2016

Meta-analysis strategies for heterogeneous studies in genome-wide association studies

https://hdl.handle.net/2144/17065

Boston University
Dissertation

META-ANALYSIS STRATEGIES FOR HETEROGENEOUS STUDIES IN GENOME-WIDE ASSOCIATION STUDIES

by

JAEGYOUNG HONG

B.S., Kookmin University, 2002
M.S., University of Illinois at Chicago, 2011

Submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

2016
Approved by

First Reader

________________________________________________________________________

Ching-Ti Liu, Ph.D.
Associate Professor of Biostatistics

Second Reader

________________________________________________________________________

Josée Dupuis, Ph.D.
Professor of Biostatistics

Third Reader

________________________________________________________________________

Kathryn L. Lunetta, Ph.D.
Professor of Biostatistics
ACKNOWLEDGMENTS

I would like to acknowledge my parents, Sung- gi Hong and Eul-sun Park, for their encouragement and support throughout my life journey. I thank my grandmother, Kyung-suk Lee, for her unlimited love. I also thank my brother, Keung-young Hong, for his support and friendship.

It has been a long educational journey towards this degree. First of all, I would like to thank my advisor Dr. Ching-Ti Liu for his guidance and encouragement in my dissertation research. This dissertation would not be completed without his expertise and guidance. I would like to thank my committee members, Dr. Josée Dupuis, Dr. Kathryn Lunetta, Dr. L Adrienne Cupples, and Michael Lavalley, for sharing their insightful discussions and statistical expertise.

I would like to thank my lovely wife, Lani Hong, for her devoted support and love. I thank my children, Elizabeth, Noah and Emily Hong, for their bottomless healthy energy and inspiration.
META-ANALYSIS STRATEGIES FOR HETEROGENEOUS STUDIES IN GENOME-WIDE ASSOCIATION STUDIES

JAELYOUNG HONG

Boston University Graduate School of Arts and Sciences, 2016

Major Professor: Ching-Ti Liu, Associate Professor of Biostatistics

ABSTRACT

Meta-analysis is a statistical technique that combines results from multiple independent studies to make inferences about parameters of interest. Although it is popular for parameter estimation and hypothesis testing, meta-analytic approaches that incorporate heterogeneous studies have not been fully developed. For heterogeneous studies, we do not expect all of the studies to have the same true underlying effect and the use of the fixed-effects model in a meta-analysis in this situation violates the assumption of homogeneity of effect size. Heterogeneity among studies can arise from multiple sources such as differences in populations by ancestry, differences in study designs, and different impacts of environmental exposures on the effect of the variable of interest. In this thesis, we introduce an analytic strategy and statistical models for meta-analysis of potentially heterogeneous studies. First, we propose a two-stage clustering approach to account for heterogeneity in trans-ethnic meta-analysis of genome-wide association studies (GWAS). Specifically, we cluster studies in the two-stage approach using cohort-specific genetic
information prior to meta-analysis to account for between-cluster heterogeneity as well as to bolster within-cluster homogeneity. An extensive simulation study shows that this approach improves power and diminishes computational intensity compared to existing methods for trans-ethnic meta-analysis. Next, under a meta-regression framework, we develop a likelihood ratio test (LRT) statistic to accommodate multiple random effects. We allow multiple sources of heterogeneity in terms of study characteristics and model the heterogeneities as random effects. We show that the proposed LRT maintains a similar or higher power than other existing methods in a simulation study especially when heterogeneity exists. We apply this new approach to meta-analyze genome-wide association data. Lastly, we derive a score test in the same context as our proposed new LRT and show the substantial advantage of the score test in computational efficiency compared to the new LRT. The introduced strategy and methodologies can effectively and efficiently aggregate the evidence from potentially heterogeneous studies in statistical genetics and other research areas.
# TABLE OF CONTENTS

ACKNOWLEDGMENTS ........................................................................................................ iv

ABSTRACT .......................................................................................................................... v

TABLE OF CONTENTS ....................................................................................................... vii

LIST OF TABLES ................................................................................................................ xi

LIST OF FIGURES ............................................................................................................. xiv

LIST OF ABBREVIATIONS ................................................................................................. xvii

Chapter 1  Introduction ....................................................................................................... 1
  1.1 Genetic Association Studies ....................................................................................... 1
  1.2 Meta-Analysis ........................................................................................................... 3
  1.3 Meta-Analysis of Potentially Heterogeneous Studies .............................................. 5
    1.3.1 Trans-ethnic Heterogeneity .............................................................................. 8
    1.3.2 Covariate Heterogeneity ............................................................................... 9
  1.4 Dissertation Outline ............................................................................................... 10

Chapter 2  Evaluation of a Two-Stage Approach in Trans-ethnic Meta-Analysis in
Genome-Wide Association Studies ..................................................................................... 13
  2.1 Introduction ............................................................................................................. 13
  2.2 Methods ................................................................................................................ 16
    2.2.1 Two-Stage Clustering Approach .................................................................... 16
3.4.2 Statistical Analysis ........................................................................ 76
3.4.3 Meta-analysis Results ...................................................................... 78
3.5 Discussion ...................................................................................... 84

Chapter 4 Score Test Approach to Statistical Model with Multiple Random Effects to Account for Heterogeneity in the Meta-Analysis ................................................. 88
4.1 Introduction ................................................................................... 88
4.2 Methods ....................................................................................... 90
  4.2.1 Linear Mixed Effects Model with Multiple Random Effects .............. 90
  4.2.2 Asymptotic Distribution .................................................................. 96
4.3 Simulation Study ........................................................................... 97
  4.3.1 Study Design of Simulation ............................................................. 97
  4.3.2 Simulation Results ......................................................................... 101
4.4 Discussion ..................................................................................... 107

Chapter 5 Conclusion and Future Work ..................................................... 110

Appendices .......................................................................................... 113

Appendix A Derivation of Maximum Likelihood Estimators for $\beta$ and $\tau i2$ .......... 113
Appendix B Candidate “Correct” Null Distribution Comparison ....................... 115
Appendix C Type I Error Rates from the Corrected Null Distribution in Two
  Additional Simulation Situations ................................................................... 119
  C.1 Varying Within-Study Variance ...................................................... 119
  C.2 Varying Z distribution ..................................................................... 121
LIST OF TABLES

Table 2.1 Pairwise dissimilarity metric in ten populations using Euclidean distance of average allele frequencies on Chromosome 21 .................................................. 30

Table 2.2 Empirical thresholds for one- and two-stage clustering approaches for FE, RE, RE-HE, BE and MANTRA methods when sample sizes are equal.............................. 31

Table 2.3 Empirical thresholds at the multiple testing adjusted threshold $\alpha=0.1$ for both one-stage and two-stage clustering approaches for FE, RE, RE-HE, BE and MANTRA methods for region-wise analyses including causal variants...................... 31

Table 2.4 Empirical thresholds at the multiple testing adjusted threshold $\alpha=0.1$ for both one- and two-stage clustering approaches for FE, RE, RE-HE, BE and MANTRA methods for region-based analyses excluding causal variants........................................ 32

Table 2.5 Type I error rates of one- and two-stage clustering approaches for FE, RE, RE-HE, BE and MANTRA methods when sample size was equal ..................................... 40

Table 2.6 Type I error rates of both one- and two-stage clustering approaches for FE, RE, RE-HE, BE and MANTRA methods when sample sizes were unequal....................... 41

Table 2.7 Empirical thresholds at the multiple testing adjusted threshold $\alpha=0.1$ for both one-stage and two-stage clustering approaches for FE, RE, RE-HE, BE and MANTRA methods when sample size was unequal............................................... 42

Table 2.8 Average computational time (in minutes) for MANTRA using the one-stage (10 Clusters) and three two-stage clustering (2 Clusters, 3 Clusters, 5 Clusters) approaches when sample size was equal ................................................................. 51
Table 3.1 Type I error rates of our approach at different α levels using the asymptotic distribution with fifty studies and the corrected distribution................................. 71

Table 3.2 Type I error rates of FE, RE, RE-HE and FE-meta-regression at α=0.01 ....... 72

Table 3.3 Variants associated with Fasting Insulin adjusting for mean age and mean BMI at p-value less than 1 × 10⁻⁶ of either our new LRT or FE models ..................... 81

Table 4.1 Type I error rates evaluated from a χ² distribution................................. 99

Table 4.2 Average computational time (in minutes) for the calculation of the LRT and the score test statistics for 10,000 simulations......................................................... 109

Table B.1 Type I error comparison for candidate “corrected” null distribution for N=10 .................................................................................................................................................................................. 115

Table B.2 Type I error comparison for candidate “corrected” null distribution for N=20 .................................................................................................................................................................................. 116

Table B.3 Type I error comparison for candidate “corrected” null distribution for N=50 .................................................................................................................................................................................. 117

Table B.4 Type I error comparison for candidate “corrected” null distribution for N=100 .................................................................................................................................................................................. 118

Table C.1 Type I error rates varying within-study variance for N=50 ...................... 120

Table C.2 Type I error rates varying Z distribution when covariates are continuous .... 121

Table C.3 Type I error rates varying Z distribution when covariates are dichotomous . 122

Table F.1 Type I error rates for univariate score test.................................................. 130

Table F.2 Average of simulated information matrix and computed covariance matrix using score functions................................................................. 131
Table F.3 Score test statistics (Ts) computed including and excluding the influential points.............................. 132
LIST OF FIGURES

Figure 2.1 Dendrogram showing the dissimilarity among ten populations from the International HapMap Project Phase 3 data ................................................................. 33

Figure 2.2 Power comparisons for ten models (two approaches of five meta-analytic methods) for the four simulated scenarios when the sample size is equal (sample size=3,000 for each study and 30,000 in total, EAF=5% and α=1e-4) ...................... 43

Figure 2.3 Power comparisons for ten models (two approaches of five meta-analytic methods) for the African-Specific Effects and Effect-size Heterogeneity scenarios for unequal vs. equal sample sizes (EAF=10% and α=1e-4) ........................................ 44

Figure 2.4 Power comparisons for ten models (two approaches of five meta-analytic methods) for the four simulated scenarios when a causal variant is included in a region (sample size=3,000 for each study and 30,000 in total, EAF=10% and multiple testing adjusted threshold α=0.1) ............................................................................ 45

Figure 2.5 Power comparisons for ten models (two approaches of five meta-analytic methods) for the four simulated scenarios when a causal variant is not included in a region (sample size=3,000 for each study and 30,000 in total, EAF=10% and multiple testing adjusted threshold α=0.1) ............................................................................ 46

Figure 2.6 Sensitivity analysis of the two-stage approach by varying the threshold of dissimilarity resulting in a different number of clusters in addition to the one-stage (10 studies) and the initial two-stage approach (3 Clusters) ............................................. 52

Figure 3.1 Power results of Scenario1: Effect-size Homogeneity scenario where random effects are zero ............................................................... 73
Figure 3.2 Power results of Scenario 2: Effect-size Heterogeneity scenario with two non-zero random effects

Figure 3.3 Power results of Scenario 3: Effect-size Heterogeneity scenario with one random effect where only one random effect is present and non-zero

Figure 3.4 Quantile-Quantile (Q-Q) plot with p-values from the empirical distribution.

