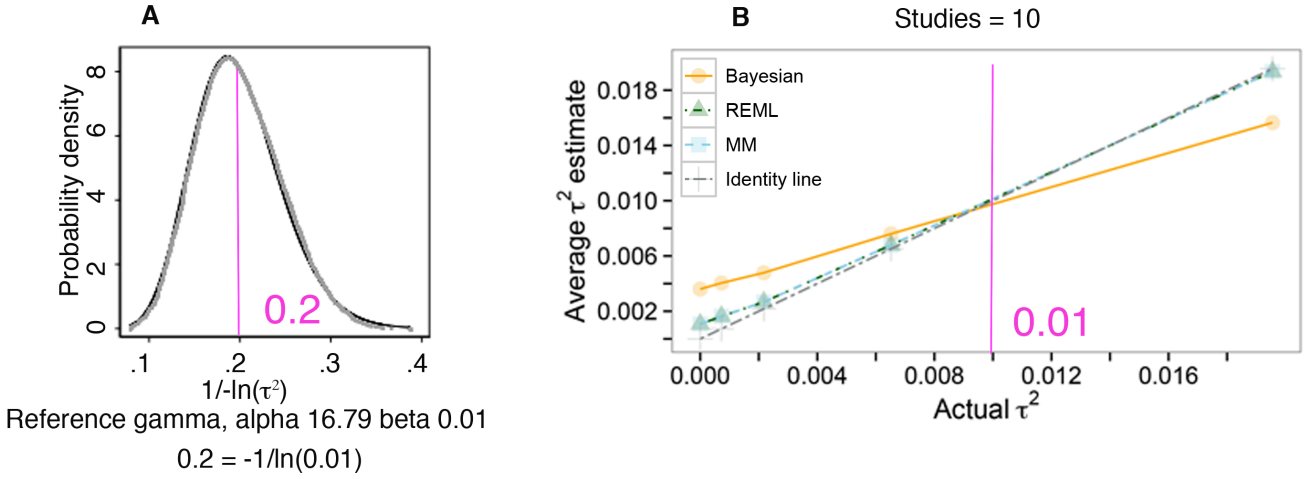


Figure 3.6: Accuracy of Bayesian heterogeneity estimates is greater near the centre of the prior distribution for heterogeneity. Panel A shows an empirically informed inverse-gamma heterogeneity prior for a genetic association meta-analysis simulation with 10 studies and 5 variants. The simulation comprised 1,000 Monte-Carlo replicates and was conducted in the same manner as that in Figure 3.3. Panel B shows a comparison of the accuracy of Bayesian and classical heterogeneity estimates. Where Bayesian random-effects meta-analyses at each of the 5 variants were informed with the empirical heterogeneity prior in panel A. By contrast, classical heterogeneity estimates were computed using restricted maximum likelihood and method-of-moments algorithms. Each data point in panel B represents the average heterogeneity estimate for an individual variant, across the 1,000 replicates in the simulation. Furthermore, the identity line provides a reference for the actual (simulated) amount of heterogeneity at individual variants and is denoted by solid crosses. Bayesian heterogeneity estimates are shown as filled circles; whilst method-of-moments and restricted maximum likelihood estimates are denoted by filled squares and triangles, respectively. Notably, Bayesian heterogeneity estimates were more accurate when values of the actual amount of heterogeneity (e.g. $\tau^2 = 0.01$) were near the dense region (e.g. $-1/ln(\tau^2) = 0.2$) of empirical heterogeneity prior and less so towards the tails. Source: Lerato Magosi (Farrall and Hopewell labs).

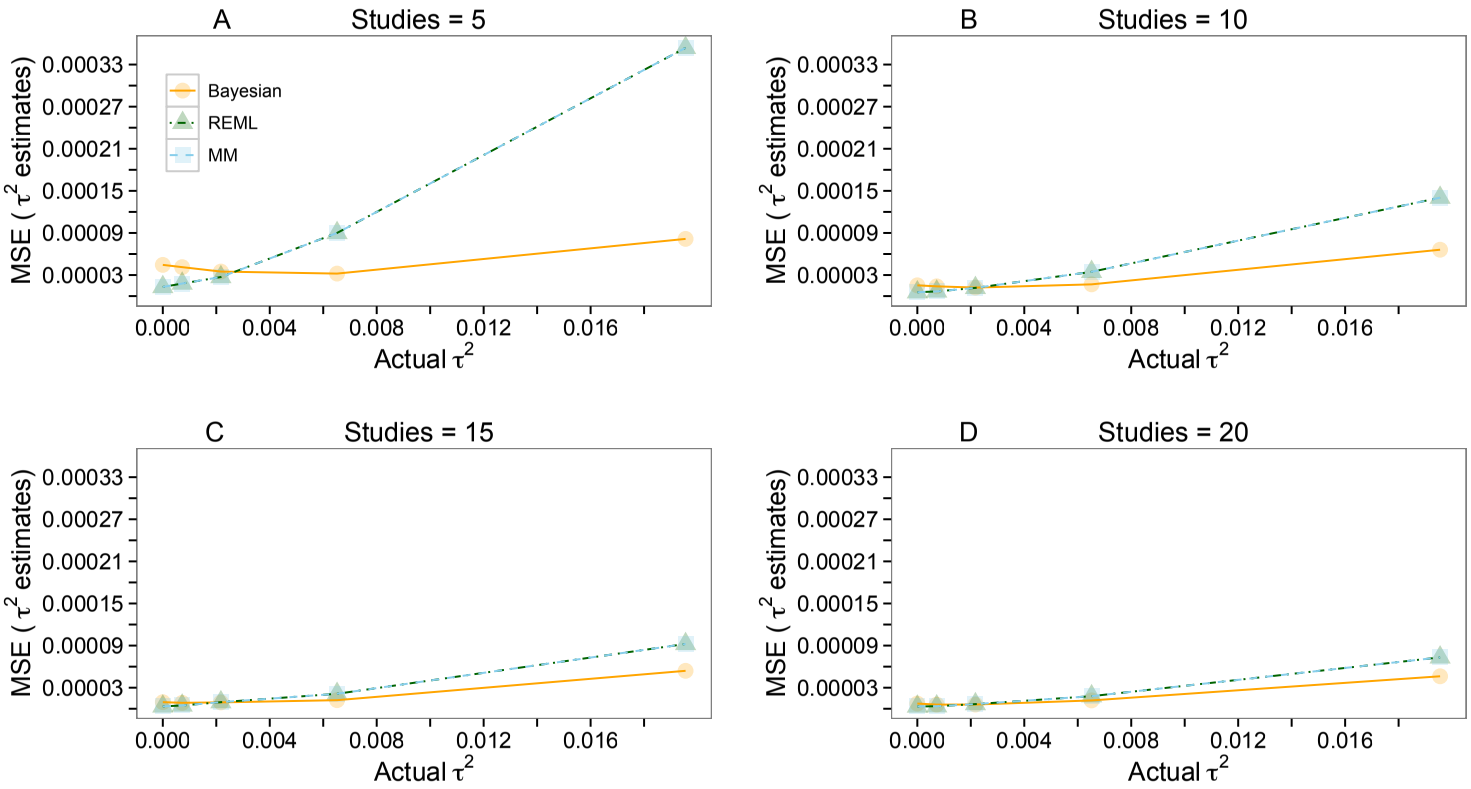


Figure 3.7. A comparison of mean squared errors of heterogeneity estimates afforded by Bayesian and classical genetic association meta-analyses. The four panels (from left to right) show Monte-Carlo simulations of genetic association meta-analysis scenarios with 5, 10, 15 and 20 studies, each based on 1,000 replicates. The meta-analysis simulations were generated in the same manner as that in Figure 3.3. Furthermore, each data point represents the mean squared error at an individual variant for the 1,000 replicates in the meta-analysis simulation. Filled circles denote mean squared error values for Bayesian heterogeneity estimates; whilst filled triangles and squares represent mean squared error values for restricted maximum likelihood and method-of-moments heterogeneity estimates, respectively. In all four meta-analysis scenarios explored, the mean squared errors for Bayesian estimates were generally smaller than those for frequentist heterogeneity estimates. Additionally, both Bayesian and frequentist approaches were more efficient at estimating heterogeneity in larger meta-analyses as demonstrated by the reduction in mean squared error values with increasing numbers of studies. Source: Lerato Magosi (Farrall and Hopewell labs).

Comparing coverage probabilities of average genetic estimates produced by Bayesian and classical random-effects meta-analyses: In each of the four genetic association meta-analysis scenarios, empirical coverage probabilities were calculated at the 95% level for average genetic effect estimates provided by Bayesian and frequentist random-effects meta-analyses. Coverage probabilities are a measure of accuracy, which evaluates the proportion of times that, confidence (or credible) intervals surrounding average genetic effect estimates capture the true average genetic effect. When empirical coverage probabilities of summary effect estimates are above the nominal level (0.95), it usually suggests conservative behaviour owing to overestimation of heterogeneity (Rhodes et al., 2016). Conversely, coverage below the nominal level typically indicates permissive (i.e. anti-conservative) behaviour due to negatively biased heterogeneity estimates (Trikalinos et al., 2013).

Figure 3.8 shows coverage probabilities of average genetic effect estimates produced by Bayesian and classical random effects meta-analyses as a function of the extent of heterogeneity. Both approaches generally showed nominal coverage with increasing heterogeneity. In addition, coverage probabilities for both Bayesian and frequentist methods were slightly more conservative in meta-analysis scenarios where participating studies were few (e.g. 5 studies) and the amount of heterogeneity was marginal (e.g. $\tau^2 = 0$); this finding was consistent with simulation studies by Chung and colleagues, which showed conservative coverage in Bayesian and frequentist summary effect estimates, when heterogeneity was minimal, (i.e. $\tau^2 = 0$) (Chung et al, 2013).

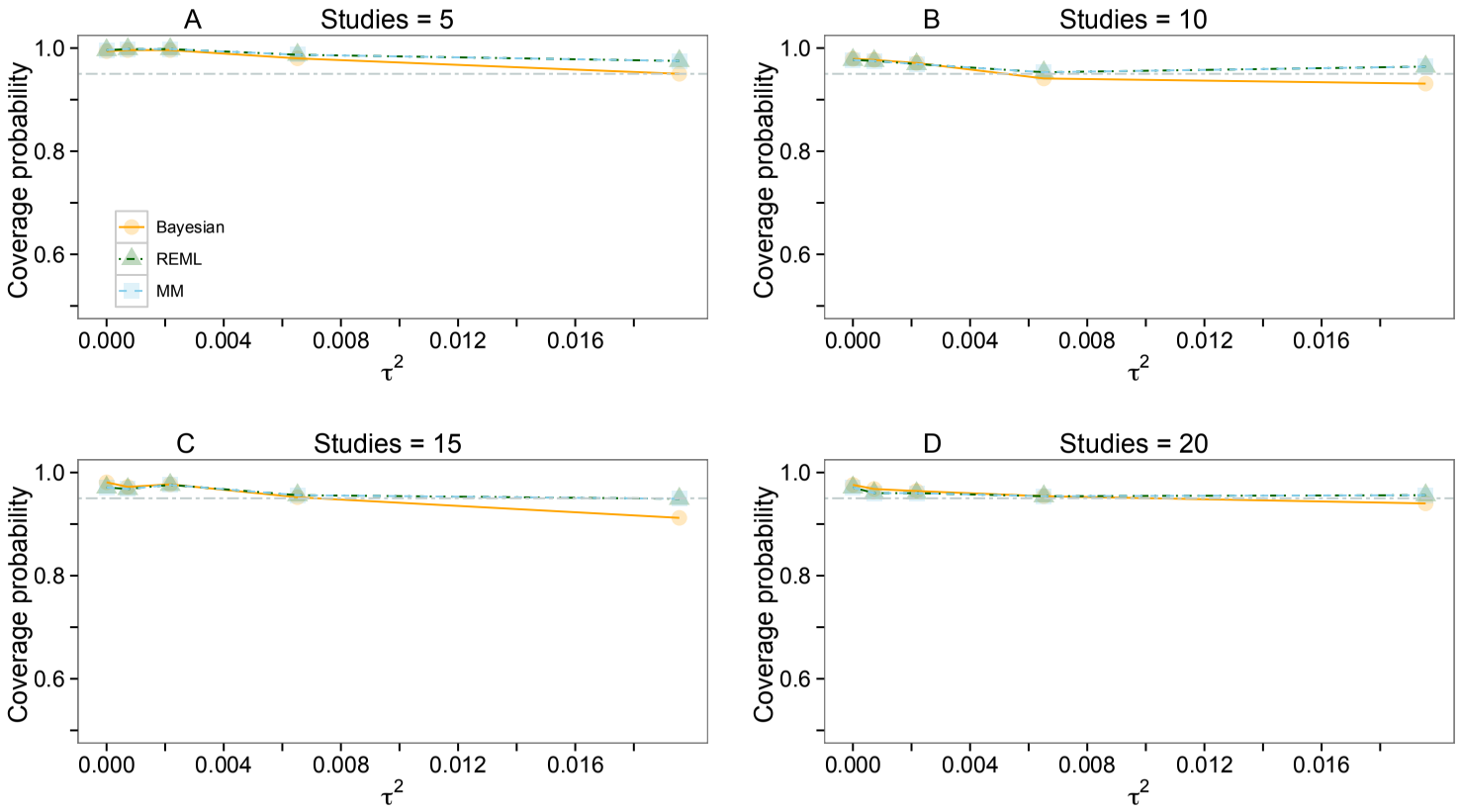


Figure 3.8: A comparison of coverage probabilities of average genetic effect estimates provided by Bayesian and classical random-effects meta-analyses. The four panels (from left to right) show Monte-Carlo simulations of genetic association meta-analysis scenarios with 5, 10, 15 and 20 studies, each based on 1,000 replicates. The meta-analysis simulations were generated in the same manner as that in Figure 3.3. Furthermore, each data point represents the coverage probability of 95% credible (or confidence) intervals for the average genetic effect estimate at an individual variant, across the 1,000 replicates in the simulation. Filled circles denote data points for Bayesian coverage probabilities; whilst filled triangles and squares represent coverage probabilities of classical random-effects meta-analyses performed using restricted maximum likelihood and method-of-moments algorithms, respectively. Additionally, the broken line (i.e. coverage probability $= 0.95$) denotes the nominal level of coverage. Coverage probabilities of Bayesian and frequentist average genetic effect estimates were generally at nominal levels. However, in genetic association meta-analysis scenarios with few studies (e.g. 5 studies) coverage probabilities of both methods tended to be more conservative i.e. higher than the nominal level when there was minimal heterogeneity (e.g. $\tau^2 = 0$). Suggesting a tendency of both methods to overestimate heterogeneity levels when the actual amount of heterogeneity is marginal. Source: Lerato Magosi (Farrall and Hopewell labs).

3.3.2 Aim 2: Comparing the precision of Bayesian and frequentist average genetic effect estimates in a meta-analysis of coronary artery disease

Simulation studies in Aim 1 (section 3.3.1) had shown that Bayesian random-effects meta-analyses employing empirically informed heterogeneity priors, produce estimates of average genetic effects and heterogeneity that are moderately more precise than those offered by frequentist methods. As such, two sets of analyses were performed on the CARDIoGRAMplusC4D GWAS meta-analysis of CAD risk (CARDIoGRAMplusC4D Consortium, 2015), with the aim of improving inference on average genetic effects and heterogeneity, at suggestively associated variants in the dataset (i.e. $P < 5 \times 10^{-5}$). Notably, these were variants in linkage equilibrium with each other (i.e. lead variants) that had showed moderate to high heterogeneity levels $(0.001 \leq \tau^2 \leq 0.06 \text{ i.e. } 13.51\% \leq I^2 \leq 90.36\%)$.

To evaluate the impact of employing GWAS-derived empirical heterogeneity priors in genetic association meta-analyses based on small numbers of studies; the first set of analyses concerned selecting a random subset of studies from the CARDIoGRAMplusC4D dataset, and comparing the precision of average genetic effect estimates produced by Bayesian and classical random-effects meta-analyses for a panel of 83 suggestively associated lead variants. Notably, the first set of analyses had a similar scope (i.e. a moderate-sized genetic association meta-analysis) to the analysis shown in Figure 3.3B.

Constructing an empirically informed heterogeneity prior using a subset of studies in the CARDIoGRAMplusC4D dataset: In order to construct a prior that was representative of the extent of heterogeneity in the CARDIoGRAMplusC4D dataset, the sample function in R®'s base package was used to generate ten subsets, where each subset comprised a random selection of 10 studies from the 48 available in the CARDIoGRAMplusC4D dataset. A GWAS meta-analysis of the common set of ~9 million variants was performed in each of the ten subsets to generate empirical heterogeneity distributions. Notably, the GWAS meta-analyses were carried out using a version of GWAMA (a software program for frequentist GWAS meta-analysis) that had been tweaked to facilitate reporting of between-study variance estimates. The empirical heterogeneity distributions were then employed to generate informed heterogeneity priors for each of the ten subsets, in accordance with section 3.2.3 of the methods.

Notably, the best fitting distributions in all but two of the ten subsets were Weibull distributions. In the two subsets where the Weibull distribution was not the best fitting distribution, visual inspection and AIC values showed that the Weibull distribution was the second best fitting distribution and only marginally out performed by the best fitting distribution. As an exemplar, a comparison of the fit of the different theoretical distributions explored to the empirical distribution of heterogeneity estimates is shown in Figure 3.9. Accordingly, the respective shape, and scale parameters of empirical heterogeneity priors for the different subsets were then averaged to afford a single Weibull heterogeneity prior, $\tau^2 \sim WEB(\nu = 5.580,\ \lambda = 0.00181)$.

Notably, the Weibull distribution being the best fitting distribution in this analysis strikes a contrast with the moderate-sized meta-analysis simulation in Figure 3.3B (empirical distribution shown in Figure 3.2) where an inverse-gamma distribution was the best fitting distribution. That said empirical distributions of heterogeneity estimates for the two sets of

analyses are broadly similar (Figures 3.2 and 3.9) in that they are both positively skewed and have comparable means and variances, where the mean and standard deviation of the empirical distribution for the meta-analysis simulation in Figure 3.3B are: mean = 5.37, SD = 1.39; and those for the empirical distribution in Figure 3.9 are: mean = 5.31, SD = 1.76.

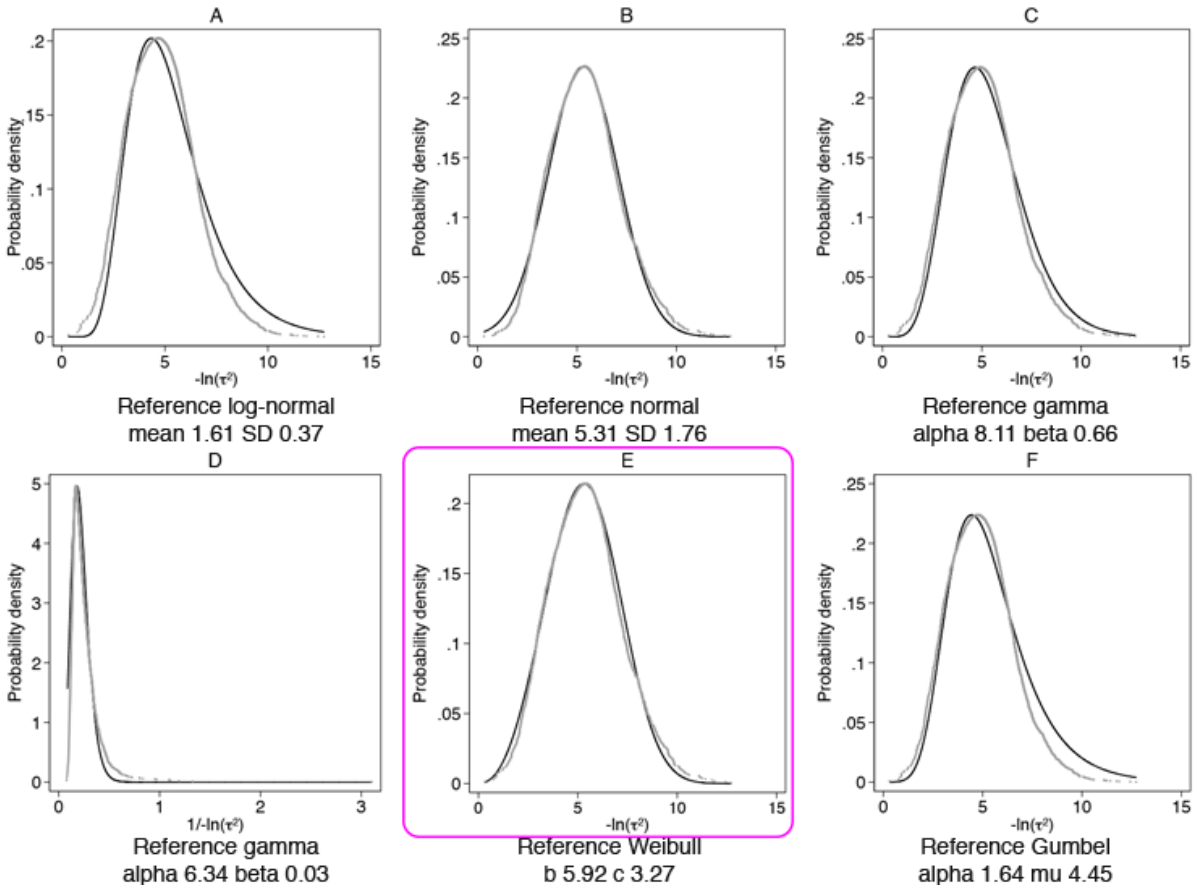


Figure 3.9: Identifying the best fitting theoretical distribution to an empirical distribution of heterogeneity estimates in a random subset of studies from the CARDIoGRAMplusC4D meta-analysis of coronary artery disease..The six panels (from left to right) show comparisons of different theoretical distributions namely: log-normal (panel A), normal (panel B), gamma (panel C), inverse-gamma (panel D), Weibull (panel E) and Gumbel (panel F) to an empirical distribution of method-of-moments heterogeneity estimates, in a random subset of ten studies in the CARDIoGRAMplusC4D meta-analysis (CARDIoGRAMplusC4D, 2015). Within each panel the theoretical distribution is denoted by a black curve whilst the empirical distribution is represented by a grey curve. Moreover, the heterogeneity estimates, τ^2 shown are scaled as log odds ratios. The parameters of the theoretical distributions explored were determined through maximum likelihood estimation. Notably, the Weibull distribution has a relatively closer fit to the empirical distribution of heterogeneity estimates compared to the other theoretical distributions. Suggesting that the Weibull distribution is the best fitting distribution, $\tau^2 \sim WEB(\nu = 5.92, \lambda = 0.001)$ (panel E). Where ν denotes the scale parameter, b whilst λ corresponds to a transformation of the shape parameter, c such that $\lambda = (1/shape)^{\nu}$. Source: Lerato Magosi (Farrall and Hopewell labs).

Estimating average genetic effects and heterogeneity: In each of the ten subsets of studies from the CARDIoGRAMplusC4D dataset, Bayesian and classical random-effects meta-analyses were performed to estimate average genetic effects and heterogeneity for a panel of 83 suggestively associated lead variants; the resulting estimates for each variant were then averaged across the ten subsets yielding 83 mean estimates for average genetic effects and heterogeneity respectively. Figure 3.10 shows credible and confidence intervals of average genetic effect estimates produced both by Bayesian and classical random effects meta-analyses, as a function of the amount of heterogeneity. At variants with relatively low heterogeneity levels ($\tau^2 \leq 0.006$, corresponding $I^2 \leq 48.39\%$), both methods provided "equally" precise average genetic effect estimates. However, for variants that had moderate to high heterogeneity levels ($0.006 < \tau^2 \leq 0.06$, i.e. $48.39\% < I^2 \leq 90.36\%$), Bayesian average genetic estimates were noticeably more precise than those afforded by the frequentist approach, almost as if a threshold had been breeched, demonstrating the utility of GWAS-derived heterogeneity priors in improving inference on average genetic effects and heterogeneity in small meta-analyses.



Figure 3.10. A comparison of the precision of Bayesian and classical average genetic effect estimates in the CARDIoGRAMplusC4D meta-analysis of coronary artery disease, using a randomly selected subset of studies. A scatterplot showing estimates of average genetic effects (y-axis) and heterogeneity (x-axis) scaled as log odds ratios at 83 suggestively associated lead variants ($P \leq 5 \times 10^{-5}$). The CARDIoGRAMplusC4D dataset comprises 48 studies. Therefore, to produce a small relatively representative subset of studies in the CARDIoGRAMplusC4D dataset that has the scope of a moderately-sized genetic association meta-analysis; 10 subsets were generated each comprising 10 randomly selected studies. Bayesian and classical random-effects meta-analyses were then performed in each subset, and estimates of average genetic effects and heterogeneity averaged across the 10 subsets. The amount of heterogeneity in the frequentist random-effects meta-analyses was estimated by method-of-moments. By contrast, Bayesian random-effects meta-analyses in each of the ten subsets were informed with an empirical Weibull heterogeneity prior, $\tau^2 \sim WEB(\nu = 5.581, \lambda = 0.002)$ formulated from the empirical distributions of method-of-moments heterogeneity estimates in the frequentist meta-analyses. Bayesian random-effects meta-analyses were performed using Markov chain Monte-Carlo methods, wherein results were based on 50,000 iterations of the Gibbs sampler following a burn-in period of 50,000 steps. Convergence was monitored with trace plots, Gelman-Rubin statistics and autocorrelation plots. Notably, the maximum Gelman-Rubin statistic value among the 83 variants (Shrink factor = 1.01) was within the conventional acceptable range of 1.00 to 1.10 (Kruschke, 2014). Bayesian and frequentist estimates were similarly precise at variants with moderate to low levels of heterogeneity (i.e. $\tau^2 \leq 0.006$, corresponding $I^2 \leq 48.39\%$). However at variants with higher amounts of heterogeneity, Bayesian average genetic effect estimates were relatively more precise than their frequentist counterparts. Source: Lerato Magosi (Farrall and Hopewell labs).

The second set of analyses involved comparing the precision of summary effect intervals (i.e. confidence and credible intervals) afforded by Bayesian and classical random-effects meta-analyses, for the same panel of suggestively associated lead variants as that used in the first set of analyses, this time employing all 48 studies in the CARDIoGRAMplusC4D dataset. Whilst there was an expectation that the large amount of data provided by the 48 participating studies would have a greater influence on the Bayesian results than the empirical heterogeneity prior; it was of interest to evaluate whether informing the Bayesian genetic association meta-analysis would offer any advantage over the frequentist approach.

Constructing an empirically informed heterogeneity prior using the full complement of studies in the CARDIoGRAMplusC4D dataset: Similar to the first set of analyses, a GWAS meta-analysis of the common set of ~9 million variants was performed to generate an empirical distribution of heterogeneity estimates; the resulting distribution was then used to formulate an empirically informed heterogeneity prior in line with section 3.2.3 of the methods, $\tau^2 \sim Gamma(shape = 11.87, rate = 1/0.52)$. A comparison of the fit of the empirical distribution to the various theoretical distributions explored is shown in Figure 3.11.

Measuring summary effects and heterogeneity: Bayesian and classical random-effects meta-analyses of the 83 suggestively associated lead variants were performed to estimate their average genetic effects and extent of heterogeneity. Both methods generally provided average genetic effect estimates that were similarly precise, except at variants with relatively higher amounts of heterogeneity ($\tau^2 \geq 0.006$, corresponding $I^2 \geq 48.39\%$), where Bayesian estimates were slightly more precise than their frequentist counterparts, Figure 3.12. This result supported the earlier hypothesis that the large amount of information across the 48 studies would likely

overwhelm the empirical heterogeneity prior, and aligns with the observations made in the meta-analysis simulation in Figure 3.3D.

Together with the Monte-Carlo simulation studies performed in Aim 1 (section 3.3.1), the example application in the CARDIoGRAMplusC4D dataset has demonstrated the advantages of GWAS-derived empirical heterogeneity priors over frequentist methods; in terms of improving inference on average genetic effects and heterogeneity in small genetic association meta-analyses, with moderate to high heterogeneity levels.

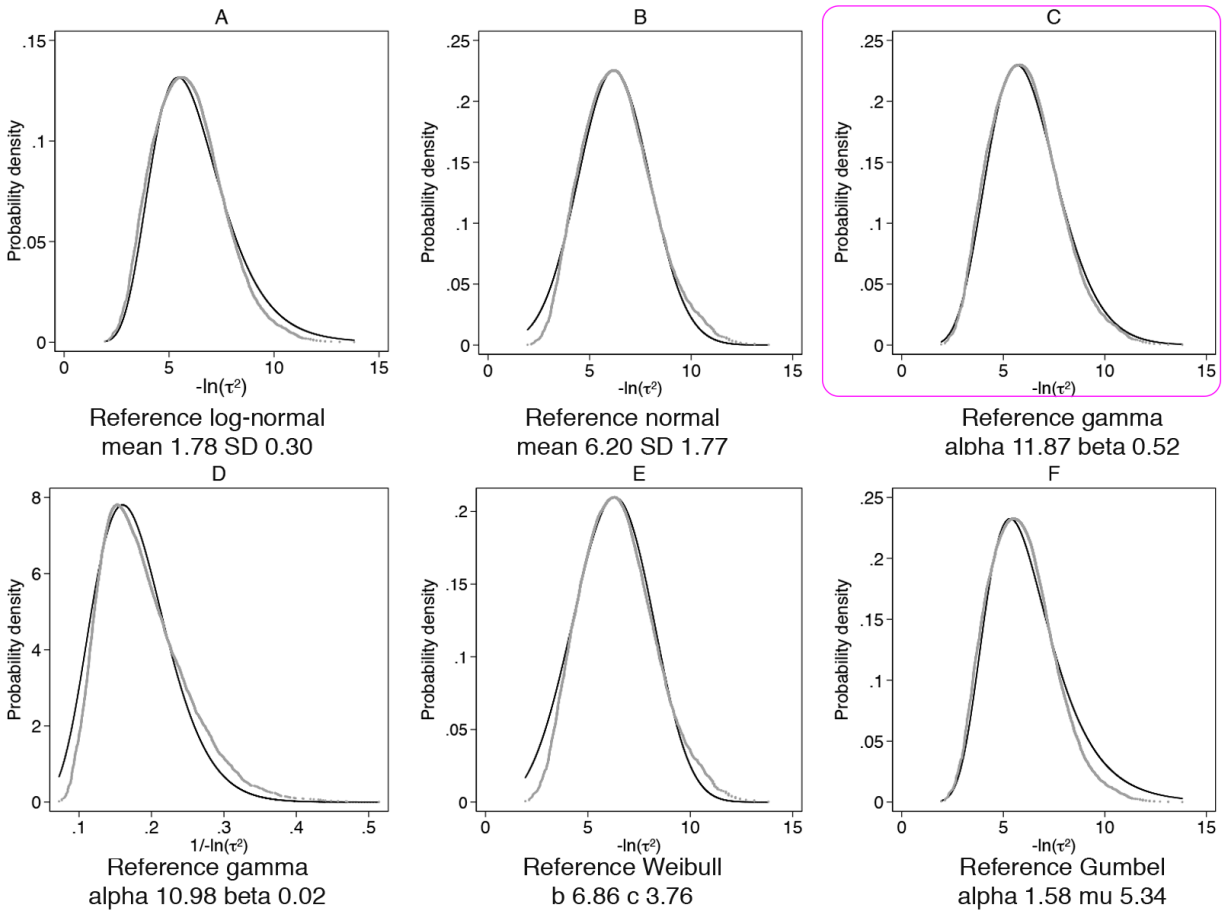


Figure 3.11: Identifying the best fitting theoretical distribution to an empirical distribution of heterogeneity estimates sourced from the full complement of studies (48) in the CARDIoGRAMplusC4D meta-analysis of coronary artery disease.. The six panels (from left to right) show comparisons of different theoretical distributions namely: log-normal (panel A), normal (panel B), gamma (panel C), inverse-gamma (panel D), Weibull (panel E) and Gumbel (panel F) to an empirical distribution of method-of-moments heterogeneity estimates, τ^2 scaled as log odds ratios. Within each panel the theoretical distribution is denoted by a black curve whilst the empirical distribution is represented by a grey curve. The parameters of the theoretical distributions explored were determined through maximum likelihood estimation and the best fitting distribution was the gamma distribution, $\tau^2 \sim$ Gamma(shape = 11.87, rate = $1/0.52$) as shown in panel C. Source: Lerato Magosi (Farrall and Hopewell labs).

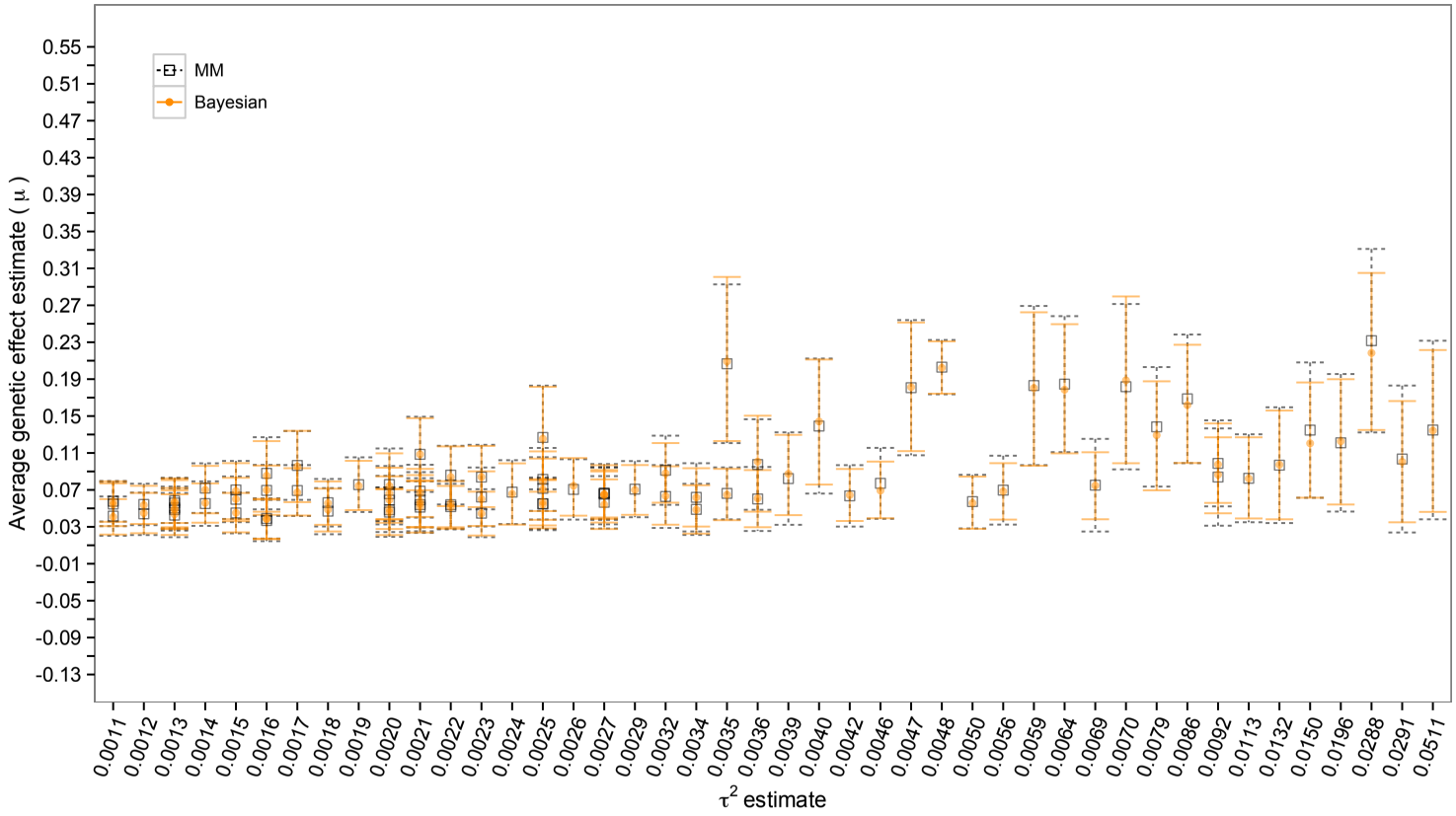


Figure 3.12. A comparison of the precision of Bayesian and classical average genetic effect estimates in the CARDIoGRAMplusC4D meta-analysis of coronary artery disease, using the full complement of 48 studies. A scatterplot showing estimates of average genetic effects (y-axis) and heterogeneity (x-axis) scaled as log odds ratios at 83 suggestively associated lead variants ($P \leq 5 \times 10^{-5}$). Bayesian and classical random-effects meta-analyses of the 48 contributing studies were performed for each of the 83 variants. The extent of heterogeneity in the frequentist random-effects meta-analyses was estimated using a method-of-moments algorithm, whislt Bayesian random-effects meta-analyses were informed with an empirical gamma heterogeneity prior, $\tau^2 \sim$ Gamma(shape = 11.87, rate = $1/0.52$) generated from the empirical distribution of method-of-moments heterogeneity estimates in the frequentist meta-analyses. Moreover, Bayesian random-effects meta-analyses were performed using Markov chain Monte-Carlo methods, wherein results were based on 50,000 iterations of the Gibbs sampler following a burn-in period of 50,000 steps. Convergence was monitored with trace plots, Gelman-Rubin statistics and autocorrelation plots. Notably, the maximum Gelman-Rubin statistic value among the 83 variants (Shrink factor = 1.02) was within the conventional acceptable range of 1.00 to 1.10 (Kruschke, 2014). Bayesian and frequentist estimates were generally equally precise, except at variants with higher heterogeneity levels (i.e. $\tau^2 \geq 0.006$, corresponding $I^2 \geq 48.39\%$) where Bayesian average genetic effect estimates were slightly more precise than their frequentist counterparts. Source: Lerato Magosi (Farrall and Hopewell labs).

3.4 Discussion

3.4.1 Rationale for improving average genetic effect estimates in smaller genetic association meta-analyses

Genome-wide association meta-analyses are widely used for the discovery of common genetic variants conferring smaller effects (e.g. per-allele disease odds ratios < 1.2 or trait variance < 0.2%) (Bush and Moore, 2012; Yang et al., 2011), and the validation of risk variants previously identified in stand-alone GWAS. However, when meta-analyses are based on small numbers of studies, estimates of average genetic effects and the extent of heterogeneity are often imprecise (Turner et al., 2015). In this work, empirically informed heterogeneity priors are proposed as a method for improving inference on average genetic effects and between-study variance in small genetic association meta-analyses. Bayesian methods allow for the inclusion of background information on the extent of heterogeneity (through informed priors) in a meta-analysis, and account for the uncertainty surrounding estimates of both summary effects and heterogeneity. Thus, my hypothesis was that informing Bayesian genetic association meta-analyses with empirical heterogeneity priors derived from GWAS meta-analysis, could yield estimates of average genetic effects and heterogeneity that were more precise than those provided by frequentist methods.

Through simulation studies and an application to a recent meta-analysis of CAD risk, I demonstrated the utility of empirically informed heterogeneity priors in improving precision of

estimates for summary effects and heterogeneity in smaller genetic association meta-analyses (i.e. ≤ 10 studies). The influence of the empirical heterogeneity priors on estimates of average genetic effects and heterogeneity was most evident in meta-analysis scenarios with moderate to high amounts of heterogeneity ($\tau^2 > 0.006$, corresponding $I^2 > 48.39\%$). Altogether, the empirically informed heterogeneity priors employed in this work, provided a modest improvement in inference of average genetic effects; in light of this, Bayesian genetic association meta-analyses with GWAS-derived empirical heterogeneity priors are likely to be more useful as a tool to refine estimates of average genetic effects and between-study variance, particularly at heterogeneous variants suggestively associated with a trait of interest.

A potential reason why the use of empirical heterogeneity priors had a moderate influence on average genetic effect estimates in this work is that, even with the inclusion of priors delineating the range of possible heterogeneity values, meta-analyses with few contributing studies have limited information to facilitate calculation of precise estimates. Furthermore, prior distributions in meta-analyses based on larger numbers of studies tend to be overwhelmed by the evidence participating studies provide; and therefore, have little influence on meta-analysis results.

3.4.2 Advantages and limitations of employing GWAS derived empirical heterogeneity priors.

The main practical benefit of performing Bayesian genetic association meta-analyses with empirically informed heterogeneity priors as proposed, is that it affords the opportunity to take advantage of the large volume of heterogeneity information that a GWAS platform provides. In addition the use of empirically informed heterogeneity priors reduces subjectivity in the declaration of prior distributions for between-study variance. Another advantage of employing Bayesian methods is that they naturally convey the level of uncertainty in heterogeneity estimates through posterior distributions; this is a key piece of information for geneticists to have when judging the reliability of meta-analysis results, as highly variable heterogeneity estimates can contribute to imprecise average genetic effect estimates.

A limitation of the proposed approach is that the parameters of prior distributions are maximum likelihood estimates, yet their uncertainty is not taken into account. An alternative Bayesian approach that accounts for variability in prior parameters would be to perform hierarchical Bayesian genetic association meta-analyses (Beaumont and Rannala, 2004). That said, the traditional Bayesian random-effects models employed in this work were favoured as a simple way of demonstrating the usefulness of empirically informed heterogeneity priors in smaller genetic association meta-analyses. One challenge associated with using Markov chain Monte-Carlo methods to draw inference on average genetic effects and heterogeneity in Bayesian meta-analyses, is in accurately assessing representative sampling and convergence of Markov chains to target posterior distributions; this is because diagnostics generally imply convergence rather than provide direct evidence of its occurrence. However, there was little

evidence to suggest that the Markov chains in the meta-analysis scenarios explored in this work failed to converge, this might have been a consequence of the relatively simple models considered where there were only two parameters of interest, namely the average genetic effect and extent of heterogeneity. On reflection, the quality of the convergence diagnostics in the meta-analysis scenarios explored taken together with the simplicity of the meta-analysis models suggests a shorter burn-in period of say, 1,000 iterations would have sufficed for adequate convergence.

Aside from MCMC methods, Bayesian random-effects meta-analyses can also be conducted using numerical integration, importance sampling or data augmentation. These approaches have different advantages over MCMC methods. For example numerical integration does not involve simulation, as such results (i.e. posterior summaries) of Bayesian meta-analyses based on numerical integration are not subject to Monte-Carlo error (Turner et al., 2015). Similar to numerical integration, data augmentation does not require simulation, thus making it quicker to run than MCMC methods. Furthermore, Bayesian meta-analyses performed using data augmentation as proposed by Turner and colleagues do not require convergence checking (Turner et al., 2015). That said, MCMC methods remain the standard procedure for performing Bayesian random-effects meta-analyses (Rhodes et al., 2016; Turner et al., 2015).

3.5 Chapter summary

In summary, this chapter proposes Bayesian estimation with GWAS-derived empirical heterogeneity priors as an option to improve inference on average genetic effects and heterogeneity in smaller genetic association meta-analysis. There were two primary findings in this chapter. First, empirically informed heterogeneity priors modestly improved precision in the estimation of summary effects and heterogeneity, particularly in smaller meta-analyses; and second, empirically informed heterogeneity priors had a greater influence on the results of Bayesian genetic association meta-analyses with moderate to high heterogeneity levels.

This page is intentionally left blank.

Chapter 4

This chapter provides a critical evaluation of the Han-Eskin random-effects model in relation to traditional (i.e. conventional) fixed and random-effects meta-analysis methods.

4.1 Hypothesis/ Research needs

As highlighted in chapter one (section 1.2), the traditional random-effects model (DerSimonian and Laird, 1986) is often outperformed by the fixed-effect method (Borenstein et al., 2010; Hardy and Thompson, 1998) in the detection of heterogeneous genetic signals (Neupane et al., 2012; Thompson et al., 2011); that is, for the same total sample size the traditional random-effects model will generally produce p-values that are larger (i.e. more conservative) than those provided by the fixed-effect method (Han and Eskin, 2011). The larger p-values suggest a higher type-2 error rate under the traditional random-effects method, and with that the knock-on effect of reduced power to detect genetic associations. The practical implication here is that it is unlikely for the conventional random-effects model to identify genetic signals that would not have otherwise been discovered using the fixed-effect method; this is important to note, given that the traditional random-effects model is typically employed in meta-analyses where heterogeneity tests (e.g. the Q-test) have detected significant variation between study-effects or heterogeneity is suspected (Han and Eskin, 2011; Neupane et al., 2012).

While it might be tempting to rely on the traditional fixed-effect model for the discovery of heterogeneous genetic signals, given its power advantage over the conventional random-

effects method, it is also important to note that the design of the traditional fixed-effect model principally facilitates detection of homogeneous genetic signals. Moreover, applying the fixed-effect model in the presence of heterogeneity can lead to unduly small p-values (Thompson et al., 2011). Thus it stands to reason that genetic associations with moderate-to-high heterogeneity might elude detection under traditional meta-analytic approaches, more so under the conventional random-effects model.

The tendency for conservative p-values under the traditional random-effects model can be ascribed to two factors. First, in contrast to the fixed-effect method, the conventional random-effects model considers two sources of variation, namely within-study and between-study variance. Consequently, standard errors of summary effect estimates under the random-effects model tend to be larger than those afforded by the fixed-effect method (Borenstein et al., 2010); thus larger studies in random-effects meta-analyses will have less influence on the average genetic effect in the presence of heterogeneity. To demonstrate, consider a meta-analysis comprising *S* studies where the genetic effect-size and corresponding standard error for a variant of interest were estimated by regression modelling in each participating study, and the summary effect-size estimate, $\hat{\theta}$ computed as the inverse-variance weighted mean of the individual study effects,

$$\hat{\theta} = \frac{\sum_{s=1}^{S} w_s y_s}{\sum_{s=1}^{S} w_s}.$$

Where y_s denotes the observed study effect in the s^{th} study. And w_s represents the weight allocated to the s^{th} study, which can be calculated as, $w_s = \frac{1}{(\sigma_s^2)}$ under the fixed-effect method

and $w_s = \frac{1}{(\sigma_s^2 + \hat{\tau}^2)}$ under the random-effects method. Notably, σ_s^2 and $\hat{\tau}^2$ denote within-study and between-study variance, respectively.

Then it follows that the variance of the summary effect-size estimate is given by $V_{\hat{\theta}} = \frac{1}{\sum_{s=1}^{S} w_s}$, such that standard errors for the summary effect-size, $SE(\hat{\theta})$ under the traditional fixed-effect (FE) and random-effects (RE) methods are,

$$SE(\hat{\theta})_{FE} = \sqrt{\left(\sum_{s=1}^{S} \frac{1}{\sigma_s^2}\right)^{-1}} \text{ and } SE(\hat{\theta})_{RE} = \sqrt{\left(\sum_{s=1}^{S} \frac{1}{(\sigma_s^2 + \hat{\tau}^2)}\right)^{-1}} \text{ respectively,}$$

where $SE(\hat{\theta})_{RE} \geq SE(\hat{\theta})_{FE}$.

Assuming the observed study effects are normally distributed then it can be expected that the summary effect estimate would be normally distributed as well. A two-tailed p-value can then be calculated for the genetic association at the variant of interest using a z-score, such that the p-value for the variant of interest as measured by the traditional fixed-effect method is given by,

$$Z_{FE} = \frac{\hat{\theta}}{SE(\hat{\theta})} = \frac{\sum_{s=1}^{S} \sigma_s^{-2} y_s}{\sqrt{\sum_{s=1}^{S} \sigma_s^{-2}}} = \frac{\sum w_s y_s}{\sqrt{\sum w_s}} ,$$

$$P_{FE} = 2\Phi(-|Z_{FE}|) .$$

Where Z_{FE} denotes the summary association test statistic under the fixed-effect method and Φ represents the cumulative density function for a standard normal distribution. P_{FE} denotes the

summary association p-value for the fixed-effect method (Han and Eskin, 2011). The p-value of the same variant under the traditional random-effects method, is given by,

$$Z_{RE} = \frac{\hat{\theta}}{SE(\hat{\theta})} = \frac{\sum_{s=1}^{S}(\sigma_s^2 + \hat{\tau}^2)^{-1} y_s}{\sqrt{\sum_{s=1}^{S}(\sigma_s^2 + \hat{\tau}^2)^{-1}}},$$

$$P_{RE} = 2\Phi(-|Z_{RE}|),$$

where $P_{RE} \geq P_{FE}$ (Han and Eskin, 2011).

The second factor contributing to conservative p-values under the traditional random-effects model is a conservative null hypothesis, which dampens power to detect genetic associations (Han and Eskin, 2011); that is, the traditional random-effects model assumes the presence of heterogeneity both under the null and alternative hypotheses. This can be better illustrated when the traditional fixed and random-effects models are presented as likelihood ratio tests. To demonstrate, we continue with the example presented above of a meta-analysis comprising *S* studies; then the likelihood for the observed effect in the s^{th} study under the null and alternative hypotheses of the traditional fixed-effect method can be described as,

$$\mathcal{L}_0 = \prod_s \frac{1}{\sqrt{2\pi\sigma_s^2}} \exp\left(-\frac{(y_s - \theta)^2}{2\sigma_s^2}\right), \text{ where } \theta = 0$$

$$\mathcal{L}_1 = \prod_s \frac{1}{\sqrt{2\pi\sigma_s^2}} \exp\left(-\frac{(y_s - \theta)^2}{2\sigma_s^2}\right), \text{ where } \theta \neq 0 .$$

Where $\mathcal{L}_0$ and $\mathcal{L}_1$ denote the likelihood under the null and alternative hypotheses, respectively. Using Wilks' theorem, that the $-2\log(\lambda)$ statistic is approximately chi-squared for large sample-sizes (Han and Eskin, 2011; Wilks, 1938), the likelihood ratio test statistic of the association under the fixed-effect method is given by:

$$-2\log(\lambda) = -2\log\left(\frac{\sup(\mathcal{L}_0)}{\sup(\mathcal{L}_1)}\right)$$

$$= \sum_s \frac{y_s^2}{\sigma_s^2} - \sum_s \frac{(y_s - \hat{\theta})^2}{\sigma_s^2}$$

expanding the quadratic term $(y_s - \hat{\theta})^2$ and cancelling out like terms,

$$= \sum_s \frac{2y_s\hat{\theta} - \hat{\theta}^2}{\sigma_s^2}$$

expressing the variance, σ_s^2 as the inverse of the sum of the weights $\frac{1}{\sum w_s}$,

$$= 2\hat{\theta}\sum w_s y_s - \hat{\theta}^2 \sum w_s$$

substituting the summary effect estimate, $\hat{\theta}$ with $\frac{\sum w_s y_s}{\sum w_s}$,

$$= \frac{(\sum w_s y_s)^2}{\sum w_s}$$

$$= Z_{FE}^2 .$$

The traditional random-effects model considers the same amount of heterogeneity, $\hat{\tau}^2$ under both the null and alternative hypotheses (Han and Eskin, 2011), such that the likelihood of the observed effect in the s^{th} study is,

$$\mathcal{L}_0 = \prod_s \frac{1}{\sqrt{2\pi(\sigma_s^2 + \hat{\tau}^2)}} \exp\left(-\frac{(y_s - \theta)^2}{2(\sigma_s^2 + \hat{\tau}^2)}\right), \text{ where } \theta = 0$$

$$\mathcal{L}_1 = \prod_s \frac{1}{\sqrt{2\pi(\sigma_s^2 + \hat{\tau}^2)}} \exp\left(-\frac{(y_s - \theta)^2}{2(\sigma_s^2 + \hat{\tau}^2)}\right), \text{ where } \theta \neq 0 .$$

The likelihood ratio test statistic for the association under the traditional random-effects model is then given by,

$$-2\log(\lambda) = \sum_s \frac{2y_s\hat{\theta} - \hat{\theta}^2}{(\sigma_s^2 + \hat{\tau}^2)} = Z_{RE}^2 .$$

4.1.1 Contemporary fixed and random-effects meta-analysis methods

In an effort to avail more powerful tools for the discovery and replication of heterogeneous genetic signals, new fixed-effect (Lebrec et al., 2010) and random-effects methods (Han and Eskin, 2011) have been proposed. These contemporary meta-analytic approaches differ from traditional methods in the hypotheses considered; in particular, for a meta-analysis with *S* contributing studies, the traditional fixed-effect method assumes there is a common genetic effect underlying all the studies, $\theta_1 = \theta_2 = \theta_3 = \cdots = \theta_s = \theta$, and tests the null hypothesis that the common genetic effect-size is zero, $H_0: \theta = 0 \ vs \ H_1: \theta \neq 0$. The summary association test statistic under the null hypothesis is given by, $Z_{FE}^2 = (\frac{\hat{\theta}}{SE(\widehat{\theta})})^2 \sim \chi_1^2$ (asymptotically), where $\hat{\theta}$ denotes the summary effect estimate and $SE(\widehat{\theta})$ represents the corresponding standard error (Neupane et al., 2012).

By contrast, the contemporary fixed-effect model tests for the presence of a statistically significant genetic effect in at least one of the participating studies in the meta-analysis. That is, $H_0: \theta_1 = \theta_2 = \theta_3 = \cdots = \theta_s = 0 \ vs \ H_1: \theta_i \neq 0$, where θ_i denotes the underlying effect in any one of the *S* participating studies. Similar to the traditional fixed-effect meta-analysis method, the summary association test statistic under the contemporary fixed-effect model is given by, $Z_{FE}^2 = (\frac{\hat{\theta}}{SE(\widehat{\theta})})^2 \sim \chi_1^2$ (asymptotically) (Neupane et al., 2012).

One practical advantage of the contemporary (or global) fixed-effect method over the traditional approach, particularly in multi-ethnic GWAS meta-analyses, is that it facilitates discovery of genetic signals that might only occur in a subset of participating studies owing to differential linkage disequilibrium patterns, epistasis or gene-environment interactions. That said there has been limited application of the global fixed-effect method in the GWAS meta-analysis literature; this might have been influenced by simulation studies showing that the global fixed-effect method tends to be more powerful than the traditional fixed-effect model, at detecting small effect loci with high heterogeneity (Neupane et al., 2012). Notably, the credibility of weak associations with high heterogeneity $(\text{e.g. } I^2 \geq 75\ \%)$ would be questioned in a discovery GWAS meta-analysis. In particular, whilst heterogeneity can hint at biologically relevant diversity among participating studies (e.g. variable linkage disequilibrium patterns, gene-gene and gene-environment interactions) as highlighted in chapter 1; weak or near null genetic associations in the presence of high heterogeneity, as shown in Figure 4.1A-B, may sometimes occur due to population structure, genotyping error or the play of chance (e.g. winner's curse phenomenon).

An exploration of heterogeneity sources in the CARDIoGRAMpluC4D meta-analysis, in chapter 2, revealed a negligible contribution of the winner's curse phenomenon and variable genotyping platforms to heterogeneity; suggesting population structure is more likely to contribute to heterogeneity in a meta-analysis. Forest plots can help to elucidate whether the distribution of study effect-estimates reflects biologically relevant diversity or a spurious association. For example, weak associations where studies with smaller sample-sizes show larger effects and the largest studies in the meta-analysis are null are more likely to indicate a spurious association.

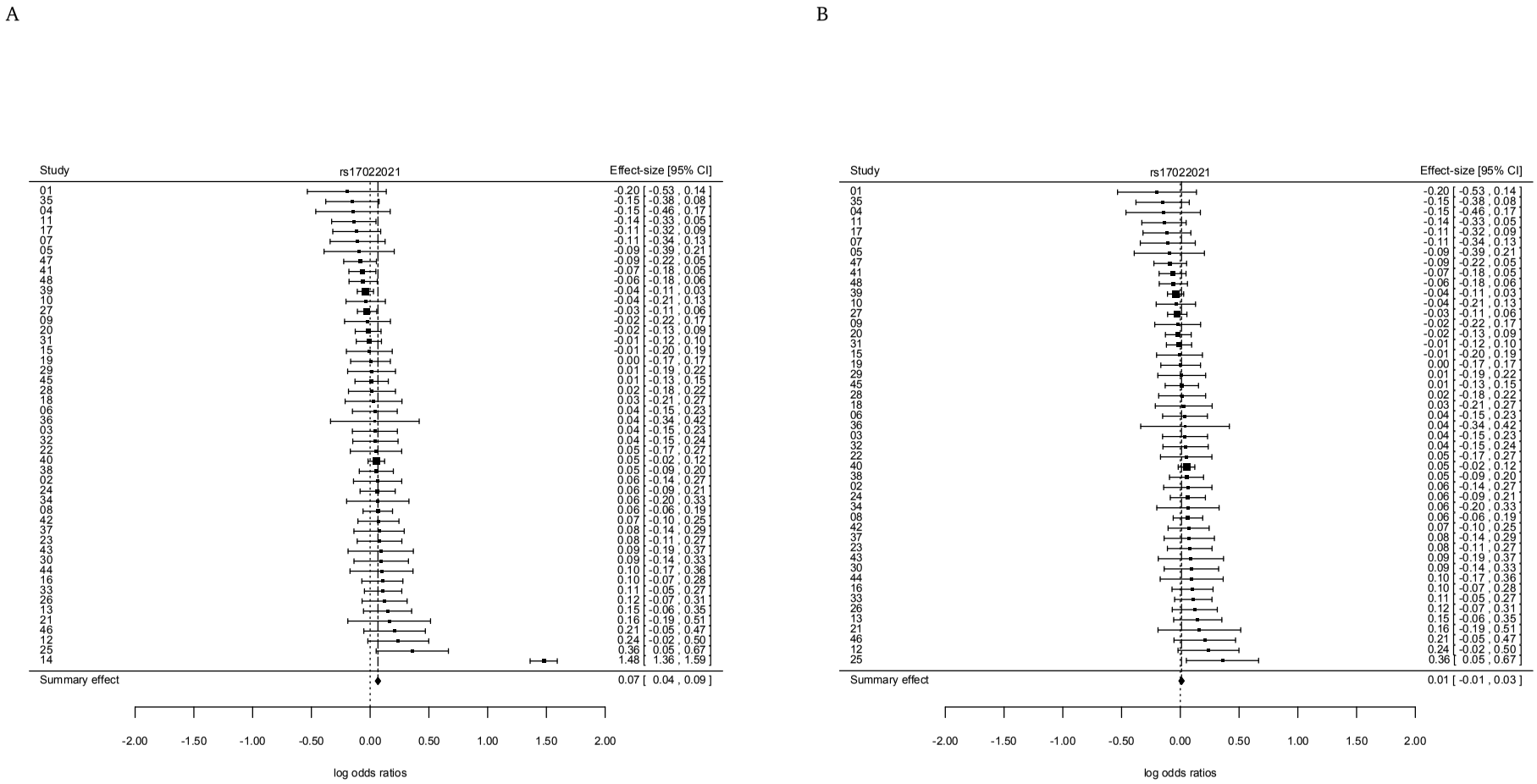


Figure 4.1: Forest plots of study effect-size estimates at rs17022021 (chromosome 3) in the CARDIoGRAMplusC4D meta-analysis, illustrating an inflated genetic association where the summary effect-size estimate is inflated by an outlier study showing extreme effects. Sorted effect-size estimates (log odds ratios) are presented for individual studies represented by filled squares with their 95% confidence intervals shown by horizontal lines; the sizes of the squares are proportional to each studies' inverse-variance weighting. A filled diamond represents the summary effect-size estimate. The forest plot in panel A shows the dispersion of study effect-size estimates at rs17022021, when all 48 studies are included in the meta-analysis ($I^2 = 92.63\%$); $P_{RE} = 0.18; P_{FE} = 7.49 \times 10^{-9}$), whilst that in panel B shows their distribution after excluding the outlier study (study no. 14) from the meta-analysis ($I^2 = 0.00\%; P_{RE} = 0.35; P_{FE} = 0.37$). Notably, the outsized genetic effect in study 14 contributes to the extensive heterogeneity and exaggerated average genetic effect estimate at rs17022021 in Panel A, leading to an apparent association with coronary artery disease. Also notable is that the Han-Eskin random-effects model reported a smaller association p-value for rs17022021 than the traditional meta-analysis methods, $P_{RE(new)} << P_{FE} << P_{RE}$. The Higgin's inconsistency index is denoted by I^2. Whilst P_{FE}, P_{RE} and $P_{RE(new)}$ represent association p-values for traditional fixed and random-effects meta-analyses and the Han-Eskin random-effects model, respectively. Source: Lerato Magosi (Farrall and Hopewell labs).