Figure 3.5 Genome-wide association plot from the fifteen MAGIC GWAS results using our new LRT method.

Figure 4.1 Empirical type I error rates and its 95% confidence intervals for the four different combinations of the study covariates at $\alpha = 0.001$.

Figure 4.2 Power of the score test and LRT tests under the Effect-size Homogeneity scenario (Scenario 1).

Figure 4.3 Power of score test and LRT tests under the Two Random Effects scenario (Scenario 2).

Figure 4.4 Power of score test and LRT tests under the One Random Effects scenario (Scenario 3).

Figure F.1 Scatter plot of score statistics vs. LRT.

Figure F.2a Forest plots for extreme and non-extreme score statistics; Extreme $T_s = 64$ and Non-Extreme $T_s = 12$.

Figure F.2b Forest plots for extreme and non-extreme score statistics; Extreme $T_s = 54$ and Non-Extreme $T_s = 11$.

Figure F.2c Forest plots for extreme and non-extreme score statistics; Extreme $T_s = 52$ and Non-Extreme $T_s = 11$. 
Figure F.3 Histogram and Quantile-Quantile (Q-Q) plots for score test p-values adding small variances................................................................. 138

Figure F.4 Histogram and Quantile-Quantile (Q-Q) plots for score test p-values calculated from the empirical null distribution............................................ 139
**LIST OF ABBREVIATIONS**

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>BE</td>
<td>Binary effects</td>
</tr>
<tr>
<td>BMI</td>
<td>Body Mass Index</td>
</tr>
<tr>
<td>FE</td>
<td>Fixed-effects</td>
</tr>
<tr>
<td>GWAS</td>
<td>Genome-wide association study</td>
</tr>
<tr>
<td>LD</td>
<td>Linkage Disequilibrium</td>
</tr>
<tr>
<td>LRT</td>
<td>Likelihood Ratio Test</td>
</tr>
<tr>
<td>MAF</td>
<td>Minor allele frequency</td>
</tr>
<tr>
<td>MAGIC</td>
<td>Meta-Analysis of Glucose- and Insulin-related traits Consortium</td>
</tr>
<tr>
<td>MANTRA</td>
<td>Meta-analysis of Tranethnic Association study</td>
</tr>
<tr>
<td>metaReg</td>
<td>Meta-regression</td>
</tr>
<tr>
<td>RE</td>
<td>Random-effects</td>
</tr>
<tr>
<td>RE-HE</td>
<td>Random-effects by Han and Eskin</td>
</tr>
<tr>
<td>SNP</td>
<td>Single Nucleotide Polymorphism</td>
</tr>
</tbody>
</table>
Chapter 1  Introduction

1.1 Genetic Association Studies

In genetic association studies, investigators are interested in testing the association of genetic determinants with complex human disease or disease-related traits. Usually, biallelic common variants, defined as genetic variants with minor allele frequency (MAF) greater than 5% or 1%, are examined in genome-wide association studies (GWAS). Depending on whether the trait is quantitative or qualitative, linear or logistic regression models are used to test the genetic association adjusting for study covariates. For correlated data, mixed effects or generalized estimating equation models may be used.

In recent years, large numbers of genetic variants have been associated with complex diseases and disease-related traits using GWAS and candidate gene studies. Despite this success, the effect sizes of identified genetic variants are small, and they only explain a small proportion of the heritability of complex disease or disease-related traits. Thus, a large number of samples are needed and an analysis in any single study is usually underpowered to detect weaker signals. The variants with weaker associations can be identified with increased power by increased sample size (McCarthy et al. 2008; Manolio et al. 2009). Therefore, investigators combine summary statistics of the individual studies and meta-analyze them to detect those weaker signals with better power. This meta-analysis approach increases power by boosting sample size and has been shown to be as efficient as combining individual participant data under most conditions (Lin and Zeng 2010).
Thus, the meta-analysis approach is an effective and more practical way to explore those weaker signals.

Although meta-analysis has been widely used for GWAS, it has often been restricted to a single ancestral population. Because genetic architectures are different across diverse populations, investigators do not want to violate the homogeneity of effects assumption of the fixed effects (FE) meta-analysis model. However, properly meta-analyzing over diverse population studies would further increase sample size and could lead to higher power to detect novel genetic loci (Y. Li and Keating 2014). Also, leveraging linkage disequilibrium (LD) information across difference ancestral populations could enhance the ability of fine-mapping to detect functional genetic variants (Liu et al. 2012; Liu et al. 2014; Boerwinkle and Heckbert 2014; Franceschini et al. 2012; Y. Li and Keating 2014).

Most complex traits are influenced by not only the genetic determinants but also environmental factors. Epidemiology studies show the association of environmental variables with complex traits. Thus, it is reasonable to explore the interaction of genetic determinants and environmental variables to find additional genetic contributors and to explain some of the unexplained heritability. Another potential source of the unexplained heritability may be due to rare variants, defined by variants with MAF less than 1%. Some studies suggest that genes with common variants with modest effect sizes on complex disease may have rare variants with larger effect size (Manolio et al. 2009).
With decreasing cost and rapid advances of sequencing technology, rare variant analysis is now available and provides opportunity in explaining the “missing heritability”.

1.2 Meta-Analysis

Meta-analysis is used to combine results from multiple independent studies to make inferences about a parameter of interest. In general, meta-analysis methods apply p-value based approaches and effect-size based approaches. Among the p-value based approaches, Fisher’s method (Fisher 1925) combines p-values from all participating studies into a pooled test statistic as

\[ \chi^2_{2K} = -2 \sum_{i=1}^{K} \ln(p_i), \]

where \( K \) is a number of studies and \( p_i \) is the p-value for testing the null hypothesis of no effect in the \( i^{th} \) study. Under the null hypothesis of no effect, the test statistic follows a Chi-square distribution with \( 2K \) degrees of freedom. Another approach is the inverse normal method, independently suggested by Stoufer et al. (Stouffer et al. 1949) and Liptak (Liptak 1958). It converts p-values using a probit function to Z-scores and combines the transformed Z-scores into a pooled test statistic. The probit function is the inverse of the standard normal cumulative distribution function and converts uniformly distributed values to normally distributed values. The test statistic is

\[ Z_1 = \sum_{i=1}^{K} \frac{Z_i}{\sqrt{K}}, \]
where $z_i = \Phi^{-1}\left(\frac{p_i}{2}\right)$ and $\Phi^{-1}$ is the probit function. The test statistic, $Z_1$, follows a standard normal distribution under the null hypothesis of no effect. Two weighted inverse normal methods are proposed by Mosteller et al. (Mosteller, Bush, and Green 1954) and Liptak (Liptak 1958). Mosteller’s weighted inverse normal method involves a nonnegative weights to the test statistic. However, the Mosteller’s method as well as the two aforementioned methods do not take into account the sign of the effects. Liptak’s weighted inverse normal method (implemented in METAL software (Willer, Li, and Abecasis 2010)) incorporates the sign of the effect and accounts for study sample sizes. The test statistic is

$$Z_2 = \frac{\sum_{i=1}^{K} w_i' z_i'}{\sqrt{\sum_{i=1}^{K} w_i'^2}},$$

where $w_i = \sqrt{n_i}, n_i$ is a sample size for the $i^{th}$ study, and $z_i' = \text{sign} (\hat{\beta}_i) \Phi^{-1}\left(\frac{p_i}{2}\right), \hat{\beta}_i$ is an effect estimate for the $i^{th}$ study. The test statistic follows a standard normal distribution under the null hypothesis of no effect.

However, the major disadvantage of all of the p-value based approaches is that they do not provide a pooled effect size estimate. This disadvantage can be overcome by combining the effect sizes of the participating studies instead of the p-values. The inverse-variance-weighted fixed-effects (FE) model is commonly used to combine the effect size estimates of the studies assuming the effect size estimates are the same across
the individual studies estimating the same underlying effect (Fleiss 1993; Hedges and Olkin 1985). The inverse-variance-weighted pooled effect estimator is

\[ \hat{\beta}_{FE} = \frac{\sum_{i=1}^{N} w_i \hat{\beta}_i}{\sum_{i=1}^{N} w_i}, \]

where the weight \( w_i = \frac{1}{s_i^2} \), and \( \hat{\beta}_i \) and \( s_i \) are the effect estimate and the standard error of the effect estimate for the \( i^{th} \) study. The standard error of \( \hat{\beta}_{FE} \) is \( \text{SE}(\hat{\beta}_{FE}) = \frac{1}{\sqrt{\sum_{i=1}^{N} w_i}} \). Hence the test statistic of FE model is

\[ Z_{FE} = \frac{\hat{\beta}_{FE}}{\text{SE}(\hat{\beta}_{FE})} = \frac{\sum_{i=1}^{N} w_i \hat{\beta}_i}{\sqrt{\sum_{i=1}^{N} w_i}}. \]

Under the null hypothesis of no effect, the test statistic of the FE model, \( Z_{FE} \), follows a standard normal distribution. The FE model assumes a common true underlying effect parameter for all studies.

1.3 Meta-Analysis of Potentially Heterogeneous Studies

The FE meta-analysis model depends on the homogeneity assumption that the true effect for all studies is the same. Thus, when the true effects across studies differ, the FE model is not appropriate. It violates the assumption of homogeneity of effect size, ignores the between-study variance, and so is likely to lead to confidence intervals of the pooled effect size estimate with lower coverage probability than expected.
We can formally test for effect size homogeneity using Cochran’s Q statistic (Cochran 1952). The Q statistic is calculated by summing the squared differences of each study’s estimated effect from the pooled effect size, weighted by its inverse variance,

\[ Q = \sum_{i=1}^{K} \frac{(\hat{\beta}_i - \hat{\beta}_{FE})^2}{Var(\hat{\beta}_i)}. \]

Under the null hypothesis of homogeneity among the effect sizes, the Q statistic follows a Chi-square distribution with \(K-1\) degrees of freedom where \(K\) indicates the number of participating studies in the meta-analysis. If the null hypothesis is not rejected, the FE model is adopted in the meta-analysis while rejecting the null hypothesis indicates that the FE model is inappropriate to use.

The Q statistic has low power when the number of studies is small. Further, it also has excessive power when the number of studies is large, meaning that clinically unimportant differences in effect across studies may produce a significant Q statistic. (Hardy and Thompson 1998). Also, the Q statistic only detects the presence or absence of heterogeneity among the effects, but does not provide a measure of the amount of heterogeneity. Thus, we can use the \(I^2\) statistic, which can be computed from the Q statistic,

\[ I^2 = \frac{Q - (K - 1)}{Q} \times 100\%, \]

where \(K\) indicates the number of studies in the meta-analysis and the interpretation is the percentage of total variation that is due to heterogeneity rather than sampling error.
(Higgins and Thompson 2002). The values of $I^2$ range from 0% to 100% where 0% indicates all variability in the effect sizes is due to sampling error and none is due to heterogeneity. The $I^2$ statistic is useful especially when the Q statistic does not reject the null hypothesis of homogeneity due to the small number of studies, the $I^2$ statistic further quantifies the impact of the heterogeneity and provides more information on how much of the total variability is due to heterogeneity in the effect sizes. If the Q statistic reject the null hypothesis $I^2$ still quantifies the impact of heterogeneity on the total variability. Higgins and Thompson (2002) suggest that a value of $I^2$ greater than 50% might indicate substantial heterogeneity across the effect sizes. $I^2$ is closely related to the between-study variance, $\tau^2$, such that as $\tau^2$ increases the value of $I^2$ increases. But the advantage of $I^2$ over $\tau^2$ is that $I^2$ values can be directly compared between meta-analyses regardless of the number of studies and effect size metric.

In the presence of heterogeneity, a random-effects (RE) model, instead of a FE model, can be used. The RE model accounts for heterogeneity in the effect sizes across the individual studies. The RE model correctly estimates the between-study variance and provides more appropriate confidence intervals for the pooled effect size estimate (Fleiss 1993; Hedges and Olkin 1985). The inverse-variance-weighted pooled effect estimator is

$$\hat{\beta}_{RE} = \frac{\sum_{i=1}^{N} w_i^* \hat{\beta}_i}{\sum_{i=1}^{N} w_i^*},$$
where the weight $w_i^* = \frac{1}{s_i^2 + \hat{t}^2}$ and the between-study variance, $\hat{t}^2$, can be estimated using DerSimonian and Laird’s method of moments (DerSimonian and Laird 1986). The standard error of $\hat{\beta}_{RE}$ is $\text{SE}(\hat{\beta}_{RE}) = \frac{1}{\sqrt{\sum_{i=1}^{N} w_i^*}}$. The test statistic of RE model is

$$Z_{RE} = \frac{\hat{\beta}_{RE}}{\text{SE}(\hat{\beta}_{RE})} = \frac{\sum_{i=1}^{N} w_i^* \hat{\beta}_i}{\sqrt{\sum_{i=1}^{N} w_i^*}}.$$ 

### 1.3.1 Trans-ethnic Heterogeneity

Until recently, most GWAS meta-analyses were conducted among studies with similar ancestry populations rather than incorporating studies from multiple ethnic populations. However, trans-ethnic meta-analysis or meta-analysis of GWAS from diverse ancestry populations is becoming increasingly popular. Although most GWAS originally focused on European-ancestry populations, the increased number of GWAS in other ethnic populations enables trans-ethnic meta-analysis. Combining GWAS from ethnically-diverse populations through meta-analysis offers many advantages, such as increased statistical power through increased total sample size, prioritizes loci that are associated with multiple populations over loci that are population-specific for secondary replication, and fine-maps underlying functional variants (Y. Li and Keating 2014).
With these opportunities and advantages, the trans-ethnic meta-analysis also presents challenges. Because genetic architectures are not the same across diverse populations, it is not clearly understood how to leverage the ancestral genetic differences. Inconsistent LD patterns along with different allele frequency spectra can result in different sets of associated markers for complex traits and limit replicability across populations. It is not clear how to optimize power for novel discovery versus fine-mapping when some loci are shared across populations and some are population-specific. Therefore, it is critical to properly incorporate heterogeneity that arises from the ancestral genetic differences in trans-ethnic studies.

1.3.2 Covariate Heterogeneity

In addition to the heterogeneity due to the genetic architecture differences in diverse populations, environmental exposures can further introduce variability in effect sizes across studies. Hence, it is important to investigate potential sources of heterogeneity when performing meta-analysis. Variability of environmental factors such as age, sex, body mass index (BMI), drug dosage, or duration of a certain exposure across studies can alter or confound treatment effects on outcomes. Hence, in clinical trials, meta-analysis can be used to examine the effect of those factors beyond the treatment effects and how those factors alter treatment effectiveness (Colditz et al. 1994). In statistical genetics, many efforts have been made to understand heterogeneity that arises from environmental factors other than genetic structure as well as the interplay of gene and environment
factors (Lebrec, Stijnen, and van Houwelingen 2010; Manning et al. 2011; X Wang et al. 2009; Kraft et al. 2007; Herbert et al. 2006).

When investigating the effects of environmental factors on the main effect of interest, the analysis approaches depend on the availability of data. Meta-regression is a useful tool when only study-level summary data are available. This regression-based method uses the estimated effect sizes from each study as a dependent variable and one or more study-level covariate measures as independent variables. Alternatively, meta-analysis of interaction-term estimates or a joint test of the main and interaction effects can be performed (Kraft et al. 2007; Manning et al. 2011). Finally, when full individual participant data is available we can model the pooled data including the interaction term using a mixed model. Although the individual participant data analysis allows an extended scope of analysis, if feasible, the collection of all data on all the participant individuals may take considerable time, resources and permissions to share such data.

1.4 Dissertations Outline

This dissertation includes three projects related to meta-analysis of potentially heterogeneous studies. Each project involves an evaluation of existing approaches and/or a novel approach, extensive simulation studies and comparisons to existing methods with a goal of properly dealing with heterogeneity that arises from genetic structure or other study characteristic differences across studies.
In chapter 2, we evaluate a two-stage approach for trans-ethnic meta-analysis in GWAS. We perform extensive simulation studies to compare five existing methods using one- and two-stage clustering approaches. We aim to show that the two-stage approach can be an effective and efficient approach using existing meta-analytic models when synthesizing GWAS from multiple populations with different ancestries.

In chapter 3, we develop a novel methodology as an extension to Han and Eskin’s approach (B. Han and Eskin 2011) to accommodate multiple sources of heterogeneity in meta-analysis. We propose a new meta-analytic model that can be used in the context of meta-analysis of heterogeneous studies. We assume that there are multiple sources of heterogeneity directly related to study-level characteristics and therefore incorporate study-level covariates as random effects using a linear mixed effects model. A Log-Likelihood Ratio Test (LRT) is compared to a corrected null distribution to evaluate p-values.

In chapter 4, we propose a Score Test as an alternative to the LRT proposed in chapter 3. We construct an empirical distribution to evaluate p-values.

In chapter 5, we conclude our findings and discuss future research directions.
In summary, we explore and develop statistical approaches to accommodate potential heterogeneity across studies to improve power and computational efficiency in meta-analysis of heterogeneous studies of GWAS.
Chapter 2  Evaluation of a Two-Stage Approach in Trans-ethnic Meta-Analysis in Genome-Wide Association Studies

2.1  Introduction

GWAS have identified many genetic loci contributing to the variation of complex human diseases (Welter et al. 2014; Mahajan et al. 2014; Morris et al. 2012; L. A. Hindorff et al. 2009; L. Hindorff et al.). Despite this success, the effect sizes of these loci are usually modest so that a large number of samples are needed and the analysis in any single study is likely under powered (McCarthy et al. 2008; Manolio et al. 2009). Moreover, the identified genetic variants from GWAS may not be the true causal variants but may only be in LD with nearby causal variants (Morris 2011).

Meta-analysis is an effective approach to combine multiple GWAS and increase power by boosting sample sizes. Such meta-analysis has typically been conducted under a fixed-effects model among studies with similar ancestry to avoid violation of the assumption of between-study effect homogeneity (Cantor, Lange, and Sinsheimer 2010; de Bakker et al. 2008; Zeggini and Ioannidis 2009). Because meta-analysis has become more popular and feasible through the use of comprehensive reference panels such as the International HapMap Project (Altshuler et al. 2010) and the 1000 Genome Project (Abecasis et al. 2010) permitting good quality genotype imputations, the number of GWAS meta-analyses, restricted to a single ancestry group and across multiple different populations, has increased. However, it remains unclear how to better leverage the allelic heterogeneity and different LD structure across populations of different ancestries since
the differences in the LD structure can cause effect heterogeneity. Properly analyzing association results from different ancestry samples can increase the sample size and hence may lead to higher power to identify novel loci that are similar across populations (Y. Li and Keating 2014). In addition, meta-analyses including populations of different genetic backgrounds may enhance the ability to fine-map causal signals by leveraging the differences in LD across ancestral groups (Franceschini et al. 2012; Liu et al. 2012; Boerwinkle and Heckbert 2014; Y. Li and Keating 2014; Liu et al. 2014). Therefore, while challenging, it is worth exploring approaches for meta-analysis of multiple GWAS across different ethnic ancestral populations.

Conventional meta-analysis methods such as FE and RE models have been used in GWAS. These methods were developed in the clinical trial setting (Fleiss 1993; Hedges and Olkin 1985). The FE model assumes that the effect sizes are consistent across individual studies. If the underlying effects are not equal, the FE model does not properly incorporate heterogeneity among studies and underestimates the standard error of the combined effect estimates. RE models account for heterogeneity, correctly estimate the standard errors, and provide more appropriate confidence intervals for the effect estimates when effect size heterogeneity exists. Although these two conventional methods are useful, they may not always be suitable for GWAS. The FE model depends on a rather strong assumption of equal effect sizes across contributing studies and hence it is not appropriate when meta-analyzing GWAS over populations of different ancestry. The RE model is designed to handle heterogeneity; however, it relies on a conservative
assumption that effect sizes are different even under the null hypothesis of no association, thereby making the procedure more conservative in detecting genetic association.

In addition to the two conventional methods, three methods were recently proposed specifically for meta-analysis of GWAS in the presence of heterogeneity of effect sizes across independent studies. To cope with low power and the conservativeness of the conventional RE model, Han and Eskin developed an alternative random-effects (RE-HE) model that assumes no heterogeneity under the null hypothesis of no association (B. Han and Eskin 2011). Han and Eskin proposed a second method called binary-effects (BE) model (B. Han and Eskin 2012). It is called binary-effects model not because it takes a binary outcome but because we assume that the true effect is either present or absent in a study using the BE model. This method is based on a new statistic called the \( m \)-value, which is the posterior probability of the non-zero effect size of a study. The \( m \)-values are calculated using all studies assuming the effect sizes are similar across studies if their effects are non-zero. The third method, Meta-analysis of Tranethnic Association studies (MANTRA), uses a Bayesian framework to account for heterogeneity (Morris 2011).