With regards to the random-effects meta-analysis models, the traditional random-effects method assumes that the underlying study effects might be identical across all participating studies or vary from study-to-study. In addition, it (the traditional random-effects method) tests the null hypothesis that the average genetic effect, μ is zero; that is, $H_0: \mu = 0 \ vs \ H_1: \mu \neq 0$. The summary association test statistic for the traditional random-effects method under the null is given by, $Z_{RE}^2 = (\frac{\widehat{\mu}}{SE(\widehat{\mu})})^2 \sim \chi_1^2$ (asymptotically).

In contrast to the traditional random-effects method which implicitly assumes the presence of heterogeneity under the null, contemporary random-effects methods such as the Han-Eskin random-effects model (Han and Eskin, 2011) test the null hypothesis that the average genetic effect is zero and there is no heterogeneity; that is, $H_0: \mu = 0 \ and \ \tau^2 = 0 \ vs \ H_1: \mu \neq 0 \ or \ \tau^2 > 0$ (asymptotically) (Neupane et al., 2012). The summary association test statistic (or likelihood ratio test statistic) under the null is then given by,

$$Z_{RE\ (new)}^2 = -2\log(\lambda)$$

$$= -2\log\left(\frac{L_0(0,0)}{L_1(\hat{\mu},\hat{\tau}^2)}\right).$$

The assumption of homogeneity of study effects under the null allows for a sharper contrast between the null and alternative hypotheses, and potentially affords the Han-Eskin random-effects model greater power than the traditional random-effects method, to discover heterogeneous genetic associations. Given that the Han-Eskin random-effects model considers the alternative hypothesis that there is either a non-null average genetic effect or extensive heterogeneity in a meta-analysis i.e. $H_1: \mu \neq 0 \ or \ \tau^2 > 0$, I hypothesised that variants with null (or near null) average genetic effects might occasionally produce significant p-values under the

Han-Eskin random-effects model owing to the presence of substantial heterogeneity (i.e. inflated genetic signals).

4.1.2 Potential heterogeneity sources for inflated genetic effects at small effect loci

Heterogeneity patterns that are likely to produce inflated genetic associations and raise credibility issues are ones where the average genetic effect at a variant is inflated by a few (i.e. one or two) outlier studies showing extreme effects, whilst the rest of the study effects are minimal (Figure 4.1A). Such heterogeneity patterns may arise due to global (genome-wide) population structure that persists even after a study-level genomic correction has been applied to adjust association statistics, or more likely as a result of local (marker-specific) ancestry differences at the variant examined. Global population structure is typically brought about by migration or random drift, whilst local ancestry differences are generally induced by natural selection (Wang et al., 2011).

Notably, the authors of the Han-Eskin random-effects model have included a correction for genomic inflation in the method, however they concede that some genomic inflation persists even after it has been applied (Han and Eskin, 2011); this was demonstrated in a meta-analysis simulation they performed on seven studies based on type-1 diabetes data from the Wellcome Trust Case Control Consortium. Where variants showing significant effects were excluded in order to compare quantile-quantile plots for the Han-Eskin random-effects model before and after genomic correction with those from traditional fixed and random-effects methods, under the null hypothesis of no association. There was minimal inflation in the tails of quantile-quantile

plots of traditional fixed and random-effects methods; however there was noticeable inflation in the tail of the uncorrected quantile-quantile plot of the Han-Eskin random-effects model. Moreover, the tail of the genomically corrected Han-Eskin quantile-quantile plot showed less inflation compared to its uncorrected counterpart, but was relatively more inflated than quantile-quantile plots of the traditional fixed and random-effects methods (Han and Eskin, 2011). Given that significant genetic signals had been removed from the outset, the inflation in the tail of the corrected Han-Eskin quantile-quantile plot was possibly a reflection of heterogeneity due to a mixture of global population structure and local ancestry differences; tying in with the hypothesis made in section 4.1.1 that the Han-Eskin random effects model is more likely to report misleadingly small p-values at variants with near-null associations (e.g. $1.00 < \text{oddsratio} < 1.06$) as the amount of heterogeneity increases.

4.1.3 Performance of traditional and contemporary fixed and random-effects methods

Simulation studies evaluating type-1 error rates $(\alpha = 0.05)$ of the four meta-analytic approaches described above, at $(\theta = 0)$ for the fixed-effect methods and $(\mu = 0;\ \tau^2 = 0)$ for the random-effects methods, showed type-1 error rates that were generally well controlled among the four methods, with the Han-Eskin random-effects model being slightly more conservative (Neupane et al., 2012). Notably, an earlier simulation study by Han and Eskin comparing type-1 error rates $(\alpha = 0.05)$ of the Han-Eskin random-effects method with those for traditional fixed and random-effects models, had shown the traditional random-effects model to be the more conservative method of the three meta-analytic approaches, especially in small meta-analyses where studies had unequal weights (Han and Eskin, 2011). Also notable, the simulation study by Han and Eskin was based on 100,000 replicates, whilst the simulation by Neupane and colleagues involved 10,000 replicates (Neupane et al., 2012); one would expect less type-1 error variation under simulation studies performed by Han and Eskin due to the higher number of replicates. A possible explanation for the conservative false-positive rate of the traditional random-effects method in smaller meta-analyses is that between-study variance estimates would be based on sparse information, leading to overestimation of heterogeneity (Han and Eskin, 2011).

In addition to type-1 error studies both sets of authors performed comparative power analyses of the Han-Eskin random-effects model and traditional fixed and random-effects methods (Han and Eskin, 2011; Neupane et al, 2012); they both assessed the performance of the

Han-Eskin random-effects model in meta-analysis simulations with different numbers of studies and levels of heterogeneity. In particular, Han and Eskin focussed on evaluating the power of the Han-Eskin random-effects model to detect loci conferring larger effects (odds ratios: 1.2, 1.3) in smaller meta-analyses (i.e. 5 and 10 studies) (Han and Eskin, 2011). Neupane and colleagues also concentrated on smaller meta-analyses (studies: 2, 3, 5, 7, 10), but interrogated a wider range of effect-sizes (odds ratios: 1.00, 1.05, 1.11, 1.16, 1.22, 1.28, 1.35) (Neupane et al., 2012). However, neither had explored the potential susceptibility of the Han-Eskin random-effects model to inflated genetic signals. In addition, the comparative power analyses in both groups did not assess the performance of the Han-Eskin random-effects model in meta-analyses scenarios with larger numbers of studies, akin to those found in recent complex trait meta-analyses projects (e.g. CARDIoGRAMplusC4D meta-analysis).

With a substantial number of complex trait meta-analysis projects focussed on the discovery of variants conferring increasingly smaller effects, the potential power advantage offered by the Han-Eskin random-effects model could play an important role in the identification of small effect heterogeneous loci, overlooked by traditional fixed and random-effects methods; with this in mind the question comparing the power of the Han-Eskin random-effects model to that of traditional fixed and random-effects methods is re-visited, to ascertain the following:

1) Can the Han-Eskin random-effects model improve on the power to detect small effect heterogeneous loci in larger meta-analyses, or is its power advantage more pronounced in meta-analyses with smaller numbers of studies?

2) Is the Han-Eskin random-effects model more vulnerable to inflated genetic signals than traditional fixed and random-effects methods?

3) Can the Han-Eskin random-effects model identify small effect heterogeneous loci overlooked by traditional fixed and random-effects methods? To that end, this chapter centres on two specific aims:

Aim 1: Assess the power of the Han-Eskin random-effects model to identify small effect heterogeneous loci and its vulnerability to inflated genetic signals, in relation to traditional fixed and random-effects methods, using Monte-Carlo simulation studies.

Aim 2: Re-appraise the CARDIoGRAMplusC4D meta-analysis of CAD risk with the Han-Eskin random-effects model, to identify risk variants overlooked by traditional fixed and random-effects methods, in the primary analysis.

4.2 Methods

4.2.1 A power comparison of the Han-Eskin random-effects model and conventional fixed and random-effects methods.

Monte-Carlo meta-analysis simulations were performed to compare the power of the Han-Eskin random-effects model to detect genetic associations, with that of traditional fixed and random-effects methods. Traditional fixed and random-effects meta-analyses were conducted with the metareg command (Harbord and Higgins, 2008) in Stata® versions 10.1 and 13.0, where the extent of heterogeneity was estimated using a method-of-moments algorithm (DerSimonian and Laird, 1986). Han-Eskin random-effects meta-analyses were performed using the METASOFT software package, version v2.0.0 (Han and Eskin, 2011).

4.2.2 Identifying small effect heterogeneous loci overlooked by traditional fixed and random-effects methods, in the CARDIoGRAMplusC4D GWAS meta-analysis.

An additive-effects only analysis of the 9,455,778 variants in the CARDIoGRAMplusC4D meta-analysis of CAD risk was performed using the Han-Eskin random-effects method, with the aim of identifying risk variants overlooked by traditional meta-analysis models. Similar to the simulation studies, the Han-Eskin random-effects meta-analysis was carried out with the METASOFT software package, version v2.0.0 (Han and Eskin, 2011).

4.3 Results

4.3.1. Aim 1: Simulation studies to assess the power of the Han-Eskin random-effects model in relation to traditional fixed and random-effects meta-analysis methods.

Three sets of Monte-Carlo simulation studies were performed to evaluate the power of the Han-Eskin random-effects method to detect small effect loci, and to assess its susceptibility to inflated genetic associations.

The first set of simulation studies involved a comparative power analysis of the Han-Eskin random-effects model and traditional fixed and random-effects methods, in the absence of heterogeneity; this was done to evaluate whether the performance (i.e. power) of the Han-Eskin random-effects model would resemble that of the traditional fixed-effect method in the absence of heterogeneity. The general expectation was that the traditional fixed-effect method would have similar power to the Han-Eskin random-effects model or slightly outperform it. This is because the design of random-effects models includes a partitioning of the total observed variation into sampling variance and between-study variance; as such, random-effects methods occasionally report non-zero heterogeneity estimates when study effects are in fact homogeneous ($\tau^2 = 0$), leading to unnecessarily conservative summary effect estimates, this is

more common in meta-analyses based on smaller numbers of studies where heterogeneity estimates are based on sparse information (e.g. 5 studies).

The traditional fixed-effect method is slightly more powerful than the Han-Eskin random-effects model in the absence of heterogeneity: To compare the Han-Eskin random-effects method's power to detect small effect homogeneous genetic signals ($\tau^2 = 0$) with that of traditional fixed and random-effects methods, three genetic association meta-analysis scenarios were simulated with different numbers of studies (5, 10, 40), each based on 300 replicates (Table 4.1). The numbers of studies were chosen to consider meta-analyses where there might be small, medium or larger numbers of participating studies. Moreover, the choice of 300 replicates was convenient in terms of the computational resources required to carry out the simulations and the precision of the 95% Binomial confidence intervals at the 5% significance level [CI: 0.0283, 0.0811]. To simplify the simulations, studies in each meta-analysis scenario were equally weighted (standard error of $\log_e$(odds ratio) = 0.08), and the variant effect-size set at ($\log_e$(odds ratio) = 0.095 i.e. odds ratio = 1.1) to model a small effect homogeneous locus ($\tau^2 = 0$). Notably, the variant effect-size and corresponding standard error were inspired by the mean genetic effect-size and median standard error in the CARDIoGRAMplusC4D consortium's meta-analysis of CAD risk (CARDIoGRAMplusC4D Consortium, 2015). Given that the three meta-analysis scenarios were simulated without heterogeneity, the observed variation simply reflected sampling variance; accordingly, the I^2 values presented in (Table 4.1) reflect within-study variation that was misinterpreted as heterogeneity.

As expected, the traditional fixed-effect method had slightly more power to detect genetic signals in the absence of heterogeneity, than the Han-Eskin random-effects model (Table

4.1). Additionally, the power to detect homogeneous genetic signals improved as the number of studies increased, akin to a sigmoid curve. Of the three meta-analytic approaches, the traditional random-effects method was the least powerful potentially due to its conservative null hypothesis. This result is consistent with simulation studies conducted by Han and Eskin, where the traditional fixed-effect method was generally more powerful than random-effects models in the absence of heterogeneity (Han and Eskin, 2011).

Table 4.1: A comparative power analysis of the Han-Eskin random-effects model and conventional fixed and random-effects methods to detect small effect genetic associations in the absence of heterogeneity.

# Studies	$I^2(\%)$	Summary effect-size, β	SE(β)	Fixed-effect p-value	Random-effects p-value	Han-Eskin Random-effects p-value	Fixed effect power (%)	Random-effects power (%)	Han-Eskin Random-effects power (%)
5	16.88	0.095	0.044	0.072	0.097	0.077	0.33	0.00	0.33
10	11.60	0.095	0.030	0.010	0.015	0.013	2.33	0.67	2.00
40	7.33	0.094	0.014	3.33E-08	2.75E-07	7.56E-08	94.00	87.33	91.33

The table shows three genetic association meta-analysis scenarios, with different numbers of studies (5, 10, 40). The three meta-analysis scenarios were generated using Monte-Carlo simulation experiments, each based on 300 replicates. Studies in each meta-analysis scenario were equally weighted (i.e. standard error of log_e(odds ratio) = 0.08) and the variant effect-size fixed at, log_e(odds ratio) = 0.095 (i.e. odds ratio = 1.1) to model a small effect homogeneous locus ($\tau^2 = 0$). In each scenario, averages across the 300 replicates are presented for summary association p-values and the heterogeneity inconsistency index, I^2. The power to detect a significant genetic association was calculated as the fraction of Monte-Carlo replicates reporting p-values $\leq 5 \times 10^{-8}$. The traditional fixed-effect method showed (slightly) greater power, to detect genetic associations in the absence of heterogeneity, across the three meta-analysis scenarios explored. Notably, τ^2: denotes the amount of heterogeneity in a meta-analysis, also referred to as between-study variance. β represents a summary effect-size expressed on the log_e(odds ratio) scale and SE(β) its corresponding standard error. Source: Lerato Magosi (Farrall and Hopewell labs).

The second set of simulation studies compared the power of the Han-Eskin random-effects model to detect heterogeneous genetic associations of varying effect-sizes and levels of heterogeneity with that of traditional fixed and random-effects methods. The expectation was that the Han-Eskin random-effects model would outperform traditional meta-analysis methods as the amount of heterogeneity increased; this was based on a combination of two factors: 1) the design of random-effects models naturally facilitates the detection of heterogeneous genetic signals (i.e. heterogeneity is acknowledged), and 2) the Han-Eskin random-effects model has a more relaxed null hypothesis than the traditional random-effects method.

The Han-Eskin random-effects model has greater power to detect small effect loci with moderate-to-high heterogeneity compared to traditional meta-analysis methods: To evaluate the power of the Han-Eskin random-effects model to identify heterogeneous genetic signals in relation to that of conventional fixed and random-effects methods; two sets of meta-analysis scenarios were simulated with different numbers of studies (5, 10, 40) and varying levels of heterogeneity (I^2 %: 25, 50, 75), yielding 18 scenarios altogether, each based on 300 replicates (Table 4.2). The first set of meta-analysis scenarios evaluated the Han-Eskin random-effects model's power to detect a small effect heterogeneous locus, ($\log_e$(odds ratio) = 0.0953; corresponding odds ratio: 1.1) whilst the second set assessed the power of the Han-Eskin random-effects model to identify a heterogeneous locus with a larger effect-size, ($\log_e$(odds ratio) = 0.2624; corresponding odds ratio: 1.3). Variant effect-sizes were inspired by magnitudes of genetic effects in the CARDIoGRAMplusC4D meta-analysis (CARDIoGRAMplusC4D Consortium, 2015). Furthermore, studies in each meta-analysis scenario were equally weighted (i.e. standard error of $\log_e$(odds ratio) = 0.08) and levels of heterogeneity chosen to reflect situations in which meta-analyses might have small ($I^2 = 25\%$), moderate ($I^2 = 50\%$) or

larger amounts ($I^2 = 75\%$) of heterogeneity. Though studies were assigned equal weights to simplify the simulation, in practice it is common for study sample-sizes to vary in a meta-analysis. Notably, a previous comparative power analysis of the Han-Eskin random effects model and traditional meta-analysis methods had shown that the relative performance of the three meta-analytic approaches in meta-analysis simulations where studies were equally precise was similar to that when participating studies had unequal sample sizes (Han and Eskin, 2011).

Table 4.2 shows that the power to detect heterogeneous genetic associations increased with the number of participating studies both under the Han-Eskin random-effects model and the traditional fixed and random-effects methods. As expected, the Han-Eskin random-effects model's power to detect small effect heterogeneous loci improved with increasing heterogeneity. Moreover, the Han-Eskin random-effects model showed greater power over traditional meta-analytic approaches to detect small effect heterogeneous loci ($\log_e$(odds ratio) = 0.0953 i.e. odds ratio = 1.1) with moderate-to-high heterogeneity ($I^2 \geq 40\%$), especially in meta-analysis scenarios comprising fewer studies ($\leq$ 10 studies). The power to identify heterogeneous genetic associations with larger effect-sizes ($\log_e$(odds ratio) = 0.2624 i.e. odds ratio = 1.3) was similar under both the Han-Eskin random-effects model and traditional fixed and random-effects methods. This result confirms the findings of simulation studies by Han and Eskin, where the Han-Eskin random-effects method was more powerful than traditional fixed and random-effects methods, at identifying small effect heterogeneous genetic associations (Han and Eskin, 2011).

Table 4.2: A comparative power analysis of the Han-Eskin random-effects model and conventional fixed and random-effects methods to detect heterogeneous genetic associations.

Simulated effect-size, OR (log_eOR)	# Studies	β log_eOR	SE(β)	$I^2(\%)$	Fixed-effect p-value	Random-effects p-value	Han-Eskin Random-effects p-value	Fixed-effect power (%)	Random-effects power (%)	Han-Eskin Random-effects power (%)
1.1 (0.0953)	5	0.0999	0.05	23.04	0.10	0.13	0.10	0.67	0.00	0.67
		0.0964	0.05	35.32	0.11	0.18	0.09	1.67	0.33	2.00
		0.0962	0.07	61.23	0.16	0.29	0.06	5.67	0.33	13.00
	10	0.0933	0.03	22.43	0.02	0.04	0.02	2.33	2.00	3.33
		0.0916	0.04	43.00	0.06	0.10	0.03	5.33	1.33	9.00
		0.1001	0.05	68.21	0.08	0.18	0.01	19.00	1.33	43.00
	40	0.0954	0.02	23.75	6.29E-07	6.27E-06	8.79E-07	90.67	62.33	90.67
		0.0938	0.02	47.05	7.35E-06	4.28E-04	2.29E-06	82.67	27.33	92.33
		0.0943	0.03	74.08	1.60E-03	1.55E-02	1.39E-10	77.67	3.67	100.00
1.3 (0.2624)	5	0.2626	0.05	22.47	7.40E-07	1.92E-04	1.06E-06	86.00	59.00	85.67
		0.2695	0.05	37.08	6.23E-05	2.56E-03	5.49E-05	85.33	43.33	86.33
		0.2571	0.07	60.21	3.61E-03	2.37E-02	1.12E-03	71.67	19.00	80.00
	10	0.2621	0.03	23.48	9.72E-12	1.21E-06	9.69E-12	99.67	95.33	100.00
		0.2605	0.04	42.21	1.34E-11	1.25E-05	1.06E-11	99.67	76.67	100.00
		0.2637	0.05	68.14	7.12E-09	6.82E-04	3.56E-10	98.33	37.67	99.67
	40	0.2626	0.02	22.51	2.96E-59	2.60E-34	2.17E-58	99.33	100.00	100.00
		0.2619	0.02	45.66	2.49E-54	7.95E-22	1.82E-57	99.67	100.00	100.00
		0.2633	0.03	73.48	1.15E-47	4.59E-12	2.88E-59	100.00	100.00	100.00

The table shows 18 meta-analysis scenarios with varying numbers of studies, (5, 10, 40) and levels of heterogeneity (I^2 % : $25, 50, 75$), each based on 300 Monte-Carlo replicates. The first set of 9 meta-analysis scenarios assess the power to detect a small effect heterogeneous locus, (log_e(odds ratio) = 0.0953; corresponding odds ratio: 1.1), whilst the second set of 9 scenarios evaluate the power to identify a heterogeneous genetic association with a larger effect-size, (log_e(odds ratio) = 0.2624; corresponding odds ratio: 1.3). Additionally, studies in each meta-analysis scenario were assigned equal weights (i.e. standard error of log_e(odds ratio) = 0.08). Furthermore, within each scenario, averages across the 300 replicates are presented for summary association p-values and the heterogeneity inconsistency index, I^2. The power to detect a significant genetic association was calculated as the fraction of Monte-Carlo replicates reporting p-values $\leq 5 \times 10^{-8}$. The Han-Eskin random-effects model generally outperforms the traditional fixed-effect method at identifying small effect heterogeneous loci, (log_e(odds ratio) = 0.0953), particularly in smaller meta-analyses where there is moderate-to-high heterogeneity ($I^2 \geq 40\%$). Notably, OR: denotes an odds ratio. β, represents a summary effect-size expressed on the log_e(odds ratio) scale and SE(β) its corresponding standard error. Source: Lerato Magosi (Farrall and Hopewell labs).

The third set of simulations compared the Han-Eskin random-effects model's vulnerability to inflated genetic signals with that of traditional fixed and random-effects meta-analysis methods. Given that the Han-Eskin random-effects method considers the null hypothesis that there is either a non-null average genetic effect or substantial heterogeneity in a meta-analysis, the expectation was that the Han-Eskin random-effects model would be more likely to report inflated genetic associations at small effect loci with high heterogeneity, compared to traditional meta-analysis methods.

Assessing the vulnerability of the Han-Eskin random-effects model to inflated genetic signals precipitated by high heterogeneity: To evaluate the Han-Eskin random-effects model's vulnerability to inflated genetic signals, nine meta-analysis scenarios were explored through Monte-Carlo simulations, each based on 300 replicates, where a single outlier study showing extreme effects was included in each meta-analysis scenario (Table 4.3). The nine scenarios explored are presented in groups of three, where each group comprises three meta-analysis scenarios with different numbers of studies (5, 10, 40). In each of the three groups, the effect-size of the outlier study is varied to show progressively larger effects (odds ratios: 1.1, 1.3, 1.6; corresponding $\log_e$(odds ratio)s: 0.095, 0.262, 0.470); whilst the non-outlier study effects are null ($\log_e$(odds ratio) = 0 i.e. odds ratio = 1.0). Additionally, studies were assigned equal weights in each meta-analysis scenario (i.e. standard error of $\log_e$(odds ratio) = 0.08; corresponding odds ratio: 1.08).

As expected, the extent of heterogeneity in the meta-analysis scenarios explored increased with the variant effect-size in the outlier study (Table 4.3). Moreover, the inflation of summary effect-size estimates was slightly greater when there were few participating studies (e.g. 5 studies); this was not surprising as smaller meta-analyses generally have less information

available to accurately estimate the amount of heterogeneity. The inclusion of an outlier study showing extreme effects had a larger impact on p-values of the Han-Eskin random-effects model than those for conventional meta-analysis methods, especially when there were relatively high amounts of heterogeneity ($I^2 > 60\%$) (Table 4.3.) In particular, the Han-Eskin random-effects method reported nominally significant p-values (i.e. $P < 0.05$) for most of the meta-analysis scenarios with higher levels of heterogeneity ($I^2 > 60\%$), despite the majority of the participating studies in each meta-analysis scenario showing null effects. Additionally, compared to the traditional fixed and random-effects models, there was a greater increase in the power of the Han-Eskin random-effects method to detect heterogeneous genetic associations as the variant effect-size of the outlier study grew more extreme.

Table 4.3: A comparison of the Han-Eskin random-effects model's vulnerability to inflated genetic signals in relation to traditional fixed and random-effects methods.

# Studies, S	Non-outlier studies, S - 1 Oddsratio	Outlier study Oddsratio	Summary effect-size, β	$I^2(\%)$	Fixed-effect p-value	Random-effects p-value	Han-Eskin Random-effects p-value	Fixed-effect power (%)	Random-effects power (%)	Han-Eskin Random-effects power (%)
5	1	1.1	1.02	22.93	0.44	0.49	0.39	11.33	7.33	12.33
10			1.01		0.45	0.50	0.39	11.00	6.33	14.00
40			1.00		0.49	0.54	0.26	5.00	3.00	24.33
5	1	1.3	1.06	67.66	0.24	0.44	0.02	32.00	8.00	89.00
10			1.03		0.35	0.49	0.05	17.67	5.67	63.33
40			1.01		0.46	0.53	0.10	8.00	2.00	60.33
5	1	1.6	1.10	86.41	0.07	0.38	3.18E-06	71.67	2.67	100.00
10			1.05		0.20	0.45	4.39E-05	38.00	3.33	100.00
40			1.01		0.40	0.51	2.64E-03	19.33	5.67	99.67

The table shows 9 genetic association meta-analysis scenarios, each based on 300 Monte-Carlo replicates. Each meta-analysis scenario, comprised a single outlier study showing extreme effects whilst the rest of the participating study effect-sizes were null (log_e(odds ratio) = 0 i.e. odds ratio = 1.0). Additionally, studies in each meta-analysis scenario were assigned equal weights (i.e. standard error of log_e(odds ratio) = 0.08; corresponding odds ratio: 1.08). The nine meta-analysis scenarios are presented row-wise in three sets of three, where each set comprises three meta-analysis scenarios with different numbers of studies (5, 10, 40). And the effect-size for the outlier study is varied across the three sets to show increasingly larger effects (odds ratios: 1.1, 1.3, 1.6; corresponding log_e(odds ratio)s: 0.095, 0.262, 0.470). The p-values presented in each meta-analysis scenario denote the average summary association p-values, for each of the three meta-analytic approaches, across the 300 replicates. Furthermore, the power to detect an inflated genetic association was calculated as the proportion of replicates (out of 300) reporting, $P \leq 0.05$. The Han-Eskin random-effects model showed greater power to detect inflated (i.e. exaggerated) genetic associations at higher heterogeneity levels ($I^2 > 60\%$). Notably, I^2 denotes the Higgin's inconsistency index and β, represents a summary effect-size expressed on the log_e(odds ratio) scale. Source: Lerato Magosi (Farrall and Hopewell labs).

Given that significant p-values under the Han-Eskin random-effects model can either represent an actual genetic association or extensive heterogeneity in a meta-analysis, this result has practical implications for the discovery of genetic signals in GWAS meta-analyses with moderate-to-high heterogeneity levels. In that, a reliance on p-values generated by the Han-Eskin random-effects method alone without the inspection of forest plots can lead to p-values that represent substantial heterogeneity - potentially due to population structure (i.e. local ancestry differences at individual variants) - being misinterpreted as significant genetic associations with the trait of interest. That is, variants in linkage equilibrium with the causal variant(s) appearing to be associated with the trait of interest as a result of their effect alleles being more common in the case groups (i.e. affected populations) of certain ancestral populations in the meta-analysis.