In this chapter, we propose a two-stage clustering strategy in the context of trans-ethnic meta-analysis of GWAS. This approach utilizes prior knowledge of genetic dissimilarity across studies in terms of their mean allele frequency. Our goal is to explore the performance of the two-stage clustering approach in various simulation scenarios using
the five methods (RE, FE, RE-HE, BE and MANTRA), to suggest the best strategy in the meta-analysis of potentially heterogeneous studies, and to examine whether the two-stage approach offers any advantages over a one-stage approach.

2.2 Methods

2.2.1 Two-Stage Clustering Approach

We apply five existing meta-analytic methods to trans-ethnic meta-analysis in a two-stage clustering approach and compare to a one-stage (i.e., no clustering) approach. We assume that there are $N$ independent studies. In the one-stage approach, we simply use the existing methods to meta-analyze over the $N$ independent GWAS. In the two-stage approach, we utilize known study-specific information and group them into $N^*$ subgroups ($N^* \leq N$) according to the (dis)similarity of genetic characteristics; for example, ethnically closely-related studies are grouped into the same cluster and distantly-related studies are separated into different clusters.

Any measure can be applied to evaluate the dissimilarity of the participating cohorts. We specifically evaluate the dissimilarity across the participating cohorts by calculating pairwise Euclidean distances of their average allele frequencies. The dissimilarity between samples from two cohorts is evaluated by the square root of the sum of squares of the average allele frequency differences of variants available in both cohorts. Initially, each cohort is assigned to its own and proceeds iteratively through a hierarchical
clustering algorithm. At each stage, it joins two closest clusters and continues until there is just a single cluster. We assume that the studies within each cluster have homogeneous effect sizes and we use an FE model among studies within each cluster. We implement the methods discussed below to combine FE-derived meta-analysis results across the different clusters in a two-stage approach. In practice, it is reasonable to use ethnic clustering such that studies from the same ethnic ancestry are allocated to the same cluster. We use clustering methods to identify those samples from similar genetic backgrounds.

2.2.2 Fixed-effects model

The FE model assumes a common true association effect parameter $\beta$ across $N$ independent studies and provides a test statistic, testing $H_0: \beta = 0$ vs. $H_1: \beta \neq 0$, using an inverse-variance-weighted pooled effect estimator and its standard error. Let $b_i$ be the effect estimate for the $i^{th}$ study assumed to follow a normal distribution with the true effect size $\beta$ and variance $s_i^2$. Suppose the variance $s_i^2$ of each study is estimated and treated as the true variance for study $i$. The inverse-variance-weighted pooled effect estimator is

$$\hat{\beta} = \frac{\sum_{i=1}^{N} w_i b_i}{\sum_{i=1}^{N} w_i},$$
where the weight $w_i = \frac{1}{s_i^2}$. The standard error of $\hat{\beta}$ is $SE(\hat{\beta}) = \frac{1}{\sqrt{\sum_{i=1}^{N} w_i}}$. Hence the test statistic of FE model is

$$Z_{FE} = \frac{\hat{\beta}}{SE(\hat{\beta})} = \frac{\sum_{i=1}^{N} w_i b_i}{\sqrt{\sum_{i=1}^{N} w_i}}.$$

Under the null hypothesis of no association, the test statistic follows a standard normal distribution (Fleiss 1993; Hedges and Olkin 1985).

### 2.2.3 Random-effects model

The RE model assumes the true effect parameter $\beta_i$ for each study follows a normal distribution with mean population effect parameter $\beta$ and between-study variance $\tau^2$. The null and alternative hypotheses are $H_0: \beta = 0, \tau^2 > 0 \text{ vs. } H_1: \beta \neq 0, \tau^2 > 0$. The between-study variance $\tau^2$ can be estimated by the method of moments using the DerSimonian and Laird approach (DerSimonian and Laird 1986). Given the estimated between-study variance $\hat{\tau}^2$, the inverse-variance-weighted pooled effect estimator $\hat{\beta}^*$ and its standard error can be estimated in a similar way to that of the FE model except that the weight includes the between-study variance. The inverse-variance-weighted pooled effect estimator is

$$\hat{\beta}^* = \frac{\sum_{i=1}^{N} w_i^* b_i}{\sum_{i=1}^{N} w_i^*}.$$
where the weight \( w_i^* = \frac{1}{s_i^2 + \tau^2} \). The standard error of \( \hat{\beta}^* \) is \( \text{SE}(\hat{\beta}^*) = \frac{1}{\sqrt{\sum_{i=1}^{N} w_i^*}} \). Hence the test statistic of RE model is

\[
Z_{RE} = \frac{\hat{\beta}^*}{\text{SE}(\hat{\beta}^*)} = \frac{\sum_{i=1}^{N} w_i^* b_i}{\sqrt{\sum_{i=1}^{N} w_i^*}}.
\]

Under the null hypothesis of no association (\( H_0: \beta = 0, \tau^2 > 0 \)), the test statistic follows a standard normal distribution (Fleiss 1993; Hedges and Olkin 1985).

### 2.2.4 Likelihood approach to RE model by Han and Eskin

Han and Eskin propose a likelihood based method that assumes no heterogeneity under the null hypothesis of no association; hence the mean and between-study variance are equal to zero, while allowing for heterogeneity under the alternative. By relaxing the conservative assumption of heterogeneity under the null hypothesis of RE, the power of the RE-HE model is increased substantially compared to the conventional RE model. The likelihoods under the null (\( H_0: \beta = 0 \) and \( \tau^2 = 0 \)) and the alternative (\( H_1: \beta \neq 0 \) and \( \tau^2 \neq 0 \)) are

\[
L_0 = \prod \frac{1}{\sqrt{2\pi s_i^2}} \exp \left( -\frac{b_i^2}{2s_i^2} \right).
\]
\[
L_1 = \prod \frac{1}{\sqrt{2\pi (s_i^2 + \tau^2)}} \exp \left( -\frac{(b_i - \beta)^2}{2(s_i^2 + \tau^2)} \right).
\]

The notation is the same as the notation used in the description of the FE and RE models. The maximum likelihood estimates (MLE) for \(\beta\) and \(\tau^2\) can be obtained by an iterative procedure suggested by Hardy and Thompson (Hardy and Thompson 1996). Given the estimated \(\hat{\beta}\) and \(\hat{\tau}^2\), the likelihood ratio test (LRT) statistic is constructed as follows

\[
T_{REHE} = -2 \log \left( \frac{\sup L_0}{\sup L_1} \right) = \sum \log \left( \frac{s_i^2}{s_i^2 + \hat{\tau}^2} \right) + \sum \frac{b_i^2}{s_i^2} - \sum \frac{(b_i - \hat{\beta})^2}{s_i^2 + \hat{\tau}^2}.
\]

The test statistic \(T_{REHE}\) asymptotically follows an equal mixture of \(\chi_1^2\) and \(\chi_2^2\) distributions when the number of studies is large. The asymptotic p-value is overly conservative when the number of studies is small due to the heavier tail of the asymptotic distribution compared to the true distribution at genome-wide significance level. Hence tabulated values provided by Han and Eskin are used for determining the statistical significance (B. Han and Eskin 2011).

### 2.2.5 Binary Effects model

The binary effects model is a new type of RE model, with newly proposed statistic called m-value and weighted z-score. The z-score \((Z_i)\) is defined by the observed effect size \((b_i)\)
divided by its standard error \((s_i)\) for the \(i^{th}\) study. The m-value, ranging from 0 to 1, can be interpreted as the posterior probability that the effect is non-zero in each study in a meta-analysis. The test statistic of the BE model \((T_{BE})\) is a weighted sum of z-scores, where greater weight is assigned to studies that are expected to have a non-zero effect (by incorporating the m-values). The test statistic is

\[
T_{BE} = \frac{\sum m_i \sqrt{w_i} Z_i}{\sqrt{\sum m_i^2 w_i}},
\]

where the m-value is computed by

\[
m_i = P(T_i = 1|b) = \frac{P(b|T_i = 1)P(T_i = 1)}{P(b|T_i = 0)P(T_i = 0) + P(b|T_i = 1)P(T_i = 1)},
\]

\(b\) is a vector of the observed effect sizes, \(T_i\) is a random variable that has a value 1 if the \(i^{th}\) study has an effect and 0 otherwise. The prior probability of each study having an effect, \(P(T_i = 1)\), assumes a beta prior and the observed effect size, \(b_i\) assumes a \(N(\beta, s_i^2)\). A Markov Chain Monte Carlo (MCMC) method is proposed to estimate the m-value (B. Han and Eskin 2012). In their proposed MCMC method, they start from a random \(T = t\) and choose a next \(t'\). \(g(t)\) is proportional to a posterior probability of \(T\) given the observed effect sizes such that

\[
g(t) = P(b|T = t)P(T = t) \propto P(T = t|b).
\]

If \(g(t) < g(t')\) then move to \(t'\). Otherwise, move to \(t'\) with probability \(g(t')/g(t)\). They repeat this MCMC method for the approximations of the m-values. \(z_i = b_i/s_i\) is the z-score and the weight can be approximated by

\[
\sqrt{w_i} \approx \sqrt{n_i \text{maf}_i(1 - \text{maf}_i)}
\]

where \(n_i\) is the sample size and \(\text{maf}_i\) is
the minor allele frequency (MAF) for the \( i^{th} \) study (Zaitlen and Eskin 2010). If the MAFs are similar among studies the weight \( \sqrt{w_i} \) approximates \( \sqrt{n_i} \) (de Bakker et al. 2008).

2.2.6 Meta-analysis of Transethnic Association studies

Morris developed this method in a Bayesian framework and implemented it in the MANTRA software (Morris 2011). It takes into account the estimated similarity among studies with respect to genetic relatedness using a Bayesian partition model (Denison and Holmes 2001; Knorr-Held and Rasser 2000) and takes genetic distance between studies into account. This approach exploits the notion that similar studies are more likely to have similar effect sizes, while dissimilar studies should, in principle, have more variable effect estimates. The statistical evidence of association is evaluated by the Bayes’ factor (BF) (Kass and Raftery 1995); that is,

\[
\Lambda = \frac{f(b, s|M_1)}{f(b, s|M_0)},
\]

where \( b \) and \( s \) are vectors of the observed allelic effects \( b_i \) and respective standard errors \( s_i \) from all studies and assume to be known values. We assume \( b_i \sim N(\beta_i, s_i) \) where \( \beta_i \) denotes the true allelic effect. Under the null model \((M_0)\), there is no association of a variant with a trait in any population, \( \beta_i = 0 \) for all \( i \), while the alternative model \((M_1)\) corresponds to not all \( \beta_i = 0 \). The marginal likelihood of the observed allelic effects under model \( M \)
\[ f(b, s|M) \propto \int_{\theta} f(b, s|\theta) f(\theta|M) d\theta \]

is given by the integration over the unknown parameters, \( \theta \), including study-specific true allelic effect \( \beta_i \) and additional hyper-parameters relating to their prior distribution, where the likelihood

\[ f(b, s|\theta) = f(b, s|\beta) = \sum_{i=1}^{N} f(b_i, s_i|\beta_i) \]

and

\[ f(b_i, s_i|\beta_i) \propto \frac{1}{s_i} \exp \left[ -\frac{(b_i - \beta_i)^2}{2s_i^2} \right] \]

The marginal likelihood \( f(b, s|M) \) cannot be directly evaluated; however, the joint posterior density can be approximated by a Metropolis-Hastings MCMC algorithm. The hyper-parameters of \( \theta \) include the number of clusters and cluster centers, and the number of clusters are addressed by a birth-death process through a reversible-jump step in the MCMC algorithm. The number of clusters determine the dimensionality of \( \theta \). The MCMC algorithm is run for initial burn-in period to allow convergence from randomly assigned starting values for \( \theta \) (Morris 2011). After convergence, the marginal likelihood
\( f(b, s|M) \) is approximated by the harmonic mean of sampled likelihood values (Newton and Raftery 1994).

### 2.3 Simulation Study Design

We performed a simulation study to compare the type I error and power of the one- and two-stage clustering approaches using the five methods, with the goal of identifying more efficient meta-analysis strategies for trans-ethnic studies.

We used haplotype data from the International HapMap Project Phase 3 (HapMap3) (Altshuler et al. 2010) haplotype data to simulate genotypes for 10 populations. In the HapMap3 data, there were four African-ancestry (AfA) cohorts (African ancestry in Southwest USA (ASW), Luhya in Webuye, Kenya (LWK), Maasai in Kinyawa, Kenya (MKK), Yoruba in Ibadan, Nigeria (YRI)) and two Asian-ancestry (AsA) cohorts (Chinese in Metropolitan Denver (CHD), Japanese in Tokyo and Han Chinese in Beijing (JPT+ CHB)), two European-ancestry (EuA) cohorts (Utah residents with North and Western European ancestry (CEU), Toscani in Italy (TSI)), one Mexican (Mexican ancestry in Los Angeles (MEX)) and one Gujarati Indian (Gujarati Indians in Houston (GIH)). The pairwise dissimilarity metric among all ten populations is presented in the Table 2.1. A smaller distance between two studies indicates closer genetic relatedness. We graphically show the two-stage clustering approach using a dendrogram (Figure 2.1). In Figure 2.1, a threshold at 0.25 of the dissimilarity metric would leave three clusters,
and studies within a cluster are relatively close to each other as measured by genetic distance, while studies in different clusters are further apart. Based on this, we grouped the two EuA cohorts with the MEX and GIH cohorts together as GEM (GIH + EuA + MEX) studies because these four cohorts cluster together as shown below. And we also grouped the African populations (AfA) and the Asian populations (AsA).

We simulated genotypes of variants on chromosome 21 using HAPGEN2 software (Su, Marchini, and Donnelly 2011) which simulated genotypes based on a reference set of haplotypes and an estimate of the recombination rate across a region. Here, we used the HapMap3 data as a reference set for each of the 10 populations. Thus, the simulated data shared LD patterns similar to the HapMap3 reference data. For each simulation replicate, we generated a continuous trait Y for a study-specific association analysis from a normal distribution with mean $\beta$ and standard deviation 1, $\beta \sim N(\beta, 1)$, where the value of $\beta$ depends on each scenario.

Under the null hypothesis of no association, we simulated 1,000 replicates (11,575,000 null single nucleotide polymorphisms (SNPs)) to evaluate Type I error for the one- and two-stage approaches for the FE, RE, RE-HE, BE and MANTRA methods at varying asymptotic thresholds (for $\alpha = 0.05$ through 0.0001). A continuous trait Y was generated from a normal distribution with mean 0 and standard deviation 1. We pooled all the variants from the 1,000 replicates and assess type I error by calculating the proportion of SNPs whose meta-analytic p-value was less than pre-specified asymptotic thresholds. For
the Bayesian method MANTRA, we only presented the proportion of SNPs with BF greater than preselected thresholds because there was no metric to convert BFs to p-values (Xu Wang et al. 2013).

For power analyses, we considered four heterogeneity scenarios: (1) Effect-size homogeneity, (2) Ancestry-specific effects, (3) Effect-size heterogeneity, and (4) Western exposure effects. In the Effect-size homogeneity scenario (scenario 1), all 10 studies have the same effects and therefore there is no heterogeneity across the studies. In the Ancestry-specific scenario (scenario 2), we considered only studies in one ancestry cluster to have non-zero effects. Specifically, we assumed that the four AfA studies have non-zero allelic effects while all others have effect size 0 and we named it as the African-specific effects scenario. In the Effect-size heterogeneity scenario (scenario 3), a SNP has an effect in multiple studies, assuming the same effect size among studies in the same ethnic group but allowing the effect size to vary across different ethnic groups (effect sizes: 0.05 for four AfA studies, 0.025 for two AsA studies and 0.0125 for four GEM studies). In the Western exposure effects scenario (scenario 4), we assumed that shared environments such as life style would result in similar genetic effects and hence we assigned equal, non-zero effect size to the causal variants only in the samples living in Europe or North America (ASW, CHD, CEU and TSI) regardless of their ethnicity. For scenarios 1, 2 and 4, we assessed power varying effect sizes from 0.05 to 0.2, which is roughly equivalent to 0.045% to 0.72% of the trait variation explained by a SNP with MAF 0.1.
We picked six causal SNPs based on their allele frequency as follows. Four SNPs had comparable effect allele frequency (EAF) (0.05, 0.1, 0.2, and 0.4, respectively) across the 10 studies. The remaining two causal SNPs had varying allele frequencies for a selected allele across studies: one SNP with EAF in a range of 0.2 to 0.5 and the other with frequency between 0.2 and 0.8. Because there was no unified threshold to compare Frequentist and Bayesian methods, we assessed empirical thresholds for all five methods for the two clustering approaches under the null hypothesis of no association by pooling the p-values or BFs of the null variants and determining the desired quantile as the empirical thresholds (Table 2.2). We calculated the power for a single variant analysis by the proportion of the simulation replicates with a causal SNP association p-value less than or BF greater than the empirical thresholds for each meta-analysis method and clustering approach.

We also calculated region-wise power using single variant analyses adjusting for multiple testing. In most genetic studies, the causal variants are often not available for analysis. In many cases, identified associated genetic variants are not the causal variants but in LD with those variants. Therefore, it is worthwhile to explore the surrounding region of a causal variant, including or excluding the causal variant, and to evaluate the performance of the two-stage clustering approach. We defined the regions of the six causal SNPs by start and end positions including all SNPs with $r^2>0.2$ with the causal SNPs on Chr21, totaling to 652 SNPs for the six causal SNPs, and assessed power in those regions with
and without the causal SNPs. We defined the regions by $r^2$ because we knew the causal SNPs and were interested in examining how LD with the causal variants would influence power. However, in practice, we may not know the causal variants and selecting a region based on location (e.g., a gene location) would be more practical. We defined statistical significance in the region-wise analysis by comparing the p-values or BF$s$ of all SNPs in a region to an empirical threshold after adjusting for multiple testing. As above, we used empirical thresholds to evaluate power. To correct for multiple comparisons, we used a minimum p-value or a maximum BF of a region from each replicate to construct the empirical distribution. We then select the empirical threshold at the multiple testing adjusted threshold $\alpha = 0.1$ based on this empirical distribution. The multiple testing adjusted threshold $\alpha = 0.1$ may be interpreted as $\alpha$ level divided by the number of SNPs to test in each region. The empirical thresholds are presented in the Tables 2.3 and 2.4.

Once the genotypes and the phenotype were simulated, we performed association analysis using SNPTEST v2 (Marchini and Howie 2010) on common variants (MAF>3%, 11,575 SNPs) for the ten studies and then ran the trans-ethnic meta-analysis using METASOFT and MANTRA software. For the one-stage (no clustering) approach, we ran FE, RE, RE-HE, BE and MANTRA over the ten studies. For the two-stage approach, we first applied FE model within each of three clusters, AfA, AsA and GEM, to get cluster-specific summary results and then ran FE, RE, RE-HE, BE and MANTRA across the three clusters. For the equal sample size scenario, each study contained a sample size of 3,000 and hence 30,000 subjects in total were available for meta-analysis.
For the unequal sample size scenario, we increased sample size in the AfA studies by 1.5 times (4,500) and decreased the sample size in the other ancestry studies by 2/3 (2,000) so that the total sample size (30,000) was the same as the equal sample size situation.
### Table 2.1: Pairwise dissimilarity metric in ten populations using Euclidean distance of average allele frequencies on Chromosome 21

<table>
<thead>
<tr>
<th></th>
<th>ASW</th>
<th>CEU</th>
<th>CHD</th>
<th>GIH</th>
<th>JPT+CHB</th>
<th>LWK</th>
<th>MEX</th>
<th>MKK</th>
<th>TSI</th>
<th>YRI</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASW</td>
<td>0</td>
<td>0.408</td>
<td>0.461</td>
<td>0.379</td>
<td>0.459</td>
<td>0.162</td>
<td>0.389</td>
<td>0.122</td>
<td>0.397</td>
<td>0.174</td>
</tr>
<tr>
<td>CEU</td>
<td>0.408</td>
<td>0</td>
<td>0.358</td>
<td>0.172</td>
<td>0.351</td>
<td>0.491</td>
<td>0.189</td>
<td>0.425</td>
<td>0.080</td>
<td>0.502</td>
</tr>
<tr>
<td>CHD</td>
<td>0.461</td>
<td>0.358</td>
<td>0</td>
<td>0.305</td>
<td>0.074</td>
<td>0.516</td>
<td>0.287</td>
<td>0.470</td>
<td>0.356</td>
<td>0.523</td>
</tr>
<tr>
<td>GIH</td>
<td>0.379</td>
<td>0.172</td>
<td>0.305</td>
<td>0</td>
<td>0.297</td>
<td>0.463</td>
<td>0.179</td>
<td>0.397</td>
<td>0.167</td>
<td>0.475</td>
</tr>
<tr>
<td>JPT+CHB</td>
<td>0.459</td>
<td>0.351</td>
<td>0.074</td>
<td>0.297</td>
<td>0</td>
<td>0.517</td>
<td>0.278</td>
<td>0.370</td>
<td>0.349</td>
<td>0.525</td>
</tr>
<tr>
<td>LWK</td>
<td>0.162</td>
<td>0.491</td>
<td>0.516</td>
<td>0.463</td>
<td>0.517</td>
<td>0</td>
<td>0.471</td>
<td>0.139</td>
<td>0.482</td>
<td>0.095</td>
</tr>
<tr>
<td>MEX</td>
<td>0.389</td>
<td>0.189</td>
<td>0.287</td>
<td>0.179</td>
<td>0.278</td>
<td>0.471</td>
<td>0</td>
<td>0.408</td>
<td>0.189</td>
<td>0.482</td>
</tr>
<tr>
<td>MKK</td>
<td>0.122</td>
<td>0.425</td>
<td>0.470</td>
<td>0.397</td>
<td>0.470</td>
<td>0.139</td>
<td>0.408</td>
<td>0</td>
<td>0.414</td>
<td>0.166</td>
</tr>
<tr>
<td>TSI</td>
<td>0.397</td>
<td>0.080</td>
<td>0.356</td>
<td>0.167</td>
<td>0.349</td>
<td>0.482</td>
<td>0.189</td>
<td>0.414</td>
<td>0</td>
<td>0.493</td>
</tr>
<tr>
<td>YRI</td>
<td>0.174</td>
<td>0.502</td>
<td>0.523</td>
<td>0.475</td>
<td>0.525</td>
<td>0.095</td>
<td>0.482</td>
<td>0.166</td>
<td>0.493</td>
<td>0</td>
</tr>
</tbody>
</table>

The dissimilarity between samples from two cohorts is evaluated by the square root of the sum of squares of the average allele frequency differences of variants available in both cohorts. A smaller distance between two studies indicates closer genetic relatedness in terms of their allele frequencies. Higher values mean greater dis-similarity.