Therefore, when using the Han-Eskin random-effects model to interrogate small effect heterogeneous loci in the presence of moderate-to-high heterogeneity, I would recommend inspecting forest plots for heterogeneity patterns that show inflated average genetic effect estimates owing to a few outlier studies showing out-sized effects; in scenarios where such heterogeneity patterns are encountered, it might be worth considering the potential contribution of genotyping error or local population structure to the observed genetic association.

Notably, Wang and colleagues have developed two association tests that can be applied to stand-alone genome-wide association studies to evaluate genetic associations whilst adjusting for both local and global population structure (Wang et al., 2011). One of the tests is a likelihood ratio test based on a conditional likelihood ratio framework, and the other involves a simple regression of the variant of interest employing the estimated local ancestry proportion at that variant as a covariate. Also notable is that simulation studies by Wang and colleagues have shown that some of the tools popularly used to correct for global population structure e.g.

genomic control (Devlin and Roeder, 1999) and EIGENSTRAT (a principal components based method) (Price et al., 2006), are relatively less efficient at adjusting for local ancestry differences at individual variants due to higher type-1 error rates; thus when examining local admixture it might be useful to consider methods better geared towards adjusting for local ancestry differences.

4.3.2 Aim 2: A re-appraisal of the CARDIoGRAMplusC4D meta-analysis of coronary artery disease, to identify small effect heterogeneous risk loci overlooked by traditional fixed and random-effects methods.

The comparative power analyses presented in Aim 1 (section 4.3.1) have shown that the Han-Eskin random-effects model outperforms traditional fixed and random-effects methods in detecting small effect genetic associations (e.g. odds ratio = 1.1) amidst moderate-to-high heterogeneity ($I^2 > 40\%$). This provides an opportunity where GWAS meta-analyses previously performed using traditional fixed and random-effects models can be re-examined with the Han-Eskin random-effects method in an effort to identify additional variants that underlie traits of interest. Given the diverse ancestry groups and ascertainment criteria in the CARDIoGRAMplusC4D meta-analysis (CARDIoGRAMplusC4D Consortium, 2015), it was of interest to search for risk variants that might have eluded detection by traditional fixed and random-effects methods, owing to heterogeneity. In particular to identify risk variants suggestively associated ($P < 5 \times 10^{-5}$) with CAD under the traditional fixed and random-effects methods, that show significant association ($P < 5 \times 10^{-8}$) under the Han-Eskin random-effects model; the rationale being that the power advantage offered by the Han-Eskin random-effects model in the presence of heterogeneity, might help to reveal CAD risk variants that were miss-specified as suggestively associated, due to inadequate power.

Prior to performing a Han-Eskin random-effects meta-analysis of the ~ 9 million variants in the CARDIoGRAMplusC4D dataset, variant effect-size estimates (i.e. β coefficients

scaled as $\log_e$(odds ratios)) were appropriately "flipped" to align effect alleles; this involved substituting β coefficients in studies where effect alleles were reversed with their additive inverse (i.e. β-flipped $= -1\ x\ \beta$). Additionally, a study-level genomic correction (Devlin and Roeder, 1999) was applied to the corresponding standard errors of the β coefficients, to minimise global (or genome-wide) inflation of association statistics. As highlighted in chapter 2 (i.e. section 2.3.2), these were variants with minor allele frequencies > 0.5% that were well imputed in a least 60% of the 48 studies in the CARDIoGRAMplusC4D dataset.

Of the 9,455,778 variants examined for additive associations under the Han-Eskin random-effects method, 8,214 were suggestively associated ($P < 5 \times 10^{-5}$) with CAD under traditional fixed and random-effects meta-analysis models (Table 4.4); and of the 8,214 suggestively associated variants, 40 showed significant associations ($P < 5 \times 10^{-8}$) with CAD under the Han-Eskin random-effects model. Grouping the 40 variants by genetic distance such that lead variants would be centred on a window of $\pm$ 0.5 centiMorgans, yielded a single discrete locus with rs688359 as the lead variant (chromosome: 6; location: 177.05 centiMorgans; $P_{RE(new)} <$ 1.34e-08; $P_{FE} <$ 1.68e-07; $P_{RE} <$ 4.02e-06) (Table 4.4). Whilst long range linkage disequilibrium relationships can extend beyond 0.5 centiMorgans; the strength of linkage disequilibrium typically decays steeply within 0.5 centiMorgans, hence a genetic distance window of $\pm$ 0.5 centiMorgans was preferred.

Table 4.4: A re-appraisal of the CARDIoGRAMplusC4D meta-analysis of coronary artery disease with the Han-Eskin random-effects model, to identify loci overlooked by traditional fixed and random-effects methods.

Category	Description	No. Variants / Loci
1	Total number of variants examined for additive associations in the Han-Eskin random-effects meta-analysis.	9,455,778
2	Subset of variants from category 1 that were suggestively associated ($P \leq 5 \times 10^{-05}$) with coronary artery disease under the traditional fixed and random-effects methods.	8,214
3	Subset of the 8,214 suggestively associated ($P \leq 5 \times 10^{-05}$) variants under the traditional fixed and random-effects methods, that showed statistically significant ($P \leq 5 \times 10^{-08}$) effects under the Han-Eskin random-effects model.	40
4	Number of loci obtained after grouping the 40 variants in category 3 by a genetic distance window of ± 0.5 centiMorgans around the lead variant at each locus	1

A single locus was identified on chromosome 6 with rs688359 as the lead variant (chromosome: 6; location: 177.05 cM; I^2: 34.06%; $P_{RE(new)} < 1.34 \times 10^{-08}$; $P_{FE} < 1.68 \times 10^{-07}$; $P_{RE} < 4.02 \times 10^{-06}$). Notably, I^2 denotes the Higgin's inconsistency index. Furthermore, $P_{RE(new)}$, P_{FE} and P_{RE} represent association p-values for the Han-Eskin random-effects model and the traditional fixed and random-effects methods, respectively. Source: Lerato Magosi (Farrall and Hopewell labs).

An inspection of the distribution of study effects at rs688359 revealed relatively low heterogeneity, $I^2 = 34.06\%$ (Figure 4.2), and there were no obvious outlier studies showing extreme effects; therefore, there was little evidence to suggest that the heterogeneity pattern observed at rs688359 was an artefact of population structure, precipitated by local ancestry differences. With this in mind, we next used regional association plots to explore the extent of correlation between rs688359 and neighbouring CAD risk loci, so as to determine whether rs688359 was linked to known CAD loci or rather constituted an independent CAD locus.

Notably, rs688359 is an intron variant of the insulin-like growth factor 2 receptor gene (*IGF2R*), and lies on the boundary of the *LPA* (lipoprotein (a)) locus. The *LPA* gene codes for the apolipoprotein(a) component of the Lp(a) lipoprotein particle (Clarke et al., 2009). In addition, the *LPA* locus strongly associates with CAD, and harbours several variants independently associated with increased Lp(a) lipoprotein level, and elevated CAD risk (Clarke et al., 2009).

As such, regional association plots of the *LPA* locus indexed on rs55730499 (location: 177.56) as the lead variant, were generated for p-values produced both by the Han-Eskin random-effects model and traditional fixed and random-effects methods, to determine whether rs688359 might be part of the extended *LPA* architecture (Figure 4.3A-C). The expectation was that the regional association plot generated with p-values from the Han-Eskin random-effects meta-analysis would have two peaks if rs688359 represented an independent CAD locus, one around rs688359 and another at rs55730499; whilst the regional association plots from the conventional meta-analysis methods would only show a single broad peak around rs55730499. There was a single broad peak around rs55730499 in each of the three regional association plots (Figure 4.3A-C), suggesting rs688359 might be part of the extended *LPA* architecture.

However, there was little correlation between rs688359 and rs55730499, leaving room for the possibility that rs688359 might be part of a non-*LPA*-linked CAD risk locus or be independently associated with CAD risk. Accordingly, a regional association plot indexed on rs688359 as the lead variant was generated to explore the level of correlation between rs688359 and other neighbouring CAD loci e.g. *LPAL2* (lipoprotein(a)-like 2) and *SLC22A3* (solute carrier family 22 member 3) (Figure 4.4); however, there was little correlation between rs688359 and other known CAD risk loci nearby. Consequently, the Han-Eskin random-effects meta-analysis might have identified an additional risk variant (rs688359) independently associated with CAD that is worthy of follow-up.

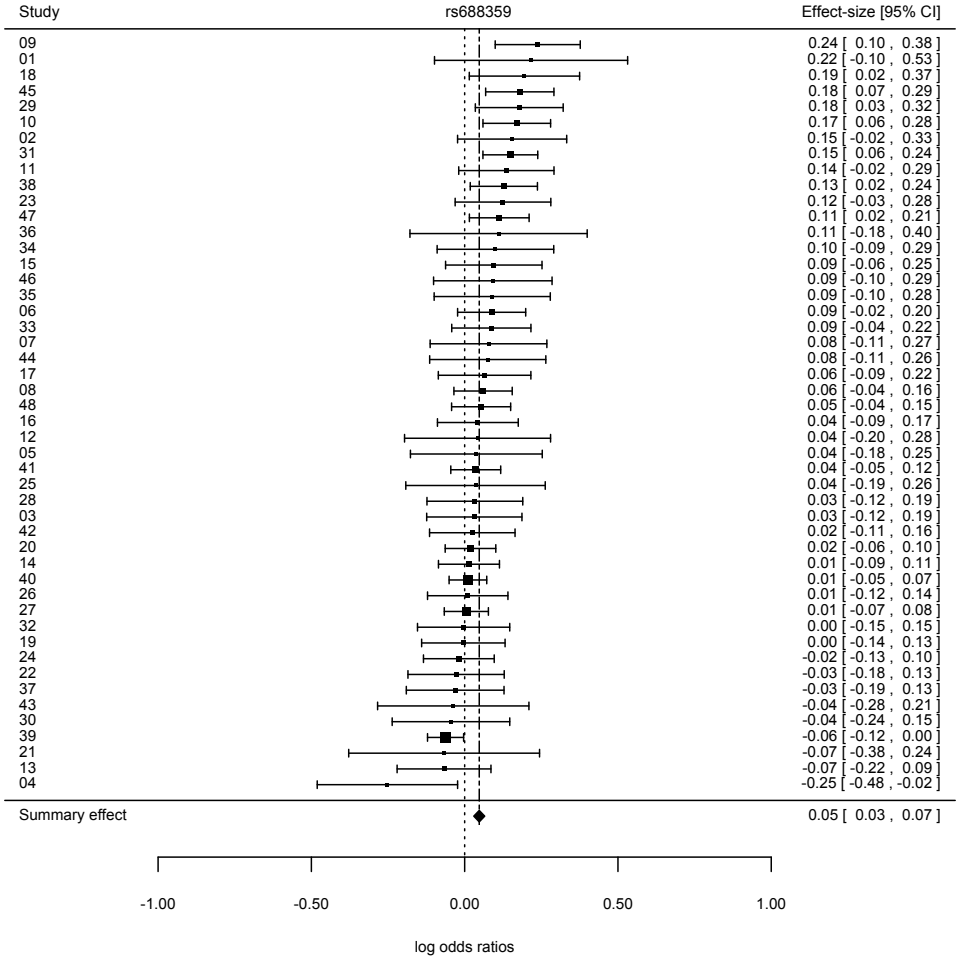


Figure 4.2: A forest plot of effect-size estimates (log odds ratios) at rs688359 (chromosome 6) highlighting locus specific heterogeneity among studies in the CARDIoGRAMplusC4D meta-analysis, $I^2 = 34.06\%$**.** Sorted effect-size estimates are presented for individual studies represented by filled squares with their 95% confidence intervals shown by horizontal lines; the sizes of the squares are proportional to each studies' inverse-variance weighting. Additionally, a filled diamond represents the summary effect-size estimate. There is relatively low heterogeneity at the rs688359 lead variant. Notably, I^2 denotes the Higgin's inconsistency index. Source: Lerato Magosi (Farrall and Hopewell labs).

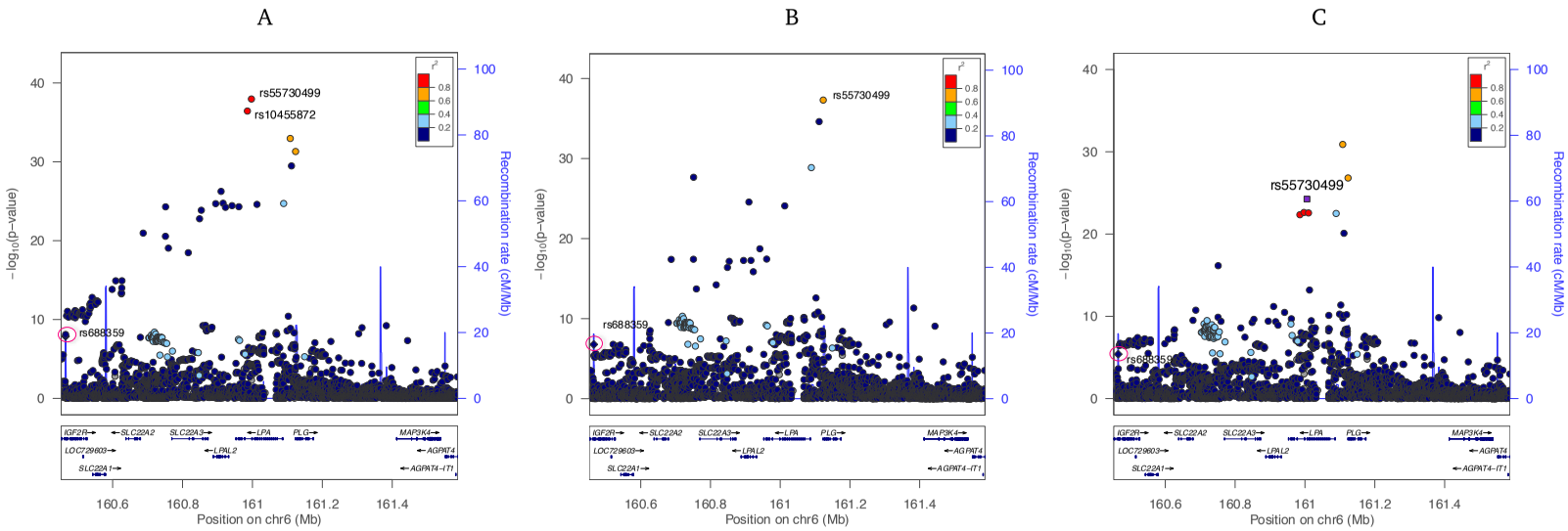


Figure 4.3: Regional association plots of the *LPA* locus highlighting rs688359. Three regional association plots of the *LPA* locus indexed on rs55730499 (lead variant) are shown for p-values produced by the Han-Eskin random-effects model (panel A), and the conventional fixed (panel B) and random-effects methods (panel C). The three regional association plots show little correlation between rs688359 and the strongly associated coronary artery disease risk locus, rs55730499. The reference population employed for LD and recombination rate calculations was hg19/1000 Genomes Nov. 2014 EUR. The figure was generated using LocusZoom. Source: Lerato Magosi (Farrall and Hopewell labs).

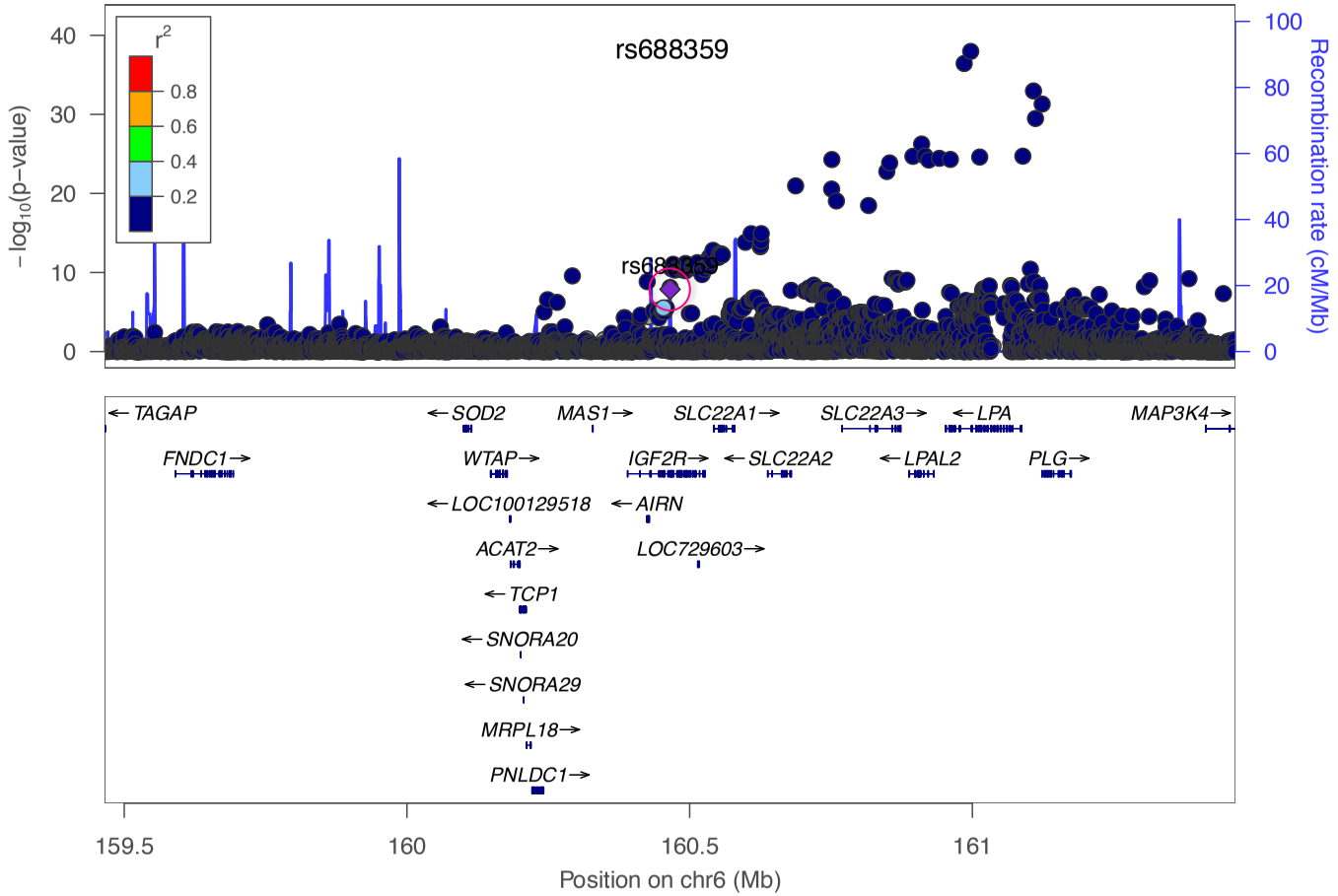


Figure 4.4: A regional association plot centred on the rs688359 (chromosome 6) lead variant. The regional association plot shows little correlation between rs688359 and other well known coronary artery disease risk loci (e.g. *SLC22A3* and *LPAL2*) in the extended *LPA* region. The reference population employed for LD and recombination rate calculations was hg19/1000 Genomes Nov. 2014 EUR. The figure was generated using LocusZoom. Source: Lerato Magosi (Farrall and Hopewell labs).

4.4 Discussion

4.4.1 Improving power to detect small effect genetic associations amidst moderate-to-high heterogeneity

Fixed-effect meta-analysis models are traditionally more powerful than random-effects methods at detecting genetic associations, even at heterogeneous loci where the random-effects approach should ideally have more power (Han and Eskin, 2011); this is because the traditional random-effects model has a conservative null hypothesis, which assumes the presence of heterogeneity under the null. By contrast, the design of fixed-effect models disregards heterogeneity under the null, allowing a sharper distinction between the null and alternative hypotheses. The conservative nature of the traditional random-effects method has led to a reliance on the fixed-effect method for discovery of variants underlying complex traits.

In order to achieve the requisite power to detect complex trait variants conferring smaller effects, meta-analysis projects have become more collaborative, assembling summary results of genome-wide association studies based on diverse ancestry groups and ascertainment criteria through international consortia. Examples of consortia established to identify genetic determinants of complex traits include: IEC for endometriosis (Painter et al., 2011; Sapkota et al., 2017), METASTROKE (Malik et al., 2016; Traylor et al., 2012), GIANT for anthropometric traits (Wood et al., 2014) and CARDIoGRAMplusC4D for CAD (CARDIoGRAMplusC4D Consortium, 2015). As these consortia continue to grow and GWAS meta-analysis datasets

become larger and likely more heterogeneous, it is reasonable to expect that genetic signals with moderate-to-high heterogeneity ($I^2 \geq 40\%$) might elude detection under the traditional fixed-effect method. Additionally, fixed-effect meta-analyses of heterogeneous variants can be misleading when there is moderate-to-high heterogeneity as the summary effect estimates do not acknowledge the extent of heterogeneity in the meta-analysis, leading to unduly small p-values (Thompson et al., 2011).

New fixed and random-effects meta-analytic approaches have recently been proposed to improve power for discovery of heterogeneous genetic associations (Han and Eskin, 2011; Lebrec et al., 2010). Simulation studies in aim 1 (section 4.3.1), confirmed that the Han-Eskin random-effects method is more powerful than the traditional fixed-effect model at identifying genetic signals in the presence of moderate-to-high heterogeneity ($I^2 \geq 40\%$), particularly when variant effect-sizes are small (e.g. $\log_e$(odds ratio) = 0.0953; corresponding odds ratio: 1.1), and meta-analyses comprise few studies (e.g. $\leq$ 10 studies). When heterogeneity levels were low ($I^2 < 40\%$) however, the traditional fixed-effect method and Han-Eskin random-effects model generally showed similar power to detect genetic associations (Table 4.2). The traditional fixed-effect method was slightly more powerful than the Han-Eskin random-effects model in the absence of heterogeneity (Table 4.1).

The power of the Han-Eskin random-effects model generally improved with increasing heterogeneity, but so did its vulnerability to inflated genetic signals (Table 4.3); that is, compared to the traditional fixed and random-effects models, the Han-Eskin random-effects method was more likely to report nominally significant p-values (i.e. $P < 0.05$) for small effect loci with relatively high heterogeneity ($I^2 > 60\%$). Considering that the genetic associations observed in

(Table 4.3) were largely driven by an outlier study in each meta-analysis scenario showing extreme effects whilst the rest of the non-outlier study effects were null; the nominal p-values produced by the Han-Eskin random-effects model were likely a reflection of the heterogeneity induced by the inflated effects in the outlier studies. Since local ancestry differences (i.e. population structure) at individual variants have the potential to induce similar heterogeneity patterns, I would recommend that reports which either identify novel lead variants or validate existing ones using the Han-Eskin random-effects method be accompanied by forest plots, in order to ensure credibility of the results; particularly for highly heterogeneous lead variants conferring smaller effect-sizes. Further to inflated genetic signals owing to population structure, inflated genetic associations can arise as a result of genotyping error, wherein the same outlier study (or set of studies) shows extreme effects at several variants in the meta-analysis; therefore I would also recommend checking whether inflated genetic associations at variants of interest might be driven by the same set of studies.

After establishing through simulation studies that the Han-Eskin random-effects model outperforms traditional meta-analysis methods in detecting genetic associations with smaller effect-sizes in the presence of moderate-to-high heterogeneity, the Han-Eskin random-effects model was then used to re-examine a recent meta-analysis of CAD risk (CARDIoGRAMplusC4D Consortium, 2015); the aim was to identify small effect heterogeneous risk variants overlooked by traditional fixed and random-effects models. A single locus was identified on chromosome 6q26 with rs688359 as the lead variant (Figure 4.2). Notably, rs688359 is an intron variant of the insulin-like growth factor 2 receptor gene (*IGF2R*), which regulates cell proliferation indirectly through clearance of IGF2 (insulin-like growth factor 2) proteins from the cell surface (Bergman et al., 2013). The insulin-like growth factor 2 receptor

is also involved in Golgi- trafficking, where it transports newly synthesised lysosome acid hydrolases tagged with mannose-6 phosphate from the trans-Golgi network to the lysosomes (Bergman et al., 2013).

Though the proximity of the rs688359 lead variant (location: 177.05 centiMorgans) to the *LPA* locus (location of lead variant, rs55730499: 177.56 centiMorgans) might have hinted at the possibility that rs688359 forms part of the extended *LPA* architecture; regional association plots of the *LPA* locus showed little correlation between rs688359 and rs55730499. Additionally, regional association plots centred on rs688359 showed little correlation between it (rs688359) and other well known CAD loci in the extended *LPA* region including: *SLC22A3* and *LPAL2*; suggesting that the Han-Eskin random-effects meta-analysis had potentially identified an additional variant independently associated with CAD risk. However, it will be important to replicate this association in an independent dataset (e.g. the UK Biobank dataset). Notably, the UK Biobank dataset is a large prospective study based on 500,000 participants aged 45-69 years. In line with this finding, Nelson and colleagues have recently demonstrated through the UK Biobank dataset, that some of the variants characterised as suggestively associated with CAD risk in the CARDIoGRAMplusC4D meta-analysis (CARDIoGRAMplusC4D Consortium, 2015) show significant association with CAD risk when re-examined in a larger, more powerful meta-analysis (Nelson et al., 2017).

The considerable overlap between variants identified as associated with CAD risk under the traditional fixed-effect method and the Han-Eskin random-effects meta-analysis suggests that the traditional fixed-effect method has adequate power to identify heterogeneous variants with smaller effect-sizes, when employed in meta-analyses based on a larger numbers of studies. Therefore, the practical benefit of performing a Han-Eskin random-effects meta-analysis to

identify small effect heterogeneous loci is likely to be greater in meta-analyses comprising fewer studies.

The introduction of the Han-Eskin random-effects model as an application for discovery of trait variants in genetic association meta-analyses was met with questions as to whether contemporary meta-analytic approaches could work well in multi-ethnic meta-analyses, where participating studies had been genotyped on different platforms, and their variants imputed to create a common set for analysis (Neupane et al., 2012). By re-appraising the CARDIoGRAMplusC4D meta-analysis with the Han-Eskin random-effects model, this work has shown that the Han-Eskin random-effects model can be used successfully in multi-ethnic GWAS meta-analyses with imputed genotypes. Lastly, there is little evidence to suggest that there might be a sizeable number of small effect heterogeneous loci that have yet to be discovered in the CARDIoGRAMplusC4D dataset.

4.4.2 Advantages, limitations and alternative methods

The Han-Eskin random-effects model's main advantage is its strength over traditional meta-analytic methods when there are few contributing studies in a meta-analysis and variant effect-sizes are very small. However, some of the limitations of the Han-Eskin random-effects model are that it can be tricky to distinguish significant p-values that are due to an actual genetic association from those that reflect substantial heterogeneity in a meta-analysis.

Additionally, whilst the substantial overlap between variants identified by the Han-Eskin random-effects model and those detected by the traditional fixed-effect method can be helpful in verifying results produced by the traditional fixed-effect method, it would be more productive if the Han-Eskin random-effects method specifically reported variants that would typically be overlooked by the traditional fixed-effect method. In line with this, Lee and colleagues have recently proposed an alternative or update to the Han-Eskin random-effects model, which now focuses on reporting variants where the Han-Eskin random-effects model produces p-values as small or smaller than those provided by the traditional fixed-effect method i.e. ($P_{RE(new)} \leq P_{FE}$) (Lee et al., 2017). However, the limitation of having to distinguish significant p-values reported by the Han-Eskin random-effects method for authentic genetic signals from those reported as a result of extensive heterogeneity remains; forest plots can be very useful in this regard, as one can tell from the distribution of observed study effects, whether the average genetic effect estimate of the variant being examined has been inflated by a few outliers showing extreme effects.

4.5 Chapter summary

To summarise, this chapter evaluated the Han-Eskin random-effects method as an alternative approach to traditional fixed and random-effects models, which could improve power to identify heterogeneous variants conferring smaller effects. There were four main findings in this chapter. First, the Han-Eskin random-effects method is generally more powerful than traditional meta-analysis methods at detecting small effect heterogeneous loci, when there are moderate-to-high heterogeneity levels and few participating studies. Second, the Han-Eskin random-effects method is more likely to report inflated genetic associations at small effect loci when there is relatively high heterogeneity in a meta-analysis. Third, the larger overlap between variants identified by the traditional fixed-effect method and the Han-Eskin random-effects model in the CARDIoGRAMplusC4D meta-analysis (CARDIoGRAMplusC4D Consortium, 2015), suggests the traditional fixed-effect method has adequate power to facilitate discovery of small effect heterogeneous loci, in meta-analyses with larger numbers of studies; and fourth, the Han-Eskin random-effects meta-analysis of the CARDIoGRAMplusC4D dataset has potentially identified an additional variant (rs688359) independently associated with CAD risk.

This page is intentionally left blank.