Population samples from: African ancestry in Southwest USA (ASW); Utah residents with North and Western European ancestry (CEU); Chinese in Metropolitan Denver (CHD); Gujarati Indians in Houston (GIH); Japanese in Tokyo and Han Chinese in Beijing (JPT+ CHB); Luhya in Webuye, Kenya (LWK); Mexican ancestry in Los Angeles (MEX); Maasai in Kinyawa, Kenya (MKK); Toscani in Italy (TSI); and Yoruba in Ibadan, Nigeria (YRI).

We used all the common variants on Chromosome 21 (11,575 variants) to calculate the dissimilarity matrix for ten studies. The computation was fast and took less than a second even using a PC (Intel® Core™ i7-2600 Processor, 8M Cache, 3.4 GHz).
Table 2.2 Empirical thresholds for one- and two-stage clustering approaches for FE, RE, RE-HE, BE and MANTRA methods when sample sizes are equal

<table>
<thead>
<tr>
<th>Threshold</th>
<th>One-stage clustering approach</th>
<th>Two-stage clustering approach</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FE</td>
<td>RE</td>
</tr>
<tr>
<td>5e-2</td>
<td>5.0E-02</td>
<td>6.4E-02</td>
</tr>
<tr>
<td>1e-2</td>
<td>1.0E-03</td>
<td>1.4E-02</td>
</tr>
<tr>
<td>1e-3</td>
<td>9.9E-04</td>
<td>1.5E-03</td>
</tr>
<tr>
<td>1e-4</td>
<td>1.0E-04</td>
<td>1.6E-04</td>
</tr>
</tbody>
</table>

The left most column is the asymptotic threshold for the one-stage and the two-stage clustering approaches. Other columns provide empirical thresholds for Fixed-Effects (FE), Random-Effects (RE), Random-Effects model of Han and Eskin (RE-HE), Binary Effects (BE) model and Meta-analysis of Transethnic Association studies (MANTRA) models.

Table 2.3 Empirical thresholds adjusting for the multiple testing for both one-stage and two-stage clustering approaches for FE, RE, RE-HE, BE and MANTRA methods for region-wise analyses including causal variants

<table>
<thead>
<tr>
<th>Causal SNP</th>
<th>One-stage clustering approach</th>
<th>Two-stage clustering approach</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FE</td>
<td>RE</td>
</tr>
<tr>
<td>EAF=0.05</td>
<td>1.30E-03</td>
<td>1.80E-03</td>
</tr>
<tr>
<td>EAF=0.1</td>
<td>2.20E-03</td>
<td>3.10E-03</td>
</tr>
<tr>
<td>EAF=0.2</td>
<td>1.30E-03</td>
<td>1.80E-03</td>
</tr>
<tr>
<td>EAF=0.4</td>
<td>1.80E-03</td>
<td>2.50E-03</td>
</tr>
<tr>
<td>EAF=0.2~0.5</td>
<td>2.80E-03</td>
<td>3.80E-03</td>
</tr>
<tr>
<td>EAF=0.2~0.8</td>
<td>5.20E-03</td>
<td>7.40E-03</td>
</tr>
</tbody>
</table>

The left most column is the effect allele frequency (EAF) or the range of the EAF of the causal SNPs across the ten populations.
Table 2.4 Empirical thresholds adjusting for the multiple testing for both one- and two-stage clustering approaches for FE, RE, RE-HE, BE and MANTRA methods for region-based analyses excluding causal variants

<table>
<thead>
<tr>
<th>Causal SNP</th>
<th>One-stage clustering approach</th>
<th>Two-stage clustering approach</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FE</td>
<td>RE</td>
</tr>
<tr>
<td>EAF=0.05</td>
<td>1.3E-03</td>
<td>1.8E-03</td>
</tr>
<tr>
<td>EAF=0.1</td>
<td>2.2E-03</td>
<td>3.1E-03</td>
</tr>
<tr>
<td>EAF=0.2</td>
<td>1.3E-03</td>
<td>1.8E-03</td>
</tr>
<tr>
<td>EAF=0.4</td>
<td>1.8E-03</td>
<td>2.5E-03</td>
</tr>
<tr>
<td>EAF=0.2~0.5</td>
<td>2.8E-03</td>
<td>3.8E-03</td>
</tr>
<tr>
<td>EAF=0.2~0.8</td>
<td>5.2E-03</td>
<td>7.4E-03</td>
</tr>
</tbody>
</table>

The left most column is the effect allele frequency (EAF) or the range of the EAF of the causal SNPs across the ten populations.
Figure 2.1 Dendrogram showing the dissimilarity among ten populations from the International HapMap Project Phase 3 data.

The distance between populations represents a Euclidean distance calculated using allele frequencies. A smaller distance between populations indicates closer genetic relatedness. The red dotted line indicates an arbitrary but clear fixed threshold at 0.25 in distance that leaves three clusters where studies within a cluster are relatively close to each other in their genetic relatedness.

Population samples are from: African ancestry in Southwest USA (ASW); Utah residents with North and Western European ancestry (CEU); Chinese in Metropolitan Denver (CHD); Gujarati Indians in Houston (GIH); Japanese in Tokyo and Han Chinese in Beijing (JPT+ CHB); Luhya in Webuye, Kenya (LWK); Mexican ancestry in Los Angeles (MEX); Maasai in Kinyawa, Kenya (MKK); Toscani in Italy (TSI); and Yoruba in Ibadan, Nigeria (YRI).
2.4 Simulation Results

2.4.1 Type I error

Table 2.5 shows that the four Frequentist methods (FE, RE, RE-HE, BE) of one- and two-stage approaches had type I error that were less than or equal to the pre-selected $\alpha$ levels. The RE model was consistently conservative. We could not evaluate type I error of MANTRA against the asymptotic thresholds because it uses BFs. Instead, we present the proportion of SNPs whose BF was greater than a few, preselected thresholds for MANTRA. In the unequal sample size scenario, the type I error rates (Table 2.6) and the empirical thresholds (Table 2.7) were similar to the equal sample size situation.

2.4.2 Power

2.4.2.1 Equal Sample Size Scenario

To evaluate power, we considered four scenarios as described above. In all power scenarios, we compared 1) the two-stage approach to the one-stage approach for the five aforementioned methods, 2) the recent methods (RE-HE, BE and MANTRA) to the conventional FE and RE methods in both one- and two-stage approaches, and 3) all of the ten combinations (five methods by the one- and two-clustering approaches) to see which one is the most powerful in each scenario. We used empirical thresholds for power analysis because there was no unified threshold to compare Frequentist and Bayesian methods. The statistical significance for comparison among the approaches and methods was assessed by comparing 95% confidence intervals.
For scenario 1, where there was no heterogeneity in effect sizes across the studies, all meta-analysis methods in both clustering approaches showed similar power. The differences between the corresponding two- and one-stage methods were trivial except in the two-stage conventional RE model (Figure 2.2A). The RE model showed slightly higher power in the one-stage analysis than in the two-stage approach.

In the Ancestry-specific effect size (scenario 2) and the Effect-size heterogeneity (scenario 3) scenarios, we introduced heterogeneity in the effect sizes according to the population ancestry. In the Ancestry-specific effect size scenario (Figure 2.2B), we set the African-ancestry studies to have equal, non-zero effect sizes. We observed a significant power increase in the 2-stage BE model when $\beta=0.15$ compared to its one-stage counterpart as their 95% confidence intervals overlapped. In general, the 2-stage approach of all the other models had either higher or similar power numerically compared to their one-stage counterparts but their 95% confidence intervals did not overlap. In both clustering approaches, the recent meta-analytic methods (RE-HE, BE and MANTRA) performed significantly better than the conventional methods (FE and RE) as their confidence intervals did not overlap. Among the recent methods, the two-stage BE model was significantly more powerful than the one-stage RE-HE and one-stage BE models as their confidence intervals did not overlap and marginally better than the two-stage RE-HE and MANTRA in the one- and two-stage approaches.
In the Effects-size heterogeneity scenario (scenario 3), we varied the size of the effect for the populations in the different ethnic clusters while the effect sizes were equal across studies within a cluster. Therefore, the heterogeneity arose from different magnitude of positive effect sizes across the clusters. In Figure 2.2C, we show power over six allele frequency scenarios. The two-stage approaches showed slightly higher or similar power compared to the same methods in the one-stage approach except for the RE model (Figure 2.2C). The two-stage BE model consistently appeared numerically more powerful than the other models but it was only significantly different from both one- and two-stage RE models. Interestingly, a variant with EAF of 0.4 (observed average EAF over 10 studies was 0.394, observed EAF=0.358~0.427) was more powerful than a variant with EAF=0.2~0.5 (observed average EAF was 0.366, observed EAF=0.206~0.494) in both two-stage and one-stage approaches. Although the variant with EAF=0.4 had slightly higher observed average EAF, a variant that was consistently present across different ethnicities in terms of allele frequency seemed to be better aggregated in a meta-analysis stage compared to a variant with somewhat diverse EAFs across different populations.