Chapter 5

5.1 General Discussion

Early applications of meta-analysis in genetic epidemiology centred on summarising evidence across genetic association studies of candidate genes with the aim of resolving conflicting findings (Ioannidis et al., 2001; Weedon et al., 2003). Since then genetic association meta-analyses have evolved to include the assembly of genome-wide association study results, often from independent research groups, to optimize power for the discovery of genetic variants underlying complex traits. As highlighted in chapter 1, differences in the methodology and/or ascertainment criteria of studies included in a meta-analysis can contribute to heterogeneity of the size and/or direction of genetic effects.

The exploration of heterogeneity in genetic association meta-analyses is paramount for two reasons. First, excessive dispersion among genetic effects of studies in a meta-analysis can impinge on the power to replicate genetic associations and discover new genetic signals. Second, moderate heterogeneity of genetic effects can illuminate important biological differences among contributing studies in a meta-analysis, such as variants preferentially associated with sub-phenotypes of the trait of interest (i.e. sub-phenotype variability). A notable example is that of a meta-analysis of endometriosis conducted by Rahmioglu and colleagues, where five of the six genome-wide significant loci (i.e. 6 SNPs) were more strongly associated with moderate-to-severe endometriosis [revised American Fertility Society (rAFS) Stage III/IV] than all endometriosis (Rahmioglu et al., 2014). One of the six SNPs showing stronger association with

moderate-to-severe endometriosis was **rs13394619 in *GREB1*** (growth regulation by oestrogen in breast cancer 1) [OR_{all} = 1.13 (95% CI: 1.07–1.20), $OR_{III/IV\ \text{enriched}}$= 1.15 (95% CI: 1.09–1.20), $OR_{III/IV\ \text{only}}$= 1.18 (95% CI: 1.11–1.24), $P = 2.9 \times 10^{-8}$] (Rahmioglu et al., 2014).

Additionally, moderate heterogeneity can reveal variants differentially associated with the trait of interest as a consequence of factors including: gender (i.e. sexual dimorphism), ancestry, family-history and age-of-onset. For example, in a meta-analysis conducted by Moffatt and colleagues to identify risk variants underpinning asthma, markers on chromosome 17q12 within the *GSDMA* (rs3894194) and *GSDMB* (rs2305480) genes were more strongly associated with childhood-onset asthma [OR_{GSDMA} = 1.26, (95% CI: 1.19–1.33); OR_{GSDMB} = 0.76, (95% CI: 0.72–0.81)] than later-onset asthma [OR_{GSDMA} = 1.02, (95% CI: 0.94–1.11); OR_{GSDMB} = 1.03, (95% CI: 0.94–1.13)] (Moffatt et al., 2010).

Dissection of heterogeneity can also inform the design of future meta-analysis projects wherein characteristics of studies showing enhanced genetic effects (i.e. genetically enriched studies) can be used as a guide when developing ascertainment criteria, in order to optimize power to detect genetic associations.

This thesis explores methods to dissect heterogeneity in GWAS meta-analysis and is motivated by three main ideas namely: 1) the development of methods to identify and explain systematic heterogeneity patterns in genetic association meta-analysis, 2) the development of methods to improve precision in the estimation of average genetic effects and heterogeneity in meta-analyses with smaller numbers of studies, and 3) the identification of small effect heterogeneous loci overlooked by traditional fixed and random-effects meta-analysis methods.

5.1.1 Identifying systematic heterogeneity patterns in genetic association meta-analyses

Discussions on the investigation of heterogeneity in genetic association meta-analyses have largely focussed on identifying and explaining heterogeneity patterns at individual variants (i.e. marker-specific heterogeneity) (Begum et al., 2012). Popular measures for detecting marker-specific heterogeneity, and quantifying its impact and extent include: the Q-statistic, Higgin's inconsistency metric, I^2 and the τ^2-statistic, respectively. Once marker-specific heterogeneity patterns have been detected, single-variant meta-regression is typically performed to identify the study characteristics that contribute to the heterogeneity patterns observed at individual variants.

A new multi-variant statistic to dissect systematic heterogeneity patterns: Assessments of the relative impact of different heterogeneity sources across genetic signals in a meta-analysis however, have been limited by the availability of appropriate methodology. To this end, a novel multi-variant statistic, *M* was presented in chapter 2 that is designed to search for systematic heterogeneity patterns across multiple genetic loci, in order to reveal studies showing consistently stronger or weaker than average genetic effects. Essentially, this new statistic integrates heterogeneity information across multiple loci through the study-wise aggregation of standardised predicted random effects (Magosi et al., 2017). The *M* statistic is based on two key assumptions, that the variants selected for heterogeneity analysis are in linkage equilibrium and are associated with the trait of interest, either at the genome-wide level (e.g. $P < 5 \times 10^{-8}$) or putatively (e.g. $P < 5 \times 10^{-5}$); the latter assumptions optimize the power of the *M* statistic to

identify systematic heterogeneity patterns. *M* analyses can be performed with lead variants at genome-wide significant loci subsequent to the completion of a meta-analysis; alternatively, a known (published) panel of lead variants for the trait of interest can be employed.

The main strength of this approach is that it not only identifies systematic heterogeneity patterns but also reveals their direction of effect; this is important as studies that consistently show weaker than average effects across the panel of variants examined in a meta-analysis can dampen power to detect genetic associations.

Applications of the *M* statistic: Practical applications of the *M* statistic include across-study quality control in larger meta-analyses based on summary data, to identify studies showing systematically weaker than average genetic effects; following which, a meta-regression of *M* statistics can be performed to identify factors within studies showing systematically weaker effects that might be lowering the power of the meta-analysis to detect genetic signals. Weaker than average effects might be induced by faulty genotype data (e.g. flipped alleles, incorrect minor allele frequencies) or variable stringency of quality control checks among participating studies in a meta-analysis (e.g. genotype call rates, Hardy- Weinberg equilibrium cut-offs, imputation quality and adjustment for population structure).

I have demonstrated through simulation studies that the *M* statistic has greater power to identify systematic heterogeneity patterns compared to the Q-statistic, a measure commonly used to detect heterogeneity in genetic association meta-analyses (Figure 2.2). Additionally, I applied the *M* statistic to a recent meta-analysis of CAD risk and was successfully able to identify genome-wide (systematic) heterogeneity patterns that could be explained by differences in ancestry, age

of CAD-onset and family-history of CAD; highlighting earlier-onset CAD, positive CAD family-history, and East Asian ancestry as factors contributing to genetic enrichment in the CARDIoGRAMplusC4D meta-analysis (Figure 2.6).

As shown in the *M* analyses of the CARDIoGRAMplusC4D meta-analysis in chapter 2 (Figure 2.6), a useful application of the *M* statistic is to measure the relative contribution of different biological factors to stronger genetic signals across the loci examined in a meta-analysis; this can be used for instance, to determine the potential gain or loss in power associated with performing a meta-analysis on studies stratified by factors including: sub-phenotype, gender or age-of-onset. For example, where complex traits have a variety of clinical phenotypes (e.g. coronary artery disease) exploratory power analysis could be performed with the *M* statistic to compare the statistical power of performing a meta-analysis with an inclusive phenotype definition (e.g. myocardial infarction (MI), chronic ischemic heart disease (IHD), percutaneous trans-luminal coronary angioplasty (PTCA), coronary artery bypass grafting (CABG) and angina) versus a more stringent one (e.g. inclusive phenotype excluding CABG and angina) (Nelson et al., 2017). The phenotype definition with stronger genetic signals across the selected informative markers could then be used as the primary dataset for meta-analysis.

To summarise, the *M* statistic can be used to explore the impact of different heterogeneity sources across multiple genetic signals in a meta-analysis, to reveal study characteristics contributing to systematically stronger genetic signals, as well as those that might lower the power to validate and discover new genetic signals. The main advantages of the *M* statistic are that it reveals the direction of systematic heterogeneity patterns; and in contrast to the, λ_{meta} heterogeneity statistic (Chen et al., 2016) which assess systematic heterogeneity on a

pairwise study basis (Figure, 2.13), the *M* statistic facilitates exploration of genome-wide heterogeneity patterns simultaneously across all participating studies in a meta-analysis. That said, the *M* statistic is based on standardised predicted random effects, *SPREs* that are mutually independent standard normal random variables, as such its (the *M* statistic) application is currently limited to independently associated variants (i.e. variants in linkage equilibrium with each other). Thus the *M* statistic is not ideal for exploring systematic heterogeneity across variants that are in weak LD.

5.1.2 Improving inference on average genetic effects and heterogeneity in smaller genetic association meta-analyses

Given that estimates of average genetic effects and heterogeneity in smaller meta-analyses (e.g. < 10 studies) are generally imprecise (Turner et al., 2015), I hypothesised that the precision of estimates at individual variants could be improved by employing empirical heterogeneity priors formulated from the distribution of heterogeneity estimates in a GWAS meta-analysis i.e. GWAS-derived empirical heterogeneity priors. Essentially, maximum-likelihood estimation is used to obtain the parameters of a distribution of heterogeneity estimates afforded by method-of-moments or restricted maximum-likelihood algorithms in a frequentist meta-analysis; the resulting parameters are then used as an informed empirical heterogeneity prior in Bayesian genetic association meta-analyses at variants of interest, with the aim of acquiring estimates of average genetic effects and heterogeneity that are more precise than what would be obtained in a stand-alone frequentist meta-analysis.

I have demonstrated through simulation studies and an application to the CARDIoGRAMplusC4D meta-analysis of CAD risk in chapter 3, that GWAS-derived empirical heterogeneity priors modestly improve precision in estimation of average effects and heterogeneity in smaller genetic association meta-analyses (Figures 3.3 and 3.7). The modest improvement in inference possibly reflects the limited information available in smaller meta-analyses to accurately estimate heterogeneity.

A practical application of this approach would be to refine average genetic effect estimates of lead variants at loci suggestively (i.e. $P < 5 \times 10^{-5}$) associated with the trait of

interest, as a preparatory step before identifying secondary genetic signals in a joint analysis. Examples where GWAS-derived empirical heterogeneity priors could be applied include cumulative meta-analysis projects which might at the time comprise a few participating studies (e.g. < 10 studies), or meta-analyses based on less frequently interrogated complex traits that may not yet have been the subject of tens of GWAS (e.g. endometriosis). Additionally, GWAS-derived empirical heterogeneity priors could be employed in niche GWAS meta-analyses concentrating on particular ancestry groups where the number of available GWAS results might be few. From the simulation studies presented in chapter 3, this work has shown that GWAS-derived empirical heterogeneity priors are more useful in improving inference on average genetic effects and heterogeneity in smaller meta-analyses (e.g. < 10 studies) rather than larger ones (e.g. ≥ 20 studies).

5.1.3 Identifying small effect heterogeneous loci overlooked by traditional fixed and random-effects methods

The genetic architecture of complex traits largely comprises small effect loci (per-allele disease odds ratios < 1.2 or trait variance < 0.2%) (Bush and Moore, 2012; Yang et al., 2011), which require sample-sizes on the order of (tens of) thousands of cases and controls for reliable detection; such sample-size requirements can either be met through large collaborative genetic association meta-analysis projects, or more recently through large prospective studies (e.g. UK Biobank). The genetic association study results combined in collaborative meta-analysis projects are usually provided by independent research groups, and tend to be based on study populations with diverse ascertainment criteria and ancestry. Importantly, heterogeneity induced by differences in ancestry and ascertainment criteria can obscure loci conferring smaller effects in meta-analyses. Given that the traditional fixed-effect meta-analysis model disregards heterogeneity and the traditional random-effects model has less power than the fixed-effect method to detect genetic signals; I hypothesised that conventional meta-analytic approaches would be underpowered to detect small effect loci in the presence of moderate-to-high heterogeneity, and would therefore occasionally overlook variants associated with traits of interest.

Han and Eskin have recently proposed a new random-effects meta-analysis method that is designed to detect small effect loci in the presence of heterogeneity (Han and Eskin, 2011), as such it was of interest to determine whether the Han-Eskin random-effects model could identify small effect loci that eluded detection by conventional fixed and random-effects methods. I noted that simulation studies comparing the power of the Han-Eskin random-effects model to

detect small effect heterogeneous loci to that of traditional fixed and random-effects methods (Han and Eskin, 2011; Neupane et al., 2012), had not evaluated the susceptibility of the Han-Eskin random-effects model to inflated genetic signals. Inflated genetic signals are defined as small effect genetic associations largely driven by a few outlier studies showing extreme effects whilst the rest of the participating studies in a meta-analysis show minimal effects (Figure 4.1). This was an important omission as the Han-Eskin random-effects model tests for a non-null average genetic effect or substantial heterogeneity in a meta-analysis. Consequently significant p-values under the Han-Eskin random-effects model may occasionally reflect the extensive heterogeneity in a meta-analysis rather than a significant genetic association. Marker-specific heterogeneity patterns where a few outliers show outsized effects might be induced by epistasis, gene-environment interactions, genotyping error or chance; however, they are more likely to occur as consequence of local ancestry differences (i.e. population structure).

I performed a critical evaluation of the Han-Eskin random-effects model using Monte-Carlo simulation studies, to compare its power to detect small effect genetic associations in the presence of moderate-to-high heterogeneity ($I^2 > 40\%$) with that of traditional fixed and random-effects methods; importantly, the critical evaluation included an assessment of the Han-Eskin random-effects model's vulnerability to inflated genetic signals in relation to traditional fixed and random-effects methods. As demonstrated in the results presented in chapter 4, the power of the Han-Eskin random-effects model to detect small effect heterogeneous loci improved with increasing heterogeneity (Table 4.2). Moreover, the Han-Eskin random-effects model was more powerful than traditional meta-analysis methods at detecting small effect loci when there was moderate-to-high heterogeneity, especially in meta-analyses with smaller numbers of studies (Table 4.2). However, the Han-Eskin random-effects model also showed

greater vulnerability to inflated genetic signals at higher heterogeneity levels ($I^2 > 60\%$), compared to the traditional fixed and random-effects methods (Table 4.3). Therefore, I would recommend that forest plots be generated for small effect heterogeneous loci, particularly ones identified with the Han-Eskin random-effects model in the presence of moderate-to-high heterogeneity. Furthermore, it might be useful to perform an association analysis which adjusts for local ancestry differences at small effect heterogeneous lead variants reporting significant p-values under the Han-Eskin random-effects model, particularly where forest plots have shown marker-specific heterogeneity patterns that are largely due to a few outliers showing extreme effects (Wang et al., 2011); as it has been shown that methods commonly employed to correct for global population structure e.g. genomic control (Devlin and Roeder, 1999) and EIGENSTRAT (Price et al., 2006) are less efficient at adjusting for local ancestry differences (Wang et al., 2011).

A re-appraisal of the CARDIoGRAMplusC4D meta-analysis of CAD risk with the Han-Eskin random-effects model to identify small effect heterogeneous loci overlooked by traditional fixed and random-effects methods, revealed an additional lead variant (rs688359) on chromosome 6 (Figure 4.2). As discussed in chapter 4, rs688359 is an intron variant of insulin-like growth factor 2 receptor gene that lies on the boundary of the *LPA* locus. Notably, the association between the *LPA* locus and elevated CAD risk is very strong (Clarke et al., 2009). There was little correlation observed between rs688359 and the lead variant at the *LPA* locus (Figure 4.3), and between rs688359 and neighbouring CAD risk loci (e.g. SLC22A3 and LPAL2) (Figure 4.4); suggesting that rs688359 is an additional risk variant that is potentially independently associated with CAD. Importantly, the association between rs688359 and elevated CAD risk will need to be replicated in an independent dataset. It will also be interesting to test

whether rs688359 correlates with increased Lp(a) level, as elevated Lp(a) levels have been shown to coincide with increased susceptibility to CAD (The Emerging Risk Factors Collaboration et al., 2009).

5.1.4 Future directions of research

Identification of an additional variant (rs688359) associated with CAD risk: In this work, the Han-Eskin random-effects model (Han and Eskin, 2011) was used to identify an additional CAD risk variant that was overlooked by traditional fixed and random-effects methods in the primary analysis of the CARDIoGRAMplusC4D meta-analysis of CAD risk (Table 4.4, Figure 4.2) (CARDIoGRAMplusC4D Consortium, 2015). Moving forward, it would be important to reproduce the association between rs688359 and increased CAD risk in another dataset, this could be achieved for example, by performing a univariate logistic regression of rs688359 and CAD risk in individuals of White British descent in the UK Biobank dataset (335,000 participants: 19,000 cases of coronary heart disease). Given that increased CAD risk has been shown to coincide with elevated plasma Lp(a) levels (The Emerging Risk Factors Collaboration et al., 2009) it would also be interesting to test whether rs688359 is associated with elevated plasma Lp(a) concentration using univariate linear regression.

Furthermore, the work done in this thesis explored whether the rs688359 variant independently associated with CAD risk or was rather part of the extended *LPA* architecture (via linkage disequilibrium) (Figure 4.3A-C, Figure 4.4). Whilst regional association plots showed little correlation between rs688359 and the lead variant (rs55730499) on the *LPA* locus, suggesting that rs688359 might be independently associated with CAD risk, it would be useful to perform a conditional analysis (Yang et al., 2012) of rs688359 and CAD risk adjusting for the genetic signal at rs55730499. Importantly, a conditional analysis would help to elucidate whether rs688359 is an independent CAD risk locus separate from the genetic signal at rs55730499.

A novel multi-variant statistic to search for genome-wide (systematic) heterogeneity patterns in meta-analyses: The work done in this thesis demonstrated an application of a novel multi-variant statistic, *M* to identify and explain systematic heterogeneity patterns in a recent meta-analysis of CAD risk (CARDIoGRAMplusC4D Consortium, 2015). Given that CAD is a binary trait, it would be interesting to apply the *M* statistic to a meta-analysis based on a quantitative trait. For example, Heid and colleagues showed in a meta-analysis to identify variants associated with waist-hip ratio that variants in certain genes (e.g. *LYPLAL1*, *HOXC13*) have a greater impact on body fat distribution in women than men (Heid et al., 2010). Given that Heid and colleagues explored the impact of gender-related heterogeneity on waist-hip ratio at individual variants, the *M* statistic could potentially be employed to test whether there might systematic effect-size differences across GWAS significant loci in the same meta-analysis attributable to differences in gender, ancestry or any of the other recorded study characteristics.

The work done in thesis has demonstrated that heterogeneity patterns may arise both at individual variants and across multiple variants in a meta-analysis. Whilst it is customary to test for the presence of heterogeneity in genetic association meta-analyses, efforts to explain the observed heterogeneity rather than exclude highly heterogeneous variants have lagged behind. For example, in a recent meta-analysis of obesity adjusted for smoking (Justice et al., 2017), the procedure for selecting lead variants at loci significantly associated with the trait of interest included removing variants with high heterogeneity index levels ($I^2 > 75\%$) prior to grouping the variants by distance ($\pm$500 kb). In cases where the set of significantly associated genetic signals includes variants with higher heterogeneity levels, rather than exclude such variants, it might be more informative to include the affected variants in the lead variant selection then conduct a meta-regression of variant effect-size estimates at each of the heterogeneous lead variants to

explore differences among the characteristics of the studies in the meta-analysis that might explain the observed heterogeneity. Additionally, the lead variants from the meta-analysis could be used as a set of informative markers in an *M* analysis to explore the impact of genome-wide heterogeneity patterns on the trait of interest; in particular, to highlight differential association by ancestry, sub-phenotype, gender or age-of-disease onset.

5.2 Final remarks

Taken together, the motivating ideas explored in this thesis have contributed to advancing knowledge in cardiovascular genetics and the broader area of statistical genetics in four ways. First, through the development of a novel statistic, *M* that identifies systematic heterogeneity patterns in genetic association meta-analyses, and by revealing systematic heterogeneity patterns in the CARDIoGRAMplusC4D meta-analysis of CAD risk underpinned by differences in ancestry, age-of-CAD onset and CAD family-history. The second contribution of this thesis was the development of GWAS-derived empirical heterogeneity priors as an alternative method to improve precision in the estimation of average genetic effects and heterogeneity in smaller genetic association meta-analyses. Third, this work draws attention to the vulnerability of the Han-Eskin random-effects model to inflated genetic signals, in meta-analyses where variants confer smaller effects and there is relatively high heterogeneity. That been said, combined with the inspection of forest plots the Han-Eskin random-effects model can provide important new insights, as demonstrated by our identification of an additional CAD risk variant (rs688359) potentially independently associated with coronary artery disease, that was overlooked by traditional fixed and random-effects methods.

References

Agresti, A. (1996). An introduction to categorical data analysis (JohnWiley & Sons. Inc).

Altshuler, D., Hirschhorn, J.N., Klannemark, M., Lindgren, C.M., Vohl, M.C., Nemesh, J., Lane, C.R., Schaffner, S.F., Bolk, S., Brewer, C.*, et al.* (2000). The common PPAR gamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. Nature Genetics *26*, 76-80.

Anderson, C.D., Nalls, M.A., Biffi, A., Rost, N.S., Greenberg, S.M., Singleton, A.B., Meschia, J.F., and Rosand, J. (2011). The effect of survival bias on case-control genetic association studies of highly lethal diseases. Circulation: Cardiovascular Genetics *4*, 188-196.

Anzures‐Cabrera, J., and Higgins, J. (2010). Graphical displays for meta‐analysis: An overview with suggestions for practice. Research Synthesis Methods *1*, 66-80.

Asimit, J., Day-Williams, A., Zgaga, L., Rudan, I., Boraska, V., and Zeggini, E. (2012). An evaluation of different meta-analysis approaches in the presence of allelic heterogeneity. European Journal of Human Genetics *20*, 709-712.

Astle, W., and Balding, D.J. (2009). Population structure and cryptic relatedness in genetic association studies. Statistical Science *24*, 451-471.

Athanasiou, T., and Darzi, A. (2011). Evidence synthesis in healthcare: a practical handbook for clinicians (Springer Science & Business Media).

Band, G., Le, Q.S., Jostins, L., Pirinen, M., Kivinen, K., Jallow, M., Sisay-Joof, F., Bojang, K., Pinder, M., Sirugo, G.*, et al.* (2013). Imputation-based meta-analysis of severe malaria in three African populations. PLoS Genetics *9*, e1003509.

Barrett, J.C., Hansoul, S., Nicolae, D.L., Cho, J.H., Duerr, R.H., Rioux, J.D., Brant, S.R., Silverberg, M.S., Taylor, K.D., Barmada, M.M.*, et al.* (2008). Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. Nature Genetics *40*, 955-962.

Beaumont, M.A., and Rannala, B. (2004). The Bayesian revolution in genetics. Nature Reviews Genetics *5*, 251-261.

Begum, F., Ghosh, D., Tseng, G.C., and Feingold, E. (2012). Comprehensive literature review and statistical considerations for GWAS meta-analysis. Nucleic Acids Research *40*, 3777-3784.

Bergman, D., Halje, M., Nordin, M., and Engström, W. (2013). Insulin-like growth factor 2 in development and disease: a mini-review. Gerontology *59*, 240-249.

Bigdeli, T.B., Lee, D., Webb, B.T., Riley, B.P., Vladimirov, V.I., Fanous, A.H., Kendler, K.S., and Bacanu, S.-A. (2016). A simple yet accurate correction for winner's curse can predict signals discovered in much larger genome scans. Bioinformatics *32*, 2598-2603.

Birge, R.T. (1932). The calculation of errors by the method of least squares. Physical Review *40*, 207.

Borenstein, M., Hedges, L.V., Higgins, J., and Rothstein, H.R. (2009). Introduction to Meta-Analysis (Wiley Online Library).

Borenstein, M., Hedges, L.V., Higgins, J., and Rothstein, H.R. (2010). A basic introduction to fixed ‐ effect and random ‐ effects models for meta ‐ analysis. Research Synthesis Methods *1*, 97-111.

Borenstein, M., Higgins, J.P.T., Hedges, L.V., and Rothstein, H.R. (2017). Basics of meta-analysis: I2 is not an absolute measure of heterogeneity. Research Synthesis Methods *8*, 5-18.

Bush, W.S., and Moore, J.H. (2012). Chapter 11: Genome-wide association studies. PLoS Computational Biology *8*, e1002822.

Cantor, R.M., Lange, K., and Sinsheimer, J.S. (2010). Prioritizing GWAS results: A review of statistical methods and recommendations for their application. American Journal of Human Genetics *86*, 6-22.

CARDIoGRAMplusC4D Consortium (2015). A comprehensive 1000 Genomes-based genome-wide association meta-analysis of coronary artery disease. Nature Genetics *47*, 1121-1130.

CARDIoGRAMplusC4D Consortium, Deloukas, P., Kanoni, S., Willenborg, C., Farrall, M., Assimes, T.L., Thompson, J.R., Ingelsson, E., Saleheen, D., Erdmann, J.*, et al.* (2013). Large-scale association analysis identifies new risk loci for coronary artery disease. Nature Genetics *45*, 25-33.

Chatterjee, N., Shi, J., and García-Closas, M. (2016). Developing and evaluating polygenic risk prediction models for stratified disease prevention. Nature Reviews Genetics *17*, 392-406.

Chen, G.B., Lee, S.H., Robinson, M.R., Trzaskowski, M., Zhu, Z.X., Winkler, T.W., Day, F.R., Croteau-Chonka, D.C., Wood, A.R., Locke, A.E.*, et al.* (2016). Across-cohort QC analyses of GWAS summary statistics from complex traits. European Journal of Human Genetics *25*, 137–146.

Chen, Y., Li, L., Zhang, Q., Clarke, R., Chen, J., Guo, Y., Bian, Z., Pan, X., Peto, R., Tao, R.*, et al.* (2014). Use of drug treatment for secondary prevention of cardiovascular disease in urban and rural communities of China: China Kadoorie Biobank Study of 0.5 million people. International journal of cardiology *172*, 88-95.

Clarke, R., Peden, J.F., Hopewell, J.C., Kyriakou, T., Goel, A., Heath, S.C., Parish, S., Barlera, S., Franzosi, M.G., Rust, S.*, et al.* (2009). Genetic variants associated with Lp(a) lipoprotein level and coronary disease. New England Journal of Medicine *361*, 2518-2528.

Cochran, W.G. (1937). Problems arising in the analysis of a series of similar experiments. Supplement to the Journal of the Royal Statistical Society *4*, 102-118.

Cochran, W.G. (1954). The combination of estimates from different experiments. Biometrics *10*, 101-129.

Colhoun, H.M., McKeigue, P.M., and Smith, G.D. (2003). Problems of reporting genetic associations with complex outcomes. Lancet *361*, 865-872.

Cox, N. (1998). WBULL: Stata module to fit Weibull distribution by maximum likelihood.

Cox, N., and Jenkins, S. (2011a). GAMMAFIT: Stata module to fit a two-parameter gamma distribution.

Cox, N., and Jenkins, S. (2011b). INVGAMMAFIT: Stata module to fit a two-parameter inverse gamma distribution.

Davison, A.C., and Hinkley, D.V. (1997). Bootstrap methods and their application, Vol 1 (Cambridge university press).

Deeb, S.S., Fajas, L., Nemoto, M., Pihlajamäki, J., Mykkänen, L., Kuusisto, J., Laakso, M., Fujimoto, W., and Auwerx, J. (1998). A Pro12Ala substitution in PPAR γ 2 associated with decreased receptor activity, lower body mass index and improved insulin sensitivity. Nature Genetics *20*, 284-287.

DerSimonian, R., and Laird, N. (1986). Meta-analysis in clinical trials. Controlled Clinical Trials *7*, 177-188.

Devlin, B., and Roeder, K. (1999). Genomic control for association studies. Biometrics *55*, 997-1004.

DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium, Asian Genetic Epidemiology Network Type 2 Diabetes (AGEN-T2D) Consortium, South Asian Type 2 Diabetes (SAT2D) Consortium, Mexican American Type 2 Diabetes (MAT2D) Consortium, Type 2 Diabetes Genetic Exploration by Nex-generation sequencing in multi-Ethnic Samples (T2D-GENES) Consortium, Mahajan, A., Go, M.J., Zhang, W., Below, J.E., Gaulton, K.J.*, et al.* (2014). Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. Nature Genetics *46*, 234-244.

Dickersin, K., and Berlin, J.A. (1992). Meta-analysis: state-of-the-science. Epidemiologic Reviews *14*, 154-176.

Evangelou, E., and Ioannidis, J.P. (2013). Meta-analysis methods for genome-wide association studies and beyond. Nature Reviews Genetics *14*, 379-389.

Evangelou, E., Trikalinos, T.A., Salanti, G., and Ioannidis, J.P. (2006). Family-based versus unrelated case-control designs for genetic associations. PLoS Genetics *2*, e123.

Fisher, S.A., and Lewis, C.M. (2008). Power of genetic association studies in the presence of linkage disequilibrium and allelic heterogeneity. Human Heredity *66*, 210-222.

Frayling, T.M., Timpson, N.J., Weedon, M.N., Zeggini, E., Freathy, R.M., Lindgren, C.M., Perry, J.R., Elliott, K.S., Lango, H., Rayner, N.W.*, et al.* (2007). A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. Science *316*, 889-894.

Gabriel, S.B., Schaffner, S.F., Nguyen, H., Moore, J.M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M.*, et al.* (2002). The structure of haplotype blocks in the human genome. Science *296*, 2225-2229.