In the Western exposure effects scenario (scenario 4), among the recent methods (RE-HE, BE and MANTRA), the two-stage approach was significantly (RE-HE and BE) or nominally (MANTRA) less powerful compared to the one-stage approach (Figure 2.2D). This finding was expected, because we clustered the studies according to the dissimilarity of mean allele frequency and therefore the true source of heterogeneity in the effect size
was not correctly taken into consideration. As a result, studies in each cluster remained heterogeneous and therefore using the FE model over the heterogeneous studies (the two-stage approach) reduced power. In this situation, the one-stage approach was more powerful. Interestingly, we observed the opposite phenomenon in the RE model. The two-stage RE model was more powerful than the one-stage RE model. This contrast suggests that, despite its power loss over heterogeneous studies, the prior use of the FE in the 2-stage RE model was more beneficial than the one-stage RE model.

2.4.2.2 Unequal Sample Size Scenario

We also considered unequal sample size scenarios since that is typical of real genetic study meta-analyses. In the unequal sample size situation, we increased the sample size (as described in the simulation study) in the AfA studies while decreased in the other studies to keep the total sample size the same. We observed improved power as expected due to the increased sample size in the AfA studies because the simulation model included a true effect in the AfA samples only, while there were no simulated effects in samples from other ancestries. We show the unequal size simulation results in the scenarios 2 and 3 in comparison to the equal sample size results (Figure 2.3A and 2.3C).

2.4.2.3 Sensitivity Analysis of The Two-Stage Clustering Approach

We also examined the sensitivity of our two-stage clustering approach in additional simulations by varying the threshold to a different value of the dissimilarity metric in Figure 2.1, resulting in a different number of clusters. Moving the threshold up from 0.25
to 0.4 results in two clusters: four AfA studies in one cluster and the rest of the studies belonging to the other cluster (2 Clusters). We can also move the threshold down such that we may have five clusters from Figure 2.1 where each cluster includes: the first two AfA studies; the other two AfA studies; two AsA studies; MEX and GIH; and two EuA studies (5 Clusters). The pair-wise distance between GIH-TSI or GIH-CEU is smaller than GIH-MEX, however, we have clustered GIH-MEX because metasoft requires at least two studies entering the software. Figure 2.6 showed a one-stage and three two-stage clustering approaches (2 Clusters, 3 Clusters and 5 Clusters). The results suggested that our 2-stage approach is quite robust to the different dissimilarity thresholds where, in fact, the three clustering scenarios yielded the same power. We performed this additional simulation study only with Effect-Heterogeneity (Scenario 3).

2.4.2.4 Region-wise analysis with causal variants included vs. excluded

Because, in most genetic studies, identified associated genetic variants were not causal variants but in LD with those variants, we explored the surrounding region of a causal variant, including or excluding the causal variant, and evaluated the performance of the two-stage clustering approach. The overall pattern of the results was similar to what we had observed except in the Effect-size Heterogeneity scenario (Scenario 3). It was evident that the inclusion of a causal variant in the analysis suggested higher power than when it was excluded (Figures 2.4 and 2.5). Although less deviation in power appeared among the models in the African-specific scenario, the two-stage RE-HE and BE models were slightly more powerful compared to their counterpart one-stage approaches.
(Figures 2.4B, 2.4C, 2.5B and 2.5C). As we previously observed, the two-stage approach had no benefits over the one-stage analysis when the heterogeneity in effect size was due to an environmental factor rather than genetic dissimilarity.
Table 2.5 Type I error rates of one- and two-stage clustering approaches for FE, RE, RE-HE, BE and MANTRA methods when sample size was equal

<table>
<thead>
<tr>
<th>Threshold^a for FE, RE, RE-HE, BE</th>
<th>Threshold^b for MANTRA</th>
<th>One-stage clustering approach</th>
<th>Two-stage clustering approach</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>FE</td>
<td>RE</td>
</tr>
<tr>
<td>5e-2</td>
<td>logBF &gt; 0</td>
<td>5.0E-02</td>
<td>3.9E-02</td>
</tr>
<tr>
<td>1e-2</td>
<td>&gt;1</td>
<td>1.0E-02</td>
<td>7.2E-03</td>
</tr>
<tr>
<td>1e-3</td>
<td>&gt;2</td>
<td>1.0E-03</td>
<td>6.8E-04</td>
</tr>
<tr>
<td>1e-4</td>
<td>&gt;3</td>
<td>1.0E-04</td>
<td>6.5E-05</td>
</tr>
</tbody>
</table>

The thresholds for the Frequentist and Bayesian methods (^a and ^b) are not comparable. The P-values and the Bayes’ factors are assessed under the null hypothesis of no association using 11,575,000 null SNPs.

^a Asymptotic thresholds for the type I error of the one-stage and the two-stage clustering approaches for the Frequentist’s methods (Fixed-Effects (FE), Random-Effects (RE), Random-Effects model by Han and Eskin (RE-HE) and Binary Effects (BE) model).

^b Thresholds for the one-stage and the two-stage clustering approaches for the Bayesian method (Meta-analysis of Transethnic Association studies (MANTRA) method).

^c We presented proportions that SNPs had BF greater than preselected thresholds.
Table 2.6 Type I error rates of both one- and two-stage clustering approaches for FE, RE, RE-HE, BE and MANTRA methods when sample sizes were unequal.

<table>
<thead>
<tr>
<th>Threshold(^a) for FE, RE, RE-HE, BE</th>
<th>Threshold(^b) for MANTRA</th>
<th>FE</th>
<th>RE</th>
<th>RE-HE</th>
<th>BE</th>
<th>MANTRA(^c)</th>
<th>FE 2-stage</th>
<th>RE 2-stage</th>
<th>RE-HE 2-stage</th>
<th>BE 2-stage</th>
<th>MANTRA 2-stage</th>
</tr>
</thead>
<tbody>
<tr>
<td>5e-2</td>
<td>logBF &gt; 0</td>
<td>5.0E-02</td>
<td>3.8E-02</td>
<td>4.9E-02</td>
<td>5.0E-02</td>
<td>1.4E-01</td>
<td>5.0E-02</td>
<td>3.7E-02</td>
<td>4.8E-02</td>
<td>4.9E-02</td>
<td>1.4E-01</td>
</tr>
<tr>
<td>1e-2</td>
<td>&gt;1</td>
<td>1.0E-02</td>
<td>7.2E-03</td>
<td>9.7E-03</td>
<td>1.0E-02</td>
<td>9.1E-03</td>
<td>1.0E-02</td>
<td>6.9E-03</td>
<td>9.5E-03</td>
<td>9.9E-03</td>
<td>9.1E-03</td>
</tr>
<tr>
<td>1e-3</td>
<td>&gt;2</td>
<td>1.1E-03</td>
<td>7.1E-04</td>
<td>9.6E-04</td>
<td>1.0E-03</td>
<td>7.9E-04</td>
<td>1.1E-03</td>
<td>7.2E-04</td>
<td>9.6E-04</td>
<td>1.0E-03</td>
<td>7.9E-04</td>
</tr>
<tr>
<td>1e-4</td>
<td>&gt;3</td>
<td>9.5E-05</td>
<td>6.2E-05</td>
<td>8.4E-05</td>
<td>8.6E-05</td>
<td>6.2E-05</td>
<td>9.5E-05</td>
<td>6.7E-05</td>
<td>9.0E-05</td>
<td>8.4E-05</td>
<td>6.2E-05</td>
</tr>
</tbody>
</table>

The sample size for the four AfA studies is 4,500 and for the other studies is 2,000. The total sample size is 30,000, same as the equal sample size scenarios. The thresholds for the Frequentist and Bayesian methods (\(^a\) and \(^b\)) are not interchangeable. The P-values and the Bayes’ factors are assessed under the null hypothesis of no association using 11,575,000 null SNPs.

\(^a\) Asymptotic thresholds for the type I error of the one-stage and the two-stage clustering approaches for the Frequentist’s methods (Fixed-Effects (FE), Random-Effects (RE), Random-Effects model by Han and Eskin (RE-HE) and Binary Effects (BE) model)

\(^b\) Thresholds for the one-stage and the two-stage clustering approaches for the Bayesian method (Meta-analysis of Transethnic Association studies (MANTRA) method).

\(^c\) We presented proportions of SNPs that had BF greater than preselected thresholds.
Table 2.7 Empirical thresholds at the multiple testing adjusted threshold $\alpha=0.1$ for both one-stage and two-stage clustering approaches for FE, RE, RE-HE, BE and MANTRA methods when sample size was unequal

<table>
<thead>
<tr>
<th>Threshold</th>
<th>FE</th>
<th>RE</th>
<th>RE-HE</th>
<th>BE</th>
<th>MANTAR</th>
<th>FE 2-stage</th>
<th>RE 2-stage</th>
<th>RE-HE 2-stage</th>
<th>BE 2-stage</th>
<th>MANTRA 2-stage</th>
</tr>
</thead>
<tbody>
<tr>
<td>5e-2</td>
<td>5.0E-02</td>
<td>6.4E-02</td>
<td>5.1E-02</td>
<td>5.0E-02</td>
<td>0.35</td>
<td>5.0E-02</td>
<td>6.7E-02</td>
<td>5.2E-02</td>
<td>5.1E-02</td>
<td>0.35</td>
</tr>
<tr>
<td>1e-2</td>
<td>1.0E-02</td>
<td>1.4E-02</td>
<td>1.0E-02</td>
<td>1.0E-02</td>
<td>0.96</td>
<td>1.0E-02</td>
<td>1.4E-02</td>
<td>1.1E-02</td>
<td>1.0E-02</td>
<td>0.96</td>
</tr>
<tr>
<td>1e-3</td>
<td>9.4E-04</td>
<td>1.4E-03</td>
<td>1.0E-03</td>
<td>9.8E-04</td>
<td>1.9</td>
<td>9.4E-04</td>
<td>1.4E-03</td>
<td>1.0E-03</td>
<td>9.7E-04</td>
<td>1.9</td>
</tr>
<tr>
<td>1e-4</td>
<td>1.0E-04</td>
<td>1.6E-04</td>
<td>1.2E-04</td>
<td>1.1E-04</td>
<td>2.8</td>
<td>1.0E-04</td>
<td>1.5E-04</td>
<td>1.1E-04</td>
<td>1.1E-04</td>
<td>2.8</td>
</tr>
</tbody>
</table>

The sample size for the four AfA studies is 4,500 and for the other studies is 2,000. The total sample size is 30,000, same as the equal sample size scenarios. The left most column is the asymptotic threshold for the one-stage and the two-stage clustering approaches for the Fixed-Effects (FE), Random-Effects (RE), Random-Effects model by Han and Eskin (RE-HE), Binary Effects (BE) model and Meta-analysis of Transehtnic Association studies (MANTRA) models.
Figure 2.2 Power comparisons for ten models (two approaches of five meta-analytic methods) for the four simulated scenarios when the sample size is equal (sample size=3,000 for each study and 30,000 in total, EAF=5% and α=1e-4)