Glass, G.V. (1976). Primary, secondary, and meta-analysis of research. Educational Researcher *5*, 3-8.

Gögele, M., Minelli, C., Thakkinstian, A., Yurkiewich, A., Pattaro, C., Pramstaller, P.P., Little, J., Attia, J., and Thompson, J.R. (2012). Methods for meta-analyses of genome-wide association studies: critical assessment of empirical evidence. American Journal of Epidemiology *175*, 739-749.

Hall, M.A., Verma, S.S., Wallace, J., Lucas, A., Berg, R.L., Connolly, J., Crawford, D.C., Crosslin, D.R., de Andrade, M., Doheny, K.F.*, et al.* (2015). Biology‐Driven Gene‐Gene Interaction Analysis of Age‐Related Cataract in the eMERGE Network. Genetic Epidemiology *39*, 376-384.

Han, B., and Eskin, E. (2011). Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. American Journal of Human Genetics *88*, 586-598.

Harbord, R.M., and Higgins, J. (2008). Meta-regression in Stata. Stata Journal *8*, 493-519.

Hardy, R.J., and Thompson, S.G. (1998). Detecting and describing heterogeneity in meta‐analysis. Statistics in Medicine *17*, 841-856.

Harville, D.A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. Journal of the American Statistical Association *72*, 320-338.

Hedges, L.V. (2016). Comment on 'Misunderstandings about Q and "Cochran's Q Test" in meta analysis'. Statistics in Medicine *35*, 496-497.

Heid, I.M., Huth, C., Loos, R.J., Kronenberg, F., Adamkova, V., Anand, S.S., Ardlie, K., Biebermann, H., Bjerregaard, P., Boeing, H.*, et al.* (2009). Meta-analysis of the INSIG2 association with obesity including 74,345 individuals: does heterogeneity of estimates relate to study design? PLoS Genetics *5*, e1000694.

Heid, I.M., Jackson, A.U., Randall, J.C., Winkler, T.W., Qi, L., Steinthorsdottir, V., Thorleifsson, G., Zillikens, M.C., Speliotes, E.K., Magi, R.*, et al.* (2010). Meta-analysis identifies 13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution. Nature Genetics *42*, 949-960.

Higgins, J., and Thompson, S.G. (2002). Quantifying heterogeneity in a meta‐analysis. Statistics in Medicine *21*, 1539-1558.

Higgins, J., Thompson, S.G., and Spiegelhalter, D.J. (2009). A re‐evaluation of random‐effects meta‐analysis. Journal of the Royal Statistical Society: Series A (Statistics in Society) *172*, 137-159.

Higgins, J., and Whitehead, A. (1996). Borrowing strength from external trials in a meta‐analysis. Statistics in Medicine *15*, 2733-2749.

Higgins, J.P., Thompson, S.G., Deeks, J.J., and Altman, D.G. (2003). Measuring inconsistency in meta-analyses. British Medical Journal *327*, 557-560.

Hirschhorn, J.N., Lohmueller, K., Byrne, E., and Hirschhorn, K. (2002). A comprehensive review of genetic association studies. Genetics in Medicine *4*, 45-61.

Hoaglin, D.C. (2016). Misunderstandings about Q and 'Cochran's Q test' in meta‐analysis. Statistics in Medicine *35*, 485-495.

Hosmer, D.W., and Lemeshow, S. (2005). Model-Building Strategies and Methods for Logistic Regression. In Applied Logistic Regression (John Wiley & Sons, Inc.), pp. 91-142.

Huedo-Medina, T.B., Sánchez-Meca, J., Marín-Martínez, F., and Botella, J. (2006). Assessing heterogeneity in meta-analysis: Q statistic or I^2 index? Psychological Methods *11*, 193-206.

Hunter, D.J. (2005). Gene-environment interactions in human diseases. Nature Reviews Genetics *6*, 287-298.

Ikram, M.A., Seshadri, S., Bis, J.C., Fornage, M., DeStefano, A.L., Aulchenko, Y.S., Debette, S., Lumley, T., Folsom, A.R., van den Herik, E.G.*, et al.* (2009). Genomewide association studies of stroke. New England Journal of Medicine *360*, 1718-1728.

International HapMap Consortium (2005). A haplotype map of the human genome. Nature *437*, 1299-1320.

International HapMap Consortium, Frazer, K.A., Ballinger, D.G., Cox, D.R., Hinds, D.A., Stuve, L.L., Gibbs, R.A., Belmont, J.W., Boudreau, A., Hardenbol, P.*, et al.* (2007). A second generation human haplotype map of over 3.1 million SNPs. Nature *449*, 851-861.

Ioannidis, J.P., Ntzani, E.E., Trikalinos, T.A., and Contopoulos-Ioannidis, D.G. (2001). Replication validity of genetic association studies. Nature Genetics *29*, 306-309.

Ioannidis, J.P., Patsopoulos, N.A., and Evangelou, E. (2007a). Heterogeneity in meta-analyses of genome-wide association investigations. PloS One *2*, e841.

Ioannidis, J.P., Patsopoulos, N.A., and Evangelou, E. (2007b). Uncertainty in heterogeneity estimates in meta-analyses. British Medical Journal *335*, 914-916.

Jenkins, S. (2013). LOGNFIT: Stata module to fit lognormal distribution by maximum likelihood.

Jorde, L.B., and Wooding, S.P. (2004). Genetic variation, classification and 'race'. Nature Genetics *36*, S28-S33.

Justice, A.E., Winkler, T.W., Feitosa, M.F., Graff, M., Fisher, V.A., Young, K., Barata, L., Deng, X., Czajkowski, J., Hadley, D.*, et al.* (2017). Genome-wide meta-analysis of 241,258 adults accounting for smoking behaviour identifies novel loci for obesity traits. Nature Communications *8*, 14977.

Kruschke, J. (2014). Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan (Academic Press).

Kulinskaya, E., and Dollinger, M.B. (2015). An accurate test for homogeneity of odds ratios based on Cochran's Q-statistic. BMC Medical Research Methodology *15*, 49.

Kulinskaya, E., and Dollinger, M.B. (2016). Commentary on 'Misunderstandings about Q and "Cochran's Q test" in meta analysis'. Statistics in Medicine *35*, 501-502.

Kulinskaya, E., Dollinger, M.B., and Bjørkestøl, K. (2011). On the moments of Cochran's Q statistic under the null hypothesis, with application to the meta‐analysis of risk difference. Research Synthesis Methods *2*, 254-270.

Lambert, P.C., Sutton, A.J., Burton, P.R., Abrams, K.R., and Jones, D.R. (2005). How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. Statistics in Medicine *24*, 2401-2428.

Langan, D., Higgins, J., and Simmonds, M. (2017). Comparative performance of heterogeneity variance estimators in meta‐analysis: a review of simulation studies. Research Synthesis Methods *8*, 181-198.

Lebrec, J.J., Stijnen, T., and van Houwelingen, H.C. (2010). Dealing with heterogeneity between cohorts in genomewide SNP association studies. Statistical Applications in Genetics and Molecular Biology *9*, Article 8.

Lee, C., Eskin, E., and Han, B. (2017). Increasing the power of meta-analysis of genome-wide association studies to detect heterogeneous effects. Bioinformatics *33*, i379-i388.

Lewis, S., and Clarke, M. (2001). Forest plots: trying to see the wood and the trees. British Medical Journal *322*, 1479-1480.

Littlejohns, T.J., Sudlow, C., Allen, N.E., and Collins, R. (2017). UK Biobank: opportunities for cardiovascular research. European Heart Journal, ehx254.

Liu, B.H. (1997). Statistical genomics: linkage, mapping, and QTL analysis (CRC press).

Lunn, D.J., Thomas, A., Best, N., and Spiegelhalter, D. (2000). WinBUGS-a Bayesian modelling framework: concepts, structure, and extensibility. Statistics and Computing *10*, 325-337.

Magosi, L.E., Goel, A., Hopewell, J.C., Farrall, M., and CARDIoGRAMplusC4D Consortium (2017). Identifying systematic heterogeneity patterns in genetic association meta-analysis studies. PLoS Genetics *13*, e1006755.

Malik, R., Traylor, M., Pulit, S.L., Bevan, S., Hopewell, J.C., Holliday, E.G., Zhao, W., Abrantes, P., Amouyel, P., Attia, J.R.*, et al.* (2016). Low-frequency and common genetic variation in ischemic stroke The METASTROKE collaboration. Neurology *86*, 1217-1226.

Mancini, F.P., Vaccaro, O., Sabatino, L., Tufano, A., Rivellese, A.A., Riccardi, G., and Colantuoni, V. (1999). Pro12Ala substitution in the peroxisome proliferator-activated receptor-gamma2 is not associated with type 2 diabetes. Diabetes *48*, 1466-1468.

Marchini, J., Cardon, L.R., Phillips, M.S., and Donnelly, P. (2004). The effects of human population structure on large genetic association studies. Nature Genetics *36*, 512-517.

McCarthy, M.I. (2008). Casting a wider net for diabetes susceptibility genes. Nature Genetics *40*, 1039-1040.

Merriman, M. (1909). A text book on the Method of Least Squares (Wiley).

Moffatt, M.F., Gut, I.G., Demenais, F., Strachan, D.P., Bouzigon, E., Heath, S., von Mutius, E., Farrall, M., Lathrop, M., Cookson, W.*, et al.* (2010). A large-scale, consortium-based genomewide association study of asthma. New England Journal of Medicine *363*, 1211-1221.

Moonesinghe, R., Khoury, M.J., Liu, T., and Ioannidis, J.P. (2008). Required sample size and nonreplicability thresholds for heterogeneous genetic associations. Proceedings of the National Academy of Sciences *105*, 617-622.

Mori, H., Ikegami, H., Kawaguchi, Y., Seino, S., Yokoi, N., Takeda, J., Inoue, I., Seino, Y., Yasuda, K., Hanafusa, T.*, et al.* (2001). The Pro12→ Ala substitution in PPAR-γ is associated with resistance to development of diabetes in the general population. Diabetes *50*, 891-894.

Morris, A.P. (2011). Transethnic meta‐analysis of genomewide association studies. Genetic Epidemiology *35*, 809-822.

Moskvina, V., Holmans, P., Schmidt, K., and Craddock, N. (2005). Design of case‐controls studies with unscreened controls. Annals of Human Genetics *69*, 566-576.

Nei, M. (1978). Estimation of average heterozygosity and genetic distance from a small number of individuals. Genetics *89*, 583-590.

Nelson, C.P., Goel, A., Butterworth, A.S., Kanoni, S., Webb, T.R., Marouli, E., Zeng, L., Ntalla, I., Lai, F.Y., Hopewell, J.C.*, et al.* (2017). Association analyses based on false discovery rate implicate new loci for coronary artery disease. Nature Genetics *49*, 1385-1391.

Neupane, B., Loeb, M., Anand, S.S., and Beyene, J. (2012). Meta-analysis of genetic association studies under heterogeneity. European Journal of Human Genetics *20*, 1174-1181.

North, B., Curtis, D., and Sham, P. (2003). A note on the calculation of empirical P values from Monte Carlo procedures. American Journal of Human Genetics *72*, 498-499.

Ntzani, E.E., Liberopoulos, G., Manolio, T.A., and Ioannidis, J.P. (2012). Consistency of genome-wide associations across major ancestral groups. Human Genetics *131*, 1057-1071.

O'rourke, K. (2007). An historical perspective on meta-analysis: dealing quantitatively with varying study results. Journal of the Royal Society of Medicine *100*, 579-582.

Painter, J.N., Anderson, C.A., Nyholt, D.R., Macgregor, S., Lin, J., Lee, S.H., Lambert, A., Zhao, Z.Z., Roseman, F., Guo, Q.*, et al.* (2011). Genome-wide association study identifies a locus at 7p15. 2 associated with endometriosis. Nature Genetics *43*, 51-54.

Park, J.-H., Gail, M.H., Weinberg, C.R., Carroll, R.J., Chung, C.C., Wang, Z., Chanock, S.J., Fraumeni, J.F., and Chatterjee, N. (2011). Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants. Proceedings of the National Academy of Sciences *108*, 18026-18031.

Plummer, M. (2016). rjags: Bayesian Graphical Models using MCMC. R package version 4-6.

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

Pullenayegum, E.M. (2011). An informed reference prior for between ‐ study heterogeneity in meta ‐ analyses of binary outcomes. Statistics in Medicine *30*, 3082-3094.

Rahmioglu, N., Nyholt, D.R., Morris, A.P., Missmer, S.A., Montgomery, G.W., and Zondervan, K.T. (2014). Genetic variants underlying risk of endometriosis: insights from meta-analysis of eight genome-wide association and replication datasets. Human Reproduction Update *20*, 702-716.

Reynolds, J., Weir, B.S., and Cockerham, C.C. (1983). Estimation of the coancestry coefficient: basis for a short-term genetic distance. Genetics *105*, 767-779.

Rhodes, K.M., Turner, R.M., and Higgins, J.P. (2015). Predictive distributions were developed for the extent of heterogeneity in meta-analyses of continuous outcome data. Journal of Clinical Epidemiology *68*, 52-60.

Rhodes, K.M., Turner, R.M., White, I.R., Jackson, D., Spiegelhalter, D.J., and Higgins, J. (2016). Implementing informative priors for heterogeneity in meta ‐ analysis using meta ‐ regression and pseudo data. Statistics in Medicine *35*, 5495-5511.

Risch, N. (2001). Implications of multilocus inheritance for gene–disease association studies. Theoretical Population Biology *60*, 215-220.

Rücker, G., Schwarzer, G., Carpenter, J.R., and Schumacher, M. (2008). Undue reliance on I 2 in assessing heterogeneity may mislead. BMC Medical Research Methodology *8*, 79.

Salanti, G., Sanderson, S., and Higgins, J.P. (2005). Obstacles and opportunities in meta-analysis of genetic association studies. Genetics in Medicine *7*, 13-20.

Sapkota, Y., Steinthorsdottir, V., Morris, A.P., Fassbender, A., Rahmioglu, N., De Vivo, I., Buring, J.E., Zhang, F., Edwards, T.L., Jones, S.*, et al.* (2017). Meta-analysis identifies five novel loci associated with endometriosis highlighting key genes involved in hormone metabolism. Nature Communications *8*, 15539.

Scott, L.J., Mohlke, K.L., Bonnycastle, L.L., Willer, C.J., Li, Y., Duren, W.L., Erdos, M.R., Stringham, H.M., Chines, P.S., Jackson, A.U.*, et al.* (2007). A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. Science *316*, 1341-1345.

Sham, P.C., and Purcell, S.M. (2014). Statistical power and significance testing in large-scale genetic studies. Nature Reviews Genetics *15*, 335-346.

Shannon, H. (2016). A statistical note on Karl Pearson's 1904 meta-analysis. Journal of the Royal Society of Medicine *109*, 310-311.

Shifman, S., Kuypers, J., Kokoris, M., Yakir, B., and Darvasi, A. (2003). Linkage disequilibrium patterns of the human genome across populations. Human Molecular Genetics *12*, 771-776.

Shoemaker, J.S., Painter, I.S., and Weir, B.S. (1999). Bayesian statistics in genetics: a guide for the uninitiated. Trends in Genetics *15*, 354-358.

Simpson, R., and Pearson, K. (1904). Report on certain enteric fever inoculation statistics. British Medical Journal *2*, 1243-1246.

Sinharay, S. (2003). Assessing convergence of the Markov chain Monte Carlo algorithms: A review. ETS Research Report Series *2003*, i-52.

Smith, T.C., Spiegelhalter, D.J., and Thomas, A. (1995). Bayesian approaches to random‐effects meta‐analysis: a comparative study. Statistics in Medicine *14*, 2685-2699.

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

The Emerging Risk Factors Collaboration, Erqou, S., Kaptoge, S., Perry, P.L., Di Angelantonio, E., Thompson, A., White, I.R., Marcovina, S.M., Collins, R., Thompson, S.G.*, et al.* (2009). Lipoprotein (a) concentration and the risk of coronary heart disease, stroke, and nonvascular mortality. Journal of the American Medical Association *302*, 412-423.

Thompson, J.R., Attia, J., and Minelli, C. (2011). The meta-analysis of genome-wide association studies. Briefings in Bioinformatics *12*, 259-269.

Thompson, S.G., and Higgins, J. (2002). How should meta‐regression analyses be undertaken and interpreted? Statistics in Medicine *21*, 1559-1573.

Thompson, S.G., and Sharp, S.J. (1999). Explaining heterogeneity in meta‐analysis: a comparison of methods. Statistics in Medicine *18*, 2693-2708.

Thorlund, K., Imberger, G., Johnston, B.C., Walsh, M., Awad, T., Thabane, L., Gluud, C., Devereaux, P., and Wetterslev, J. (2012). Evolution of heterogeneity (I2) estimates and their 95% confidence intervals in large meta-analyses. PloS One *7*, e39471.

Traylor, M., Farrall, M., Holliday, E.G., Sudlow, C., Hopewell, J.C., Cheng, Y.C., Fornage, M., Ikram, M.A., Malik, R., Bevan, S.*, et al.* (2012). Genetic risk factors for ischaemic stroke and its subtypes (the METASTROKE collaboration): a meta-analysis of genome-wide association studies. Lancet Neurology *11*, 951-962.

Trikalinos, T.A., Hoaglin, D.C., and Schmid, C.H. (2013). Empirical and Simulation-Based Comparison of Univariate and Multivariate Meta-Analysis for Binary Outcomes (Agency for Healthcare Research and Quality (US), Rockville (MD)).

Turner, R.M., Davey, J., Clarke, M.J., Thompson, S.G., and Higgins, J.P. (2012). Predicting the extent of heterogeneity in meta-analysis, using empirical data from the Cochrane Database of Systematic Reviews. International Journal of Epidemiology *41*, 818-827.

Turner, R.M., Jackson, D., Wei, Y., Thompson, S.G., and Higgins, J. (2015). Predictive distributions for between‐study heterogeneity and simple methods for their application in Bayesian meta‐analysis. Statistics in Medicine *34*, 984-998.

Unoki, H., Takahashi, A., Kawaguchi, T., Hara, K., Horikoshi, M., Andersen, G., Ng, D.P., Holmkvist, J., Borch-Johnsen, K., Jorgensen, T.*, et al.* (2008). SNPs in KCNQ1 are associated with susceptibility to type 2 diabetes in East Asian and European populations. Nature Genetics *40*, 1098-1102.

Veroniki, A.A., Jackson, D., Viechtbauer, W., Bender, R., Bowden, J., Knapp, G., Kuss, O., Higgins, J.P.T., Langan, D., and Salanti, G. (2016). Methods to estimate the between-study variance and its uncertainty in meta-analysis. Research Synthesis Methods *7*, 55-79.

Viechtbauer, W. (2005). Bias and Efficiency of Meta-Analytic Variance Estimators in the Random-Effects Model. Journal of Educational and Behavioral Statistics *30*, 261-293.

Viechtbauer, W., and Cheung, M.W.L. (2010). Outlier and influence diagnostics for meta‐analysis. Research synthesis methods *1*, 112-125.

von Hippel, P.T. (2015). The heterogeneity statistic I 2 can be biased in small meta-analyses. BMC Medical Research Methodology *15*, 35.

Wang, X., Zhu, X., Qin, H., Cooper, R.S., Ewens, W.J., Li, C., and Li, M. (2011). Adjustment for local ancestry in genetic association analysis of admixed populations. Bioinformatics *27*, 670-677.

Watkins, H., and Farrall, M. (2006). Genetic susceptibility to coronary artery disease: from promise to progress. Nature Reviews Genetics *7*, 163-173.

Weedon, M.N., Schwarz, P.E., Horikawa, Y., Iwasaki, N., Illig, T., Holle, R., Rathmann, W., Selisko, T., Schulze, J., Owen, K.R.*, et al.* (2003). Meta-analysis and a large association study confirm a role for calpain-10 variation in type 2 diabetes susceptibility. American Journal of Human Genetics *73*, 1208-1212.

Wellcome Trust Case Control Consortium, Craddock, N., Hurles, M.E., Cardin, N., Pearson, R.D., Plagnol, V., Robson, S., Vukcevic, D., Barnes, C., Conrad, D.F.*, et al.* (2010). Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. Nature *464*, 713-720.

Wellcome Trust Case-Control Consortium, and International Stroke Genetics Consortium (2010). Failure to validate association between 12p13 variants and ischemic stroke. New England Journal of Medicine *362*, 1547-1550.

Wilks, S.S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. The Annals of Mathematical Statistics *9*, 60-62.

Winkler, T.W., Day, F.R., Croteau-Chonka, D.C., Wood, A.R., Locke, A.E., Magi, R., Ferreira, T., Fall, T., Graff, M., Justice, A.E.*, et al.* (2014). Quality control and conduct of genome-wide association meta-analyses. Nature Protocols *9*, 1192-1212.

Wood, A.R., Esko, T., Yang, J., Vedantam, S., Pers, T.H., Gustafsson, S., Chu, A.Y., Estrada, K., Luan, J., Kutalik, Z.*, et al.* (2014). Defining the role of common variation in the genomic and biological architecture of adult human height. Nature Genetics *46*, 1173-1186.

Wright, A.F., and Hastie, N.D. (2001). Complex genetic diseases: controversy over the Croesus code. Genome Biology *2*, COMMENT2007.

Wright, S. (1968). Evolution and the Genetics of Populations, Vol 1 (Chicago: University of Chicago Press).

Yang, G., Wang, Y., Zeng, Y., Gao, G.F., Liang, X., Zhou, M., Wan, X., Yu, S., Jiang, Y., Naghavi, M.*, et al.* (2013). Rapid health transition in China, 1990-2010: findings from the Global Burden of Disease Study 2010. Lancet *381*, 1987-2015.

Yang, J., Ferreira, T., Morris, A.P., Medland, S.E., Genetic Investigation of ANthropometric Traits (GIANT) Consortium, DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium, Madden, P.A.F., Heath, A.C., Martin, N.G., Montgomery, G.W.*, et al.* (2012). Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nature Genetics *44*, 369-375, S361-363.

Yang, J., Manolio, T.A., Pasquale, L.R., Boerwinkle, E., Caporaso, N., Cunningham, J.M., de Andrade, M., Feenstra, B., Feingold, E., Hayes, M.G.*, et al.* (2011). Genome partitioning of genetic variation for complex traits using common SNPs. Nature Genetics *43*, 519-525.

Young, A.I., Wauthier, F., and Donnelly, P. (2016). Multiple novel gene-by-environment interactions modify the effect of FTO variants on body mass index. Nature Communications *7*, 12724.

Zeggini, E., and Ioannidis, J.P. (2009). Meta-analysis in genome-wide association studies. Pharmacogenomics *10*, 191-201.

Zeggini, E., Scott, L.J., Saxena, R., Voight, B.F., Marchini, J.L., Hu, T., de Bakker, P.I., Abecasis, G.R., Almgren, P., Andersen, G.*, et al.* (2008). Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. Nature Genetics *40*, 638-645.

Zondervan, K.T., and Cardon, L.R. (2004). The complex interplay among factors that influence allelic association. Nature Reviews Genetics *5*, 89-100.

Contributions of Collaborators

Data curation of the CARDIoGRAMplusC4D dataset was performed in collaboration with Anuj Goel (Wellcome Centre for Human Genetics, England).

Appendix

Electronic versions of the source code for the getmstatistic Stata command and R package can be accessed at the following URL: https://github.com/magosil86/getmstatistic

note: red arrows (↪) indicate line wrapping

getmstatistic: Stata module

getmstatistic_stata_0.1.0/getmstatistic.ado

```
capture program drop getmstatistic
*! getmstatistic v0.1.0 LEMagosi 25may2016
program define getmstatistic, rclass
    version 10.0
    syntax varlist(min=4 max=4) [if] [in] [, noPRInt LATexout NOGRaph SAVEdataset MM REML EB]
    *marksample touse
    tokenize `varlist'

    preserve

    /* Identifies samples for computation based on [if] [in] */

    if "`if'`in'" != "" {

        keep `if' `in'
    }


     /* Check data types for input variables */


        capture confirm numeric variable `1' `2'
        if _rc!=0 {
        display "Error: beta and its corresponding standard error should be of type double"
        exit _rc
        }

        capture confirm string variable `3' `4'
        if _rc!=0 {
        display "Error: variant and study names should be of type string"
        exit _rc
        }


    /* Check whether a specific method has been selected for estimation of tau2 */
    local wc : word count "`mm' `reml' `eb'"
    if `wc' > 1 {

        di as error "Either mm or reml can be specified, but not both"
        exit 198

        }

    if ("`mm'" == "" & "`reml'" == "" & "`eb'" == "") { /* mm is the default metareg_method */

        local metareg_method "mm"
        di as text "tau2 estimated by default method: `metareg_method'"


        }


    if ("`mm'" != "") { /* applying method-of_moments option as metareg_method */

        local metareg_method "mm"
        di as text "Using option: method-of-moments to estimate tau2"

    }

    else if ("`reml'" != "") { /* applying reml option as metareg_method */
```

```
    local metareg_method "reml"
    di as text "Using option: residual maximum likelihood to estimate tau2"

}


else if ("`eb'" != "") { /* applying eb option as metareg_method */

    local metareg_method "eb"
    di as text "Using option: empirical bayes to estimate tau2"

}


/* Main code: Compute the multi-variant heterogeneity statistic (M) */

quietly {


            /*
             * Required variable  Type       Comment
             * beta               double     effect-size
             * se                 double     lambda corrected std error for beta
             * variants           string     rsID
             * studies            string     names of participating studies
             *
             * Usage:
             * getmstatistic beta se variants studies
             * getmstatistic beta se variants studies [, noPRInt LATEXOUT NOGRAPH]
             */

            /*
            Expectations:
            1. M identifies outlier studies showing systematically stronger or weaker effects than
            average.
            2. M delineates the direction of effects of the outlier studies.
            */

            set more off

            *assign user arguments to macros
            local beta `1'
            local se_lambda_corr `2'
            local variants `3'
            local studies `4'


            *assign study numbers
            egen study = group(`studies')

            *calculate no. of studies and assign the value to the macro nstudies
            quietly: duplicates report study
            local nstudies = r(unique_value)


            *assign snp numbers
            egen snp = group(`variants')

            *calculate no. of snps and assign the value to the macro nsnps
            quietly: duplicates report snp
            local nsnps = r(unique_value)


            di "Summary: Heterogeneity analysis is based on `nsnps' SNPs and `nstudies' studies"
```

```
**************** end of part 1: assign snp and study numbers      ***************


*align study effects i.e. betas
foreach snp of numlist 1(1)`nsnps' {
quietly: metareg `beta' if snp == `snp',wsse(`se_lambda_corr') `metareg_method'
matrix B = e(b)
local ES B[1,1]
if `ES'<0 replace `beta' = -`beta' if snp == `snp'
}

*run metareg to extract standardized predicted random effects(SPREs) i.e. usta
gen xb = .
gen usta = .
gen xbu = .
gen stdxbu = .
gen hat = .
gen tau2 = .
gen I2 = .
gen Q = .

foreach snp of numlist 1(1)`nsnps' {
quietly: metareg `beta' if snp == `snp',wsse(`se_lambda_corr') `metareg_method'
quietly: replace tau2 = e(tau2) if e(sample)
quietly: replace I2 = e(I2) if e(sample)
quietly: replace Q = e(Q) if e(sample)
quietly: predict b if e(sample),xb
quietly: predict x if e(sample),ustandard
quietly: predict y if e(sample),xbu
quietly: predict z if e(sample),stdxbu
quietly: predict h if e(sample),hat
quietly:replace xb = b if snp == `snp'
quietly:replace usta = x if snp == `snp'
quietly:replace xbu = y if snp == `snp'
quietly:replace stdxbu = z if snp == `snp'
quietly:replace hat = h if snp == `snp'
drop x y z h b

}

***** end of part 2: run metareg to extract tau2, I2, Q, usta, xb, xbu, stdxbu and hat ***

*Computing M statistic (usta_mean)

*Mstat: aggregate the ustas by taking the mean and calculate CI
gen usta_mean = .
gen usta_sd = .
gen ustamean_se = .
gen lowerbound = .
gen upperbound = .

label variable usta_mean "Mstatistic"
label variable lowerbound "usta_mean 95% CI lowerbound"
label variable upperbound "usta_mean 95% CI upperbound"


foreach study of numlist 1(1) `nstudies' {
quietly: summ usta if study == `study'
local umean = r(mean)
local uobs = r(N)
if (`uobs' == 1) {
    local usd = 0
}
else {
    local usd = r(sd)
    }
```

```
cii `uobs' `umean' `usd', level(95)
di "completed study: `study'"


*Top up allocates missing snps the average mean and sd
*local topup = (`nsnps'-`uobs')*(1/`uobs')*0
local topup = (`nsnps'-`uobs')*((1/`uobs')*0)


local ulb = r(lb)
local uub = r(ub)
local ustderr = r(se)
quietly:replace usta_mean = `umean' + `topup' if study == `study'
quietly:replace usta_sd = `usd' if study == `study'
quietly:replace ustamean_se = `ustderr' if study == `study'
quietly:replace lowerbound = `ulb' if study == `study'
quietly:replace upperbound = `uub' if study == `study'
di "Mstatistic for study `study' is: `umean'"
}


*The expected mean under the Ho:
*local Mmu = 50*(1/50)*0
local Mmu = `nsnps'*(1/`uobs')*0

*The expected spread under the Ho:
*local Msd = ((50*(1/50)^2)*1)^0.5
local Msd = ((`nsnps'*(1/`uobs')^2)*1)^0.5

di _n
di "expected mean M statistic = `Mmu', SD = `Msd', Snps = `nsnps'"


*Ranking M statistics
*sort aggregated ustas to determine rank
sort usta_mean
egen rank = group(usta_mean)


**** end of part 3:calculate Mstatistic (ustamean) i.e. aggregate ustas into means ******


local alpha = 0.05

generate pval_ustamean = .
label variable pval_ustamean "un-corrected pvalue for Mstatistic"

generate bonfpval_ustamean = .
label variable bonfpval_ustamean "2-sided bonferroni corrected pvalue for Mstatistic"

gen zustamean = (usta_mean - `Mmu')/`Msd'


*using 2sided pval
quietly: replace pval_ustamean = 2*(normal(-abs(zustamean)))
quietly: replace bonfpval_ustamean = pval_ustamean * `nstudies'
quietly: replace bonfpval_ustamean = 1 if bonfpval_ustamean > 1
qqvalue pval_ustamean,method(simes) qvalue(q_simes_ustamean)


local zustamean_threshold = invnormal((`alpha'/`nstudies')/2)


di "critical z-value: `zustamean_threshold'"


local Mstatistic_threshold = (`zustamean_threshold' * `Msd') + `Mmu'
```

```
            di "critical M statistic value: `Mstatistic_threshold'"


            *** end of part 4b:Generate ustamean bonf_pvalues and calculate Mstatistic (usta_mean)
                ↪ threshold ***


            *Compute average variant effect size

            gen study_mean_beta = .
            gen study_N_beta = .

            *mean variant/snp effect size per study
            tabstat `beta', by (study) stat(mean n) save


            foreach study of numlist 1(1) `nstudies' {


            matrix mean_beta_mat=r(Stat`study')  // has 1 column (beta) and 2 rows(mean, N)

            local mean_beta=mean_beta_mat[1,1] // mean beta for the study
            local N_beta=mean_beta_mat[2,1] // N number of variants/snps

            di "`mean_info'"

            quietly: replace study_mean_beta = `mean_beta' if study == `study'
            quietly: replace study_N_beta = `N_beta' if study == `study'
                    }  // end of loop: that finds the mean variant/snp effect size in each study


            *take anti log of beta(log odds ratio) to determine the percentage strength of the outlier
                ↪ i.e.
            * how much more stronger the outlier is than average
            gen oddsratio = exp(study_mean_beta)


            ***** end of part 4:calculate average variant effect size for each study *****************



}



    rename usta_mean Mstatistic
    rename ustamean_se M_se
    rename usta_sd M_sd
    rename bonfpval_ustamean bonfpvalue
    rename study_mean_beta average_variant_effectsize


/* Save results: save M statistic results */

    *generate timestamp
    local c_date = c(current_date)
    local c_time = c(current_time)

    local c_time_date = "`c_date'"+"_"+"`c_time'"
    local time_string = subinstr("`c_time_date'", ":", "_", .)
    local time_string = subinstr("`time_string'", " ", "_", .)

if ("`savedataset'" == "savedataset") {

            *labeling the variables for easy comprehension
            label variable tau2 "tau_squared, estimates of between-study heterogeneity"
```

```
            label variable I2  "I_squared, proportion of total variation due to between study variance
                ↪ "
            label variable Q "Cochran's Q"
            label variable xb "fitted values excluding random effects"
            label variable usta "standardized predicted random effect (SPRE)"
            label variable xbu "fitted values including random effects"
            label variable stdxbu "standard error of prediction (fitted values) including random
                ↪ effects"
            label variable hat "diagonal elements of the projection hat matrix"
            label variable average_variant_effectsize "mean of beta values in each study"
            label variable oddsratio "average variant effect size expressed as oddsratio"
            label variable M_se "M statistic standard errors"
            label variable M_sd "M statistic standard deviations"
            label variable study "study number"
            label variable snp "variant number"

    savesome Mstatistic M_se M_sd `beta' `se_lambda_corr' `variants' `studies' lowerbound upperbound
        ↪ bonfpvalue tau2 I2 Q xb usta xbu stdxbu hat study snp average_variant_effectsize oddsratio
        ↪ using Mstatistic_results_`time_string'.dta

    }


/* Print results: identify and display outlier studies */


if ("`print'" != "noprint") {

            di as text _n "Multi-variant heterogeneity statistic (M)"

            di as text _n "Number of variants = " as result %7.0f `nsnps' /*
            */ _col(55) as text "Number of studies = " _col(70) as result %7.0f `nstudies'

            di as text "Expected Mean = " as result %12.4g `Mmu' /*
            */ _col(55) as text "Expected SD = " _col(70) as result %12.4g `Msd'

            di as text "M critical = +/-" as result %12.4g abs(`Mstatistic_threshold') /*
            */ _col(55) as text "(alpha = 0.05, 2-sided Bonferroni correction)"



            if (`Mstatistic_threshold' > 0) {
                di as text _n "Influential studies (Systematically stronger than average)"
                capture tabstat Mstatistic lowerbound upperbound bonfpvalue if Mstatistic >= `
                    ↪ Mstatistic_threshold', by(`studies') nototal stat(mean)
                if (_rc == 0) {
                        tabstat Mstatistic lowerbound upperbound bonfpvalue if Mstatistic >= `
                            ↪ Mstatistic_threshold', by(`studies') nototal stat(mean)
                    }
                    else {
                        di as result "No influential studies found at alpha = 0.05"
                    }

                    di as text _n "Underperforming studies (Systematically weaker than average)"
                    capture tabstat Mstatistic lowerbound upperbound bonfpvalue if Mstatistic <= -`
                        ↪ Mstatistic_threshold', by(`studies') nototal stat(mean)
                    if (_rc == 0) {
                        tabstat Mstatistic lowerbound upperbound bonfpvalue if Mstatistic <= -`
                            ↪ Mstatistic_threshold', by(`studies') nototal stat(mean)
                    }
                    else {
                        di as result "No underperforming studies found at alpha = 0.05"
                    }

                }

                else {
```

```
        di as text _n "Underperforming studies (Systematically weaker than average)"
        capture tabstat Mstatistic lowerbound upperbound bonfpvalue if Mstatistic <= `
            ↪ Mstatistic_threshold', by(`studies') nototal stat(mean)
        if (_rc == 0) {
                tabstat Mstatistic lowerbound upperbound bonfpvalue if Mstatistic <= `
                    ↪ Mstatistic_threshold', by(`studies') nototal stat(mean)
            }
            else {
                di as result "No underperforming studies found at alpha = 0.05"
                }

            di as text _n "Influential studies (Systematically stronger than average)"
            capture tabstat Mstatistic lowerbound upperbound bonfpvalue if Mstatistic >= -`
                ↪ Mstatistic_threshold', by(`studies') nototal stat(mean)
            if (_rc == 0) {
                tabstat Mstatistic lowerbound upperbound bonfpvalue if Mstatistic >= -`
                    ↪ Mstatistic_threshold', by(`studies') nototal stat(mean)
            }
            else {
                di as result "No influential studies found at alpha = 0.05"
            }

        } //end of if: (`Mstatistic_threshold' > 0)



    if ("`latexout'" == "latexout") {

        di as text _n "Displaying latex output: "

            if (`Mstatistic_threshold' > 0) {
                    di as text _n "Influential studies (Systematically stronger than average)"
                    capture latabstat Mstatistic lowerbound upperbound bonfpvalue if usta_mean
                        ↪  >= `Mstatistic_threshold', by(`studies') nototal stat(mean) cap(
                        ↪ Underperforming studies showing \\ systematically weaker effects
                        ↪ than average)
                    if (_rc == 0) {
                        latabstat Mstatistic lowerbound upperbound bonfpvalue if usta_mean >=
                            ↪ `Mstatistic_threshold', by(`studies') nototal stat(mean) cap(
                            ↪ Underperforming studies showing \\ systematically weaker
                            ↪ effects than average)
                    }
                    else {
                        di as result "No influential studies found at alpha = 0.05"
                    }

                    di as text _n "Underperforming studies (Systematically weaker than average
                        ↪ )"
                    capture latabstat Mstatistic lowerbound upperbound bonfpvalue if usta_mean
                        ↪  <= -`Mstatistic_threshold', by(`studies') nototal stat(mean) cap(
                        ↪ Influential studies showing \\ systematically stronger effects than
                        ↪  average)
                    if (_rc == 0) {
                    latabstat Mstatistic lowerbound upperbound bonfpvalue if usta_mean <= -`
                        ↪ Mstatistic_threshold', by(`studies') nototal stat(mean) cap(
                        ↪ Influential studies showing \\ systematically stronger effects than
                        ↪  average)
                    }
                    else {
                        di as result "No underperforming studies found at alpha = 0.05"
                    }

                    }

                    else {
```

```
                        di as text _n "Underperforming studies (Systematically weaker than average)"
                        capture latabstat Mstatistic lowerbound upperbound bonfpvalue if Mstatistic <=
                            ↪ `Mstatistic_threshold', by(`studies') nototal stat(mean) cap(
                            ↪ Underperforming studies showing \\ systematically weaker effects than
                            ↪ average)
                        if (_rc == 0) {
                                latabstat Mstatistic lowerbound upperbound bonfpvalue if Mstatistic <=
                                    ↪ `Mstatistic_threshold', by(`studies') nototal stat(mean) cap(
                                    ↪ Underperforming studies showing \\ systematically weaker
                                    ↪ effects than average)
                            }
                            else {
                                    di as result "No underperforming studies found at alpha = 0.05"
                                    }

                            di as text _n "Influential studies (Systematically stronger than average)"
                            capture latabstat Mstatistic lowerbound upperbound bonfpvalue if
                                ↪ Mstatistic >= -`Mstatistic_threshold', by(`studies') nototal stat(
                                ↪ mean) cap(Influential studies showing \\ systematically stronger
                                ↪ effects than average)
                            if (_rc == 0) {
                                    latabstat Mstatistic lowerbound upperbound bonfpvalue if
                                        ↪ Mstatistic >= -`Mstatistic_threshold', by(`studies')
                                        ↪ nototal stat(mean) cap(Influential studies showing \\
                                        ↪ systematically stronger effects than average)
                            }
                            else {
                                    di as result "No influential studies found at alpha = 0.05"
                            }
                            } // end of if: (`Mstatistic_threshold' > 0)

                    } // end of if: ("`latexout'" == "latexout")


            } // end of if: ("`print'" != "noprint")


/* Generate plots */

if ("`nograph'" != "nograph") {


    *generate lines marking 5 percent significance threshold
    local line1 `Mstatistic_threshold'
    local line2 = -1 * `Mstatistic_threshold'

            *plot M statistic(usta_mean) against average variant effect size
            scatter Mstatistic average_variant_effectsize, xscale(log) mlabel(study) yline(`line1' `
                ↪ line2', lcolor(gs15)) ylabel(, nogrid) xtitle("Average variant effect-size (log
                ↪ oddsratio)") graphregion(color(gs15)) mcolor(gs12) mlabcolor(gs4) name("M_nobars_`
                ↪ time_string'")
            *twoway (scatter Mstatistic average_variant_effectsize, xscale(log) mlabel(study) yline(`
                ↪ line1' `line2', lcolor(gs15)) ylabel(, nogrid) ytitle("M statistic") xtitle("
                ↪ Average variant effect-size (log oddsratio)") graphregion(color(gs15)) mcolor(gs12)
                ↪  mlabcolor(gs4)) (rcap upperbound lowerbound average_variant_effectsize, lcolor(
                ↪ gs13) lwidth(vthin)), legend(off) name("M_`time_string'")


            *Generate a histogram of the M statistics
            twoway histo Mstatistic, name("M_histogram_`time_string'")


    }


/* Stored results */
```

```
        return scalar number_variants = `nsnps'
        return scalar number_studies = `nstudies'
        return scalar M_expected_mean = `Mmu'
        return scalar M_expected_sd = `Msd'
        return scalar M_critical_alpha_0_05 = `Mstatistic_threshold'
        return local `varlist'

    restore

*And that's all folks

end
```

getmstatistic: R package

getmstatistic_R_0.1.1/compute_mstatistics.R

```
# Author:     Lerato E. Magosi
# R version: 3.1.0 (2014-04-10)
# Platform:  x86_64-apple-darwin10.8.0 (64-bit)
# Date:      20Apr2017


# Goal: Copute M statistics to identify systematic (non-random multi-variant) heterogeneity
#       patterns in GWAS meta-analysis.


# Required libraries ------------------------
# foreign           # needed to load stata datasets
# metafor           # needed to estimate standardized predicted random effects
# ggplot2           # needed to plot M statistics
# psych             # needed for summary statistics
# gridExtra         # needed for pretty tables
# stargazer         # needed for exporting dataframes to latex
# stats             # needed for the following functions: qt, qnorm, pnorm, p.adjust
# utils             # needed for the following functions: str, head, tail
# grDevices         # needed for the following functions: dev.new, rainbow
# graphics          # needed for the hist function


# Calling globalVariables on the following variables to address
# the note: "no visible binding for global variable" generated by "R CMD check"
utils::globalVariables(c("usta_mean", "study_names", "study", "yval", "usta_lb", "usta_ub",
                         "xb", "tau2", "xbse", "rawresid", "uncondse", "oddsratio", "zustamean",
                         "pval_ustamean", "bonf_pval_ustamean"))


# Function: compute_m_statistics ------------------------
#
# goal: use standardized predicted random effects (obtained via metafor) to compute M statistics
#
# parameters:
#
#
#   beta_in                 (numeric)   vector of effect-sizes,
#   lambda_se_in            (numeric)   vector of standard errors genomically corrected at study-level,
#   study_names_in          (character) vector of study names,
#   variant_names_in        (character) vector of variant neames,
#   tau2_method             (character) method to estimate heterogeneity: either "DL" or "REML",
#   x_axis_increment_in     (numeric)   value by which x-axis of M scatterplot will be incremented,
#   x_axis_round_in         (numeric)   value to which x-axis labels of M scatterplot will be rounded
#
#
# returns: a list containing:
#
#
#       M_expected_mean           (numeric),
#       M_expected_sd             (numeric),
#       M_crit_alpha_0_05 (numeric),
#   number_variants           (numeric),
#   number_studies            (numeric),
#       M_dataset                 (dataframe),
#       influential_studies_0_05  (dataframe),
#       weaker_studies_0_05       (dataframe)
#
# ------------------------------------------------------------------------------------------



compute_m_statistics <- function(beta_in, lambda_se_in, variant_names_in,
                                 study_names_in, tau2_method = "DL",
```

```
                            x_axis_increment_in = 0.02, x_axis_round_in = 2,
                            verbose_output = FALSE) {

    # Assemble dataset

    beta <- base::as.numeric(beta_in)
    lambda_se <- base::as.numeric(lambda_se_in)
    study_names <- base::factor(study_names_in)
    variant_names <- base::factor(variant_names_in)

    m <- base::data.frame(beta, lambda_se, variant_names, study_names)

# View dataset structure
if (verbose_output) utils::str(m)

# ------------------------------


# Assign study numbers
    m$study <- m$study_names
    base::levels(m$study) <- base::seq_along(base::levels(m$study))

# Calculate no. of studies
nstudies <- base::nlevels(m$study_names)

# Assign snp/variant numbers
m$snp <- m$variant_names
base::levels(m$snp) <- base::seq_along(base::levels(m$snp))

# Calculate no. of variants
nsnps <- base::nlevels(m$variant_names)

# Set alpha at 5% level
    usta_alpha <- 0.05


# Print numbers of studies and snps
    if (verbose_output) {

            base::writeLines("\n")
            base::print(base::paste("Summary: This heterogeneity analysis is based on ", nsnps, " SNPs
                ↪  and ", nstudies, " studies"))
            base::print("***************** end of part 1: assign snp and study numbers
                ↪ ****************")
            base::writeLines("")

            }


    # ------------------------------------


# Define function to align study effects i.e. betas
align_betas <- function(dframe_current_snp) {

    if (tau2_method == "DL") {

                    metafor_res <- metafor::rma.uni(yi = dframe_current_snp[, "beta"], sei =
                        ↪ dframe_current_snp[, "lambda_se"], weighted = T, knha = T, method =
                        ↪ tau2_method)

                    if (metafor_res$b[1] < 0) {

                            if (verbose_output) { base::print(base::paste0("Aligning study effects in
                                ↪ variant: ", base::unique(dframe_current_snp[, "variant_names"]))) }

                            dframe_current_snp$beta <- -dframe_current_snp$beta
```

```
                    }
    }

    else {

                    metafor_res <- metafor::rma.uni(yi = dframe_current_snp[, "beta"], sei =
                        ↪ dframe_current_snp[, "lambda_se"], weighted = T, knha = T, method =
                        ↪ tau2_method, control=list(stepadj=0.5, maxiter=10000))

                    if (metafor_res$b[1] < 0) {

                            if (verbose_output) { base::print(base::paste0("Aligning study effects in
                                ↪ variant: ", base::unique(dframe_current_snp[, "variant_names"]))) }

                            dframe_current_snp$beta <- -dframe_current_snp$beta

                    }
    }


    dframe_current_snp

}


# Split dataset by snp to obtain mini-datasets for each snp, then align study effects i.e. betas
list_snp_minidatasets <- base::split(m, m$variant_names)
align_betas_results <- base::lapply(list_snp_minidatasets, align_betas)

    # ----------------------------------



# Define function to extract standardized shrunken residuals i.e. usta


metafor_weighted_least_squares_rand_effects_regression <- function(dframe_current_snp) {

    # Run random effects meta-analysis (inverse variance weighted least sq regression) on current snp

    if (tau2_method == "DL") {

                    metafor_results <- metafor::rma.uni(yi = dframe_current_snp[, "beta"], sei =
                        ↪ dframe_current_snp[, "lambda_se"], weighted = T, knha = T, method =
                        ↪ tau2_method)
    }
    else {

                    metafor_results <- metafor::rma.uni(yi = dframe_current_snp[, "beta"], sei =
                        ↪ dframe_current_snp[, "lambda_se"], weighted = T, knha = T, method =
                        ↪ tau2_method, control=list(stepadj=0.5, maxiter=10000))
    }


    # Compute predicted values for the metafor_results
    metafor_results_predict <- metafor::predict.rma(metafor_results, digits = 8)

            # Warning note: Although the prediction's including random effects were similar i.e.
            # (Empirical Bayes estimates in metafor and xbu in Stata's metareg) the values of
            # their corresponding standard errors were quite different.


    # Compute Best Linear Unbiased Predictions i.e. Blups for the metafor_results
    metafor_results_blup <- metafor::blup.rma.uni(metafor_results, digits = 8)

    # Compute outlier diagnostics
```

```
    metafor_results_influence <- metafor::influence.rma.uni(metafor_results)

    # Extract hat values i.e. leverage a.k.a diagonals of hat matrix
    metafor_results_influence_hat <- (metafor_results_influence$inf)$hat

    # Print metafor results
            if (verbose_output) {

                base::writeLines("")
        base::print(paste("Random effects meta-analysis for variant: ", base::unique(
            ↪ dframe_current_snp$variant_names)))
        base::writeLines("")
        base::print(paste("snp_no: ", base::unique(dframe_current_snp$snp)))
        base::writeLines("")

        base::print("metafor_results: ")
                base::print(metafor_results)
                base::writeLines("")

                }


    # Create a data.frame comprising results to return
    output <- base::data.frame(
        dframe_current_snp,
        tau2 = metafor_results$tau2,           # estimate of between-study variance
        I2 = metafor_results$I2,               # Higgins inconsistency metric
        Q = metafor_results$QE,                # Q-statistic
        xb = metafor_results_predict$pred,     # predicition excluding random effects
        xbse = metafor_results_predict$se,     # standard error of prediction excluding random effects
        xbu = metafor_results_blup$pred,       # predictions including random effects
        stdxbu = metafor_results_blup$se,      # corresponding std. error of Empirical Bayes estimates
            ↪  in metafor
        hat = metafor_results_influence_hat,   # leverage a.k.a diagonals of hat matrix
        row.names = base::with(dframe_current_snp, base::interaction(variant_names, study_names)))

    output

}



metafor_weighted_least_squares_rand_effects_regression_results <- base::lapply(align_betas_results,
    ↪ metafor_weighted_least_squares_rand_effects_regression)

# View structure
    if (verbose_output) {

            base::writeLines("\n")
            base::writeLines("Dataset structure for individual variants: ")
            base::writeLines("")
            utils::str(metafor_weighted_least_squares_rand_effects_regression_results)
            base::writeLines("")

            }

# Append snp mini-datasets into a single file
meta_analysis_results <- base::data.frame(base::do.call(rbind,
    ↪ metafor_weighted_least_squares_rand_effects_regression_results))


# Compute raw residuals, unconditional standard error and ustas (a.k.a spres: standardized predicted
    ↪ random effects)
# notes: R. M. Harbord and J. P. T. Higgins, pg. 517 sbe23, Stata journal
    # Reference: Meta-regression in Stata, The Stata Journal (2008) 8, Number 4, pp. 493 to 519

meta_analysis_results_incl_spres <- meta_analysis_results
```

```
meta_analysis_results_incl_spres$rawresid <- base::with(meta_analysis_results_incl_spres, beta - xb)
meta_analysis_results_incl_spres$uncondse <- base::with(meta_analysis_results_incl_spres, base::sqrt(
    ↪ lambda_se**2 + tau2 - xbse**2))
meta_analysis_results_incl_spres$usta <- base::with(meta_analysis_results_incl_spres, rawresid /
    ↪ uncondse)



# View structure
    if (verbose_output) {

            base::writeLines("\n")

            base::writeLines("Dataset structure for whole dataset: ")
            base::writeLines("")

            utils::str(meta_analysis_results_incl_spres)
            base::writeLines("")


            }


    # ------------------------------------


# Define function to compute M-statistics for each study by taking the mean of their ustas
compute_ustamean <- function(dframe_current_study) {

    # Extract mean, sd and number of observations for usta variable in current study
    usta_describe <- psych::describe(dframe_current_study$usta)

            umean <- usta_describe$mean
            umean_se <- usta_describe$se
            usd <- usta_describe$sd
            uobs <- usta_describe$n

            # Determine lower and upper bound of Bonferroni corrected usta mean confidence interval
                ↪ for current study
            usta_df <- uobs - 1
            usta_tcrit_95 <- stats::qt(1 - ((usta_alpha / 2) / nstudies), usta_df)
            ulb <- umean - (usta_tcrit_95 * usd / base::sqrt(uobs))
            uub <- umean + (usta_tcrit_95 * usd / base::sqrt(uobs))

            # Top up allocates missing snps the average mean and sd
            #topup <- (nsnps - uobs) * (1 / uobs) * 0
            umean_topup <- (nsnps - uobs) * ((1 / uobs) * 0)



            # Create a data.frame comprising results to return
    output_compute_ustamean <- base::data.frame(dframe_current_study,
                                            usta_mean = umean + umean_topup,      # M-statistics
                                            usta_mean_se = umean_se,              # M-statistic
                                                ↪ standard errors
                                            usta_sd = usd,                        # M-statistic
                                                ↪ standard deviations
                                            usta_lb = ulb,                        # M-statistic
                                                ↪ lower bound
                                            usta_ub = uub,                        # M-statistic
                                                ↪ upper bound
                                                    row.names = base::with(
                                                        ↪ dframe_current_study, base::
                                                        ↪ interaction(variant_names,
                                                        ↪ study_names)))  # label row names
```

```
                                                    ↪ with study and variant names

            if (verbose_output) {

                base::print(paste0("study_name: ", base::unique(dframe_current_study$study_names),
                "; Mstatistic(usta_mean): ", base::unique(output_compute_ustamean$usta_mean),
                "; Mstatistic std.error(usta_mean_se): ", base::unique(
                    ↪ output_compute_ustamean$usta_mean_se),
                "; CI: ", base::unique(output_compute_ustamean$usta_lb), " ", base::unique(
                    ↪ output_compute_ustamean$usta_ub)))
                base::writeLines("\n")
                base::print("Summary statistics of the usta variable")
                base::writeLines("\n")
                base::print(usta_describe, digits = 8)
                base::writeLines("\n")

                }


    output_compute_ustamean



}


# Split dataset by study to obtain mini-datasets for each study, then compute M-statistics
list_study_minidatasets <- base::split(meta_analysis_results_incl_spres,
    ↪ meta_analysis_results_incl_spres$study_names)
compute_ustamean_results <- base::lapply(list_study_minidatasets, compute_ustamean)


    # View structure
    if (verbose_output) {

            base::writeLines("\n")

            base::writeLines("Dataset structure for individual studies: ")
            base::writeLines("")

            utils::str(compute_ustamean_results)

            }

# Append snp mini-datasets into a single file
mstatistic_results <- base::data.frame(base::do.call(rbind, compute_ustamean_results))


    if (verbose_output) {

            base::writeLines("Dataset structure after computing M-statistics i.e. ustamean: ")
            base::writeLines("")

            utils::str(mstatistic_results)

            }

    # ------------------------------------


    # Calculating the expected mean and sd for the M-statistic

    # The expected mean under the Ho:
    # If have 50 variants, Mmu = 50 * (1 / 50) * 0
    Mmu <- nsnps * (1 / nsnps) * 0
```

```
    # The expected spread under the Ho:
    # If have 50 variants, Msd = ((50 * (1 / 50)^2) * 1)^0.5
    Msd <- ((nsnps * (1 / nsnps)^2) * 1)^0.5

base::writeLines("")
    if (verbose_output) base::print(base::paste("expected mean for M statistic = ", Mmu, "; SD = ",
        ↪ Msd, "; Snps = ", nsnps))

    # ------------------------------------



# Compute average variant effect-size

compute_avg_variant_effectsize <- function(dframe_mstatistic_results) {

            # Sort aggregated ustas (i.e. ustamean) to determine rank
            dframe_mstatistic_results_srtdby_usta_mean <- dframe_mstatistic_results[base::order(
                ↪ dframe_mstatistic_results[, "usta_mean"]), ]

            # Rank studies by usta_mean
            dframe_mstatistic_results_srtdby_usta_mean$rank <- base::factor(
                ↪ dframe_mstatistic_results_srtdby_usta_mean[, "usta_mean"])
            base::levels(dframe_mstatistic_results_srtdby_usta_mean$rank) <- base::seq_along(base::
                ↪ levels(dframe_mstatistic_results_srtdby_usta_mean$rank))


    # Extract beta values to compute average variant effect-size
            list_betas_splitby_study_names <- base::split(dframe_mstatistic_results_srtdby_usta_mean[,
                ↪  "beta"], dframe_mstatistic_results_srtdby_usta_mean[, "study_names"])
            list_describe_betas <- base::lapply(list_betas_splitby_study_names, psych::describe)


            if (verbose_output) {
                base::writeLines("")
                    base::print(" Summary statistics for effect-sizes (betas) in each study")
                    base::print(list_describe_betas)
            }



    # Compute average variant effect-size and number of variants in each study
    populate_study_mean_beta_vec <- function(current_study_betas) {

        current_study_mean_beta <- current_study_betas$mean

    }

    study_mean_beta <- base::unlist(base::lapply(list_describe_betas, populate_study_mean_beta_vec))



    populate_study_n_beta_vec <- function(current_study_betas) {

        current_study_n_beta <- current_study_betas$n

    }

    study_n_beta <- base::unlist(base::lapply(list_describe_betas, populate_study_n_beta_vec))

    # Note: list_describe_betas is sorted by study_names hence need to
    #       re-sort dframe_mstatistic_results_srtdby_usta_mean before assembling the output
    #       data.frame

    dframe_mstatistic_srtd_by_study_names <- dframe_mstatistic_results_srtdby_usta_mean[base::order(
        ↪ dframe_mstatistic_results_srtdby_usta_mean[, "study_names"]), ]
```

```
            output_compute_avg_variant_effectsize <- base::data.frame(
                    study_names = base::unique(dframe_mstatistic_srtd_by_study_names$study_names),
                    usta_mean = base::unique(dframe_mstatistic_srtd_by_study_names$usta_mean),
                        ↪ # M-statistics
                    usta_mean_se = base::unique(dframe_mstatistic_srtd_by_study_names$usta_mean_se),
                        ↪ # M-statistic standard errors
                    study_mean_beta,
                        ↪ # average variant effect-sizes
                    oddsratio = base::exp(study_mean_beta),
                        ↪ # average variant effect-sizes expressed as oddsratios
                    study_n_beta,
                        ↪ # number of loci in each study
                    study = base::unique(dframe_mstatistic_srtd_by_study_names$study),
                        ↪ # study numbers
                    rank = base::unique(dframe_mstatistic_srtd_by_study_names$rank),
                        ↪ # study ranks determined by size of M-statistics
                    row.names = NULL)


    output_compute_avg_variant_effectsize <- output_compute_avg_variant_effectsize[base::order(
        ↪ output_compute_avg_variant_effectsize[, "study_names"]), ]

            output_compute_avg_variant_effectsize


}


compute_avg_variant_effectsize_results <- compute_avg_variant_effectsize(mstatistic_results)

    # ------------------------------------



# Generate plots

    # Histogram of M statistics
    filename_histogram_mstats <- base::paste0("Histogram_Mstatistics_", nstudies, "studies_", nsnps, "
        ↪ snps.tif")
    grDevices::tiff(filename_histogram_mstats, width = 17.35, height = 23.35, units = "cm", res = 300,
        ↪  compression = "lzw", pointsize = 14)
    usta_mean_hist <- base::unique(mstatistic_results[, c("study_names", "usta_mean")])
    graphics::hist(usta_mean_hist[, "usta_mean"], main="Histogram of M statistics", xlab="M statistics
        ↪ ")
    grDevices::dev.off()

    # ** * **


    # Plot M statistic(ustamean) against average variant effect size



    # Subset compute_avg_variant_effectsize_results into strong and weak studies
    compute_avg_variant_effectsize_results_srtby_usta_mean <- compute_avg_variant_effectsize_results[
        ↪ base::order(compute_avg_variant_effectsize_results[, "usta_mean"]), ]

    usta_mean_scatter_strong <- base::subset(compute_avg_variant_effectsize_results_srtby_usta_mean,
        ↪ compute_avg_variant_effectsize_results_srtby_usta_mean$usta_mean >= 0)
    usta_mean_scatter_strong$strength <- base::rep("strong", base::nrow(usta_mean_scatter_strong))

    usta_mean_scatter_weak <- base::subset(compute_avg_variant_effectsize_results_srtby_usta_mean,
        ↪ compute_avg_variant_effectsize_results_srtby_usta_mean$usta_mean < 0)
usta_mean_scatter_weak$strength <- base::rep("weak", base::nrow(usta_mean_scatter_weak))

# Now that we have succesfully added the strength column, we shall
```

```
    # combine the usta_mean_scatter_strong and usta_mean_scatter_weak subsets

    usta_mean_scatter_strength <- base::rbind(usta_mean_scatter_weak, usta_mean_scatter_strong)


    # Calculating expected i.e. theoretical Mstatistic threshold
    mstat_threshold_zscore <- base::abs(stats::qnorm((usta_alpha / nstudies) / 2))
    mstat_threshold_zscore

    mstat_threshold <- mstat_threshold_zscore * base::sqrt((1 / nsnps^2) * nsnps) + ((1 / nsnps) * 0)
    mstat_threshold


    if (verbose_output) {

        base::writeLines("\n")
        base::print("Table of Mstatistics highlighting strong (M >= 0) and weak (M < 0) studies: ")
        base::writeLines("\n")
        base::print(usta_mean_scatter_strength)
        base::writeLines("\n")
        base::print(base::paste0("Mstatistic threshold: ", mstat_threshold))


        }


  # Set ustamean threshold
  dat_hlines <- data.frame(strength = base::levels(base::as.factor(usta_mean_scatter_strength$strength)
      ↪ ), yval = c(-mstat_threshold, mstat_threshold))

    # Plotting - with study numbers
    x_axis_min <- base::min(base::log(usta_mean_scatter_strength$oddsratio))
    x_axis_max <- base::max(base::log(usta_mean_scatter_strength$oddsratio))

    filename_mstats_vs_avg_effectsize <- base::paste0("Mstatistics_vs_average_variant_effectsize_",
        ↪ nstudies, "studies_", nsnps, "snps.tif")
    grDevices::tiff(filename_mstats_vs_avg_effectsize, width = 23.35, height = 17.35, units = "cm",
        ↪ res = 300, compression = "lzw", pointsize = 14)
    h <- ggplot2::ggplot(usta_mean_scatter_strength, ggplot2::aes(base::log(oddsratio), usta_mean,
        ↪ colour = usta_mean, label = study_names)) + ggplot2::geom_point(size = 4.5) + ggplot2::
        ↪ geom_text(ggplot2::aes(label = study), hjust = 1.2, vjust = -0.5, size = 2.5, colour = "
        ↪ azure4") + ggplot2::scale_colour_gradientn(name = "M statistic", colours = grDevices::
        ↪ rainbow(11)) + ggplot2::scale_x_continuous(trans="log", limits=c(x_axis_min, x_axis_max),
        ↪ breaks=base::round(base::seq(x_axis_min, x_axis_max, x_axis_increment_in),x_axis_round_in),
        ↪  minor_breaks=ggplot2::waiver(), labels = base::round(base::exp(base::seq(x_axis_min,
        ↪ x_axis_max, x_axis_increment_in)),x_axis_round_in)) + ggplot2::theme_bw() + ggplot2::
        ↪ scale_fill_hue(c = 45, l = 40) + ggplot2::xlab("Average effect size (oddsratio)") + ggplot2
        ↪ ::ylab("M statistic") + ggplot2::theme(panel.grid.minor = ggplot2::element_blank(), panel.
        ↪ grid.major = ggplot2::element_blank()) + ggplot2::theme(axis.title.x = ggplot2::
        ↪ element_text(size = 14), axis.text.x = ggplot2::element_text(size = 14)) + ggplot2::theme(
        ↪ axis.title.y = ggplot2::element_text(size = 14), axis.text.y = ggplot2::element_text(size =
        ↪  14))
    hi <- h + ggplot2::geom_hline(ggplot2::aes(yintercept = yval), data = dat_hlines, colour = "grey80
        ↪ ", linetype = "dashed", lwd = 0.4) + ggplot2::theme(legend.text = ggplot2::element_text(
        ↪ size = 10)) + ggplot2::theme(legend.title = ggplot2::element_text(size = 12)) + ggplot2::
        ↪ theme(legend.position = "bottom")
    base::print(hi + ggplot2::geom_hline(ggplot2::aes(yintercept = c(0,0)), data = dat_hlines, colour
        ↪ = "grey80", linetype = "solid", lwd = 0.4))
    grDevices::dev.off()

    # ------------------------------------


# Generate tables

 # Compute 2-sided Bonferroni correction of M statistic p-values
```

```
    # Using 2sided pval
usta_mean_scatter_strength$zustamean <- base::with(usta_mean_scatter_strength, (usta_mean - Mmu) / Msd
    ↪ )
usta_mean_scatter_strength$pval_ustamean <- base::with(usta_mean_scatter_strength, 2 * (stats::pnorm(-
    ↪ base::abs(zustamean))))
usta_mean_scatter_strength$bonf_pval_ustamean <- base::with(usta_mean_scatter_strength, pval_ustamean
    ↪ * nstudies)
usta_mean_scatter_strength$bonf_pval_ustamean[usta_mean_scatter_strength$bonf_pval_ustamean > 1] <- 1
usta_mean_scatter_strength$qval_ustamean <- base::with(usta_mean_scatter_strength, stats::p.adjust(
    ↪ bonf_pval_ustamean, method = "fdr"))


# Diagnose influential studies and underperforming studies
usta_mean_influential_studies <- base::subset(usta_mean_scatter_strength,
    ↪ usta_mean_scatter_strength$usta_mean >= mstat_threshold)

usta_mean_influential_studies <- usta_mean_influential_studies[, c("study_names", "usta_mean", "
    ↪ bonf_pval_ustamean")]
base::names(usta_mean_influential_studies) <- c("Study", "M", "Bonferroni_pvalue")

if (base::nrow(usta_mean_influential_studies) >= 1) {
        title_influential_studies <- "Table: M and Bonferroni p-values \nof systematically stronger
            ↪ studies \nat the 5% significant level."

            filename_influential_studies <- base::paste0("table_influential_studies.tif")
            grDevices::tiff(filename_influential_studies, width = 17.35, height = 23.35, units = "cm",
                ↪ res = 300, compression = "lzw", pointsize = 14)
    table_influential_studies <- draw_table(body = usta_mean_influential_studies, heading =
        ↪ title_influential_studies)
    grDevices::dev.off()

} else {
    base::writeLines("")
    base::print("Table not generated as there were no influential \nstudies found at alpha = 0.05")

}



    # Latex output
    base::writeLines("")
    if (base::nrow(usta_mean_influential_studies) >= 1) {
        stargazer::stargazer(usta_mean_influential_studies, summary = FALSE, rownames = FALSE, title =
            ↪ "M statistics and Bonferroni p-values showing systematically stronger studies at the 5
            ↪ percent significant level.", style = "aer")
} else {
    base::writeLines("")
    base::print("Latex table not generated as there were no influential studies found at alpha =
        ↪ 0.05")

}

    # ---- ** ----


usta_mean_underperforming_studies <- base::subset(usta_mean_scatter_strength,
    ↪ usta_mean_scatter_strength$usta_mean <= -mstat_threshold)

usta_mean_underperforming_studies <- usta_mean_underperforming_studies[, c("study_names", "usta_mean",
    ↪ "bonf_pval_ustamean")]
base::names(usta_mean_underperforming_studies) <- c("Study", "M", "Bonferroni_pvalue")

if (base::nrow(usta_mean_underperforming_studies) >= 1) {
        title_underperforming_studies <- "Table: M and Bonferroni p-values \nof systematically weaker
            ↪ studies \nat the 5% significant level"
```

```
            filename_underperforming_studies <- base::paste0("table_underperforming_studies.tif")
            grDevices::tiff(filename_underperforming_studies, width = 17.35, height = 23.35, units = "
                ↪ cm", res = 300, compression = "lzw", pointsize = 14)
    table_underperforming_studies <- draw_table(body = usta_mean_underperforming_studies, heading =
        ↪ title_underperforming_studies)
    grDevices::dev.off()

} else {
    base::writeLines("")
    base::print("Table not generated as there were no \nunder-performing studies found at alpha =
        ↪ 0.05")

}


    # Latex output
    base::writeLines("")
    if (base::nrow(usta_mean_underperforming_studies) >= 1) {
        stargazer::stargazer(usta_mean_underperforming_studies, summary = FALSE, rownames = FALSE,
            ↪ title = "M statistics and Bonferroni p-values showing systematically weaker studies at
            ↪ the 5 percent significant level.", style = "aer")
} else {
    base::print("Latex table not generated as there were no under-performing studies found at alpha =
        ↪ 0.05")

}


    # ------------------------------------



# Merge usta_mean_scatter_strength with mstatistic_results to generate Mdataset

# Since usta_mean, ustamean_se and study already appear in mstatistic_results, we shall exclude them
    ↪ from the merge
# get indices corresponding to usta_mean, ustamean_se and study
idx_ustamean <- base::which(base::names(usta_mean_scatter_strength) == "usta_mean")
idx_usta_mean_se <- base::which(base::names(usta_mean_scatter_strength) == "usta_mean_se")
idx_study <- base::which(base::names(usta_mean_scatter_strength) == "study")

mstatistic_results_incl_usta_mean_scatter_strength <- base::merge(mstatistic_results,
    ↪ usta_mean_scatter_strength[, -c(idx_ustamean, idx_usta_mean_se, idx_study)], by = "study_names
    ↪ ")

mstatistic_dataset <- mstatistic_results_incl_usta_mean_scatter_strength[, c("study_names", "beta", "
    ↪ lambda_se", "variant_names", "usta_mean", "usta_sd", "usta_mean_se",
                                                    "usta_lb", "usta_ub", "bonf_pval_ustamean", "
                                                        ↪ qval_ustamean", "tau2", "I2", "Q", "xb",
                                                        ↪  "usta",
                                                    "xbu", "stdxbu", "hat", "study", "snp", "
                                                        ↪ study_mean_beta", "oddsratio", "
                                                        ↪ study_n_beta")]

names(mstatistic_dataset) <- c("study_names_in", "beta_in", "lambda_se_in", "variant_names_in", "M", "
    ↪ M_sd", "M_se", "lowerbound", "upperbound", "bonfpvalue", "qvalue", "tau2", "I2", "Q", "xb", "
    ↪ usta",
                                                    "xbu", "stdxbu", "hat", "study", "snp", "
                                                        ↪ beta_mean", "oddsratio", "beta_n")
# View structure
base::writeLines("Structure for dataset of computed M statistics: ")
base::writeLines("")

utils::str(mstatistic_dataset)
base::writeLines("")
```

```
        # ------------------------------------


        # List of items to return
        list(M_expected_mean = Mmu,
        M_expected_sd = Msd,
        M_crit_alpha_0_05 = mstat_threshold,
        number_variants = nsnps,
        number_studies = nstudies,
        M_dataset = mstatistic_dataset,
        influential_studies_0_05 = usta_mean_influential_studies,
        weaker_studies_0_05 = usta_mean_underperforming_studies)


}
```

getmstatistic_R_0.1.1/data.R

```
#' heartgenes214.
#'
#' heartgenes214 is a multi-ethnic GWAS meta-analysis dataset for coronary
#' artery disease.
#'
#' It comprises summary data (effect-sizes and their corresponding standard
#'     errors) for 48 studies (68,801 cases and 123,504 controls), at 214 lead
#'     variants independently associated with coronary artery disease
#'     (P < 0.00005, FDR < 5\%). Of the 214 lead variants, 44 are genome-wide
#'     significant (p < 5e-08). The meta-analysis dataset is based on
#'     individuals of: African American, Hispanic American, East Asian,
#'     South Asian, Middle Eastern and European ancestry.
#'
#' The study effect-sizes have been flipped to ensure alignment of the
#'     effect alleles.
#'
#' Standard errors were genomically corrected at the study-level.
#'
#'
#' @format A data frame with seven variables:
#' \describe{
#' \item{\code{beta_flipped}}{Effect-sizes expressed as log odds ratios. Numeric}
#' \item{\code{gcse}}{Standard errors}
#' \item{\code{studies}}{Names of participating studies}
#' \item{\code{variants}}{Names of genetic variants/SNPs}
#' \item{\code{cases}}{Number of cases in each participating study}
#' \item{\code{controls}}{Number of controls in each participating study}
#' \item{\code{fdr214_gwas46}}{Flag indicating GWAS significant variants,
#'       1: Not GWAS-significant, 2: GWAS-significant}
#' }
#'
#' @source Magosi LE, Goel A, Hopewell JC, Farrall M, on behalf of the
#'         CARDIoGRAMplusC4D Consortium (2017) Identifying systematic
#'         heterogeneity patterns in genetic association meta-analysis studies.
#'         PLoS Genet 13(5): e1006755. https://doi.org/10.1371/journal.pgen.1006755.
#'
#'         \url{https://magosil86.github.io/getmstatistic/}
#'
"heartgenes214"
```

getmstatistic_R_0.1.1/driver_compute_mstatistics.R

```
# Goal: Generate and export S3 class/methods for getmstatistic and document using roxygen tags
#
#' Quantifying Systematic Heterogeneity in Meta-Analysis.
#'
#' \code{getmstatistic} computes \emph{M} statistics to assess the contribution
#' of each participating study in a meta-analysis. The \emph{M} statistic
#' aggregates heterogeneity information across multiple variants to, identify
#' systematic heterogeneity patterns and their direction of effect in
#' meta-analysis. It's primary use is to identify outlier studies, which either
#' show "null" effects or consistently show stronger or weaker genetic effects
#' than average, across the panel of variants examined in a GWAS meta-analysis.
#'
#' In contrast to conventional heterogeneity metrics (Q-statistic, I-squared and
#' tau-squared) which measure random heterogeneity at individual variants,
#' \emph{M} measures systematic (non-random) heterogeneity across multiple
#' independently associated variants.
#'
#' Systematic heterogeneity can arise in a meta-analysis due to differences in
#' the study characteristics of participating studies. Some of the differences
#' may include: ancestry, allele frequencies, phenotype definition,
#' age-of-disease onset, family-history, gender, linkage disequilibrium and
#' quality control thresholds. See the getmstatistic website for statistical
#' theory, documentation and examples.
#'
#' \code{getmstatistic} uses summary data i.e. study effect-sizes and their
#' corresponding standard errors to calculate \emph{M} statistics (One \emph{M}
#' for each study in the meta-analysis).
#'
#' In particular, \code{getmstatistic} employs the inverse-variance weighted
#' random effects regression model provided in the \code{metafor} R package
#' to extract SPREs (standardized predicted random effects) which are then
#' aggregated to formulate \emph{M} statistics.
#'
#'
#' @seealso \code{\link[metafor]{rma.uni}} function in \code{metafor} for random
#'     effects model, and \url{https://magosil86.github.io/getmstatistic/} for
#'     getmstatistic website.
#'
#' @aliases getm getmstat getmstatistic Rgetmstatistic
#'
#' @param beta_in A numeric vector of study effect-sizes e.g. log odds-ratios.
#' @param lambda_se_in A numeric vector of standard errors, genomically
#'       corrected at study-level.
#' @param study_names_in A character vector of study names.
#' @param variant_names_in A character vector of variant names e.g. rsIDs.
#' @param tau2_method A character scalar, method to estimate heterogeneity:
#'       either "DL" or "REML" (Optional). Note: The REML method uses the
#'       iterative Fisher scoring algorithm (step length = 0.5, maximum
#'       iterations = 10000) to estimate tau2.
#' @param x_axis_increment_in A numeric scalar, value by which x-axis
#'       of M scatterplot will be incremented (Optional).
#' @param x_axis_round_in A numeric scalar, value to which x-axis labels of
#'       M scatterplot will be rounded (Optional).
#' @param verbose_output An optional boolean to display intermediate output.
#' @param ... Further arguments.
#'
#' @return Returns a list containing:
#' \itemize{
#'   \item Mstatistic_expected_mean   , A numeric scalar for the expected mean for M
#'   \item Mstatistic_expected_sd     , A numeric scalar for the expected standard deviation
#'         for M
#'   \item number_studies             , A numeric scalar for the number of studies
#'   \item number_variants            , A numeric scalar for the number of variants
#'   \item Mstatistic_crit_alpha_0_05 , A numeric scalar of the critical M value at the 5
#'         percent significance level.
#'   \item M_dataset (dataframe) A dataset of the computed M statistics, which
```

```
#'         includes the following fields:
#'   \itemize{
#'         \item M           , Mstatistic
#'         \item M_sd        , standard deviation of M
#'         \item M_se        , standard error of M
#'         \item lowerbound  , lowerbound of M 95% confidence interval
#'         \item upperbound  , upperbound of M 95% confidence interval
#'         \item bonfpvalue  , 2-sided bonferroni pvalues of M
#'         \item qvalue      , false discovery rate adjusted pvalues of M
#'         \item tau2        , tau_squared, DL estimates of between-study
#' heterogeneity
#'         \item I2          , I_squared, proportion of total variation due to
#' between study variance
#'         \item Q           , Cochran's Q
#'         \item xb          , fitted values excluding random effects
#'         \item usta        , standardized predicted random effect (SPRE)
#'         \item xbu         , fitted values including random effects
#'         \item stdxbu      , standard error of prediction (fitted values)
#' including random effects
#'         \item hat         , diagonal elements of the projection hat matrix
#'         \item study       , study numbers
#'         \item snp         , variant numbers
#'         \item beta_mean   , average variant effect size
#'         \item oddsratio   , average variant effect size as oddsratio
#'         \item beta_n    , number of variants in each study
#'         }
#'   \item influential_studies_0_05 (dataframe) A dataset of influential studies
#'       significant at the 5 percent level.
#'   \item weaker_studies_0_05 (dataframe) A dataset of under-performing studies
#'       significant at the 5 percent level.
#' }
#'
#' @examples
#' library(getmstatistic)
#' library(gridExtra)
#'
#'
#' # Basic M analysis using the heartgenes214 dataset.
#' # heartgenes214 is a multi-ethnic GWAS meta-analysis dataset for coronary artery disease.
#' # To learn more about the heartgenes214 dataset ?heartgenes214
#'
#' # Running an M analysis on 20 GWAS significant variants (p < 5e-08) in the first 10 studies
#'
#' heartgenes44_10studies <- subset(heartgenes214, studies <= 10 & fdr214_gwas46 == 2)
#' heartgenes20_10studies <- subset(heartgenes44_10studies,
#'     variants %in% unique(heartgenes44_10studies$variants)[1:20])
#'
#' getmstatistic_results <- getmstatistic(heartgenes20_10studies$beta_flipped,
#'                                        heartgenes20_10studies$gcse,
#'                                        heartgenes20_10studies$variants,
#'                                        heartgenes20_10studies$studies)
#' getmstatistic_results
#'
#' # Explore results generated by getmstatistic function
#'
#' # Retrieve dataset of M statistics
#' dframe <- getmstatistic_results$M_dataset
#'
#' \donttest{
#'
#' str(dframe)
#'}
#'
#' # Retrieve dataset of stronger than average studies (significant at 5% level)
#' getmstatistic_results$influential_studies_0_05
#'
#' # Retrieve dataset of weaker than average studies (significant at 5% level)
```

```
#' getmstatistic_results$weaker_studies_0_05
#'
#' # Retrieve number of studies and variants
#' getmstatistic_results$number_studies
#' getmstatistic_results$number_variants
#'
#' # Retrieve expected mean, sd and critical M value at 5% significance level
#' getmstatistic_results$M_expected_mean
#' getmstatistic_results$M_expected_sd
#' getmstatistic_results$M_crit_alpha_0_05
#'
#'
#' # Additional examples: These take a little bit longer to run
#'
#' \donttest{
#'
#' # Run M analysis on all 214 lead variants
#' # heartgenes214 is a multi-ethnic GWAS meta-analysis dataset for coronary artery disease.
#' getmstatistic_results <- getmstatistic(heartgenes214$beta_flipped,
#'                                        heartgenes214$gcse,
#'                                        heartgenes214$variants,
#'                                        heartgenes214$studies)
#' getmstatistic_results
#'
#'
#' # Subset the GWAS significant variants (p < 5e-08) in heartgenes214
#' heartgenes44 <- subset(heartgenes214, heartgenes214$fdr214_gwas46 == 2)
#'
#' # Exploring getmstatistic options:
#' #     Estimate heterogeneity using "REML", default is "DL"
#' #     Modify x-axis of M scatterplot
#' #     Run M analysis verbosely
#' getmstatistic_results <- getmstatistic(heartgenes44$beta_flipped,
#'                                        heartgenes44$gcse,
#'                                        heartgenes44$variants,
#'                                        heartgenes44$studies,
#'                                        tau2_method = "REML",
#'                                        x_axis_increment_in = 0.03,
#'                                        x_axis_round_in = 3,
#'                                        verbose_output = TRUE)
#' getmstatistic_results
#'
#'
#' }
#'
#' @export
getmstatistic <- function(beta_in, lambda_se_in, study_names_in, variant_names_in, ...) UseMethod("
    ↪ getmstatistic")

#' @describeIn getmstatistic Computes M statistics
#' @export
getmstatistic.default <- function(beta_in, lambda_se_in, study_names_in,
                                  variant_names_in, tau2_method = "DL",
                                  x_axis_increment_in = 0.02, x_axis_round_in = 2,
                                  verbose_output = FALSE, ...) {

    # Check whether all required variables are present
    if (missing(beta_in))
        stop("Beta values missing, need to specify study effect-sizes.")

    if (missing(lambda_se_in))
        stop("Standard errors missing, need to specify standard errors.")

    if (missing(study_names_in))
        stop("Study names missing, need to specify study names.")

    if (missing(variant_names_in))
```

```
        stop("Variant names missing, need to specify variant names.")


    # Verify datatypes of required variables
    if (!is.numeric(c(beta_in, lambda_se_in))) {

        stop("beta_in and lambda_se_in should be of type, numeric.")

    }


    if (!is.character(c(variant_names_in, study_names_in))) {

        stop("variant_names_in and study_names_in should be of type, character.")

    }

    compute_m_results <- compute_m_statistics(beta_in, lambda_se_in, study_names_in, variant_names_in,
                                              tau2_method, x_axis_increment_in, x_axis_round_in,
                                              verbose_output)

    compute_m_results$call <- match.call()

    class(compute_m_results) <- "getmstatistic"

    compute_m_results

}

#' @export
print.getmstatistic <- function(x, ..., verbose_output = FALSE) {

    cat("Call:\n")
    print(x$call)

    cat("\nM_expected_mean:\n")
    print(x$M_expected_mean)

    cat("\nM_expected_sd:\n")
    print(x$M_expected_sd)

    cat("\nM_crit_alpha_0_05:\n")
    print(x$M_crit_alpha_0_05)

    cat("\ninfluential_studies_0_05:\n")
    print(x$influential_studies_0_05)

    cat("\nweaker_studies_0_05:\n")
    print(x$weaker_studies_0_05)


}
```

getmstatistic_R_0.1.1/getmstatistic_welcome.R

```
.onAttach <- function(libname, pkgname) {
  packageStartupMessage("Welcome to getmstatistic \n see https://magosil86.github.io/getmstatistic/ for
      ↪ examples and documentation")
}


.onLoad <- function(libname, pkgname) {
  op <- options()
  op.getmstatistic <- list(
    getmstatistic.path = "",
    getmstatistic.install.args = "",
    getmstatistic.name = "Lerato E. Magosi",
    getmstatistic.desc.author = '"Lerato E. Magosi <magosil86@gmail.com> [aut, cre]"',
    getmstatistic.desc.license = "MIT",
    getmstatistic.desc.suggests = NULL,
    getmstatistic.desc = list()
  )
  toset <- !(names(op.getmstatistic) %in% names(op))
  if(any(toset)) options(op.getmstatistic[toset])

  invisible()
}
```

getmstatistic_R_0.1.1/util_draw_table.R

```
#' Helper function to draw table grobs.
#'
#' \code{draw_table()} Pre and post version: 2.0.0 gridExtra packages
#' handle drawing tables differently. \code{draw_table()} determines
#' the installed version of gridExtra and applies the appropriate
#' syntax. If gridExtra version < 2.0.0 then it uses old gridExtra
#' syntax to build table Grob(graphical object) else uses new syntax.
#' \code{draw_table()}
#'
#' prints tables without rownames.
#'
#' @section Acknowledgements: Thanks to Ryan Welch,
#' https://github.com/welchr/LocusZoom/issues/16
#'
#'
#' @param body A dataframe. Table body.
#' @param heading A string. Table title.
#' @param ... Further arguments to control the gtable.
#'
#' @examples
#' # note: not exported hence examples are not run
#' library(gridExtra)
#'
#' \donttest{
#'
#' # Table of iris values
#' iris_dframe <- head(iris)
#' title_iris_dframe <- paste("Table: Length and width measurements (cm) of sepals ",
#'                        "and petals, for 50 flowers from 3 species of iris ",
#'                        "(setosa, versicolor, and virginica). ", sep="")
#' table_influential_studies <-  draw_table(body = iris_dframe, heading = title_iris_dframe)
#'
#' # Table of mtcars values
#' mtcars_dframe <- head(mtcars)
#' title_mtcars_dframe <- paste("Table: Motor Trend US magazine (1974) automobile ",
#'                         "statistics for fuel consumption, automobile ",
#'                         "design and performance. ", sep="")
#' table_influential_studies <-  draw_table(body = mtcars_dframe,
#'                                          heading = title_mtcars_dframe)
#'}

#' # @export
draw_table <- function(body, heading, ...) {

  ge_version <- as.character(utils::packageVersion("gridExtra"))
  vcomp = utils::compareVersion(ge_version,"2.0.0")

  if (vcomp == -1) {
    table_old <- gridExtra::tableGrob(

      body,
      show.rownames = FALSE,
      gpar.corefill = grid::gpar(fill = "white", col = "black"),
      gpar.colfill = grid::gpar(fill = "white", col = "black"),
      ...
    )

    # Assigning the var h the height of our table
        height_table_old <- grid::grobHeight(table_old)

        title_old <- grid::textGrob(heading, y=grid::unit(0.5, "npc") + 0.7*height_table_old,vjust=0,gp=
            ↪ grid::gpar(fontsize=12.0))
        gt_table_old <- grid::gTree(children=grid::gList(table_old, title_old))

        print(grid::grid.newpage())
        print(grid::grid.draw(gt_table_old))
```

```
  } else {

    tt <- gridExtra::ttheme_default(
        core=list(bg_params = list(fill = "white", col="black")),
        colhead=list(bg_params = list(fill = "white", col="black")))

    table <- gridExtra::tableGrob(

      body,
      theme = tt,
      rows = NULL,
      ...
    )

    title <- grid::textGrob(heading, gp=grid::gpar(fontsize=12), vjust=0)

    padding <- grid::unit(0.2,"npc")

    table <- gtable::gtable_add_rows(table, heights = grid::grobHeight(title) + padding, pos = 0)

    table <- gtable::gtable_add_grob(table, title, t=1, b=nrow(table), l=1, r=ncol(table))

   #rectangle grob dimensions: height:row3 --> nrows, width:col1 --> ncols
    table <- gtable::gtable_add_grob(table, grobs = grid::rectGrob(gp = grid::gpar(fill = NA, lwd = 2)), t
        ↪  = 3, b = nrow(table), l = 1, r = ncol(table))

   #rectangle grob dimensions: height:row2 --> row2, width:col1 --> ncols
    table <- gtable::gtable_add_grob(table, grobs = grid::rectGrob(gp = grid::gpar(fill = NA, lwd = 2)), t
        ↪  = 2, b = 2, l = 1, r = ncol(table))

    print(grid::grid.newpage())
    print(grid::grid.draw(table))


  }
}
```