The four scenarios of heterogeneity of allelic effects: (A) Scenario1: Effect-size Homogeneity scenario – equal, non-zero allelic effects on the causal variant across the ten studies; (B) scenario 2: African-Specific Effects scenario – equal, non-zero allelic effects on the causal variant only in four AfA studies (AWS, LWK, MKK and YRI); (C) scenario 3: Effect-size Heterogeneity scenario – non-zero allelic effects in multiple studies, equal effect size among studies in the same ethnic group and the different effect size across the different ethnic groups (effect sizes: 0.05 for four AfA studies, 0.025 for two AsA studies and 0.0125 for four GEM studies); and (D) Western Exposure Effects scenario – equal, non-zero allelic effects on the causal variant only in the study samples living in Europe or USA (AWS, CHD, CEU and TSI).
Figure 2.3 Power comparisons for ten models (two approaches of five meta-analytic methods) for the African-Specific Effects and Effect-size Heterogeneity scenarios for unequal vs. equal sample sizes (EAF=10% and $\alpha=1e^{-4}$)

For equal sample size simulation, the sample size=3,000 for each study and 30,000 in total. For unequal sample size simulation, the sample size in the AfA studies is increased by 1.5 times (4,500) and the sample sizes in the other ancestry studies are decreased by 2/3 (2,000) but the total sample size (30,000) remains the same. (A) African-Specific Effects scenario using unequal sample size; (B) African-Specific Effects scenario using equal sample size; (C) Effect-size Heterogeneity scenario using unequal sample size; (D) Effect-size Heterogeneity scenario using equal sample size.
Figure 2.4 Power comparisons for ten models (two approaches of five meta-analytic methods) for the four simulated scenarios when a causal variant is included in a region (sample size=3,000 for each study and 30,000 in total, EAF=10% and multiple testing adjusted threshold $\alpha=0.1$)

The four scenarios of heterogeneity of allelic effects: (A) Scenario 1: Effect-size Homogeneity scenario; (B) scenario 2: African-Specific Effects scenario; (C) scenario 3: Effect-size Heterogeneity scenario; and (D) Western Exposure Effects scenario.
Figure 2.5 Power comparisons for ten models (two approaches of five meta-analytic methods) for the four simulated scenarios when a causal variant is not included in a region (sample size=3,000 for each study and 30,000 in total, EAF=10% and multiple testing adjusted threshold α=0.1)

The four scenarios of heterogeneity of allelic effects: (A) Scenario1: Effect-size Homogeneity scenario; (B) scenario 2: African-Specific Effects scenario; (C) scenario 3: Effect-size Heterogeneity scenario; and (D) Western Exposure Effects scenario.
2.5 Discussion

Meta-analysis has been predominantly performed among European-ancestry population studies. However, with the use of comprehensive reference panels, the number of genome-wide association studies in other ethnic populations has increased. While meta-analyzing over multiple GWAS across genetically different populations may provide an opportunity to enhance power by increasing sample size and to explore genetic variants that are either transferable across populations or unique to a certain ethnicity (Liu et al. 2013; Liu et al. 2012), this approach also presents challenges such as potential heterogeneity due to differences in linkage disequilibrium, differences in underlying causal variants, differences in genotype platforms leading to differences in imputation accuracy, and possible environmental influence on genetic impacts.

With the opportunities and challenges of trans-ethnic meta-analysis, we proposed a two-stage clustering approach instead of the one-stage approach and assessed the performance of five meta-analytic methods that are currently available and widely being used. Our two-stage approach was applied to account for between-cluster heterogeneity as well as to bolster within-cluster homogeneity. We showed in a simulation study that our two-stage approach can improve power in RE-HE and BE models in the presence of heterogeneity among multiple GWAS due to the genetic dissimilarity and can also reduce the computational intensity by reducing the number of studies entering the meta-analysis in MANTRA while maintaining similar power. (Table 2.8) The average computational time was 5.9 minutes to analyze 652 variants when the number of studies included in the
meta-analysis in MANTRA was three in the two-stage approach while it took 10.9 minutes for the same number of variants when the number of studies was ten in the one-stage approach.

We have examined the power and type I error of five meta-analytic methods. The type I error rates of all Frequentist methods were well controlled in one- and two-stage approaches; we lacked the ability to evaluate the type I error in the Bayesian method MANTRA. For a fair power comparison, we calculated empirical thresholds for all five methods in the two approaches and used empirical thresholds for the power computation. For the region-wise simulation studies, the empirical thresholds are computed using minimum p-value or maximum BF. The results showed that our two-stage approach has noticeable improvement in power especially for BE and RE-HE models and in computational efficiency for MANTRA compared to the crude approach when we assumed that heterogeneity arises from different genetic structure across diverse populations. Thus, the results suggested the two-stage approach may be an effective intermediate step in a trans-ethnic meta-analysis of GWAS that would improve power.

Note that MANTRA already takes into account the estimated similarity among studies with respect to genetic relatedness using a Bayesian partition model. In comparison, our two-stage approach showed no discernable loss in power but reduced the computational burden by reducing the total number of studies entering MANTRA analysis. Therefore, the two-stage clustering approach using the prior knowledge of genetic relatedness in
terms of mean allele frequency improves power when the difference between studies is
due to genetic dissimilarity, and enhances the computational efficiency by diminishing
the computation burden for the Bayesian method MANTRA.

We acknowledge that there may be limitations in our two-stage approach. Because we
use genetic distance as prior information to cluster, the cluster classification is pre-fixed
and is not updated using other information such as allelic effect estimates as in
MANTRA. Further, the fixed threshold applied to the genetic distance in the dendrogram
for clustering is rather subjective and arbitrary, and may be prone to mis-clustering of
genetically heterogeneous studies. However, as shown in Figure 2.6, the two-stage
approach is robust to mis-clustering by taking a sub-optimal dissimilarity value for the
threshold. In practice, clustering using the prior knowledge of ancestry of study samples
is recommended. Alternatively, a data-driven approach can be used to find an optimal
threshold for clustering. Assessing the power in the region-wise analysis using the
empirical threshold computed by minimum p-value or maximum BF in a region would be
more precise if a larger number of replicates were evaluated. Lastly, when heterogeneity
arises from environmental factors rather than genetic dissimilarity, our two-stage
approach tends to lose power due to incorrect clustering. In such a case, further
refinement of current approaches need to be developed, such as introducing another layer
of clustering or treating those environmental factors as random effects, to properly
account for the multiple sources of the heterogeneity and to retain good power.
In conclusion, with the growing interest and availability of GWAS results from diverse populations, the use of appropriate strategies and analytic tools in trans-ethnic meta-analysis will be crucial. In such meta-analyses, our two-stage approach accounts for potential heterogeneity due to genetic dissimilarity and therefore boosts power to detect genetic signals. In addition, our simulation study shows that the conventional FE and RE models may not be suitable for the trans-ethnic meta-analysis in GWAS because the assumption of the FE model may be violated and because the RE model may be too conservative. In that case, the three recently developed methods have great advantages over the conventional methods. They are robust and more powerful especially when the studies in the meta-analysis are potentially heterogeneous. However, inappropriate and incorrect clustering of studies can result in power loss. Therefore the use of the three robust recent meta-analytic methods in the appropriate two-stage clustering strategy may enhance the ability to uncover and understand genetic signals contributing to human complex disease.
Table 2.8 Average computational time (in minutes) for MANTRA using the one-stage (10 Clusters) and three two-stage clustering (2 Clusters, 3 Clusters, 5 Clusters) approaches when sample size was equal

<table>
<thead>
<tr>
<th>(In minutes)</th>
<th>One-stage clustering approach</th>
<th>Two-stage clustering approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>652 SNPs</td>
<td>10 Clusters: 10.9</td>
<td>2 Clusters: 5.0</td>
</tr>
<tr>
<td>11,575 SNPs</td>
<td>192.3</td>
<td>89.5</td>
</tr>
</tbody>
</table>

We used Intel® Xeon® Processor E5-2650 v2 (20M Cache, 2.60 GHz) for this comparative computational time analysis.
Figure 2.6 Sensitivity analysis of the two-stage approach by varying the threshold of dissimilarity resulting in a different number of clusters in addition to the one-stage (10 studies) and the initial two-stage approach (3 Clusters)

In the 2 Clusters scenario, we included four African studies in one cluster and the rest of the studies in the other cluster. In the 5 Clusters scenario, we had five clusters with studies in each cluster as follows: the first two African studies from Figure 2.1; the other two African studies; the two Asian studies; MEX and GIH; and the two EuA studies. In total, twenty models (four clustering approaches of five meta-analytic methods) were presented with Effect-Heterogeneity (Scenario 3) simulation setting when the sample size is equal (sample size=3,000 for each study and 30,000 in total, $\alpha=1e^{-4}$).
Chapter 3  Statistical Model with Multiple Random Effects to Account for Heterogeneity

3.1  Introduction

The popularity of meta-analysis has consistently grown in the fields of clinical trials, epidemiological studies, statistical genetics, and many other areas of medical research (Lee, Bausell, and Berman 2001; DerSimonian and Laird 2015; Sutton and Higgins 2008; Zeggini and Ioannidis 2009). Meta-analysis refers to a statistical approach to aggregate evidence from multiple independent studies to specifically address a common hypothesis and gain a better understanding of treatment effects. The sample size of a single individual study is often too small to detect and provide a reliable estimate of a treatment effect. The primary advantage of meta-analysis is to increase power by increasing sample size and to derive a robust pooled estimate closest to the true but unknown parameters of interest. Ideally, the underlying effect sizes in each participating study are the same, and the variation of the study-specific effect estimates is due to sampling variation alone. However, such a simple setting is not always the case in reality and variation due to multiple sources other than chance may contribute to the complex setting of heterogeneous studies. The goal of this paper is to develop a statistical framework for meta-analysis to account for potential heterogeneity from multiple sources which may be quantified by cohort-specific characteristics.
Heterogeneity across participating studies may be explored via meta-analysis. There has been increasing interest in determining which study covariates may influence on the main effects of interest. In clinical and epidemiological studies, one may assume that factors such as geographical location, drug dosage or duration of treatment might change the effects of treatment and cause heterogeneity in effect size across studies (Simmonds and Higgins 2007; Sterne et al. 2002; Abrams, Gillies, and Lambert 2005; Colditz et al. 1994). Heterogeneity can also arise from the use of different techniques of measuring outcomes and study designs (S G Thompson 1994; Abrams, Gillies, and Lambert 2005; Simon G. Thompson and Sharp 1999). In statistical genetics, differences in genetic architecture such as linkage disequilibrium (LD), differences in underlying functional variants and differences in allele frequencies of functional variants can result in heterogeneity in the effect sizes in the meta-analysis of GWAS (Morris 2011; Y. Li and Keating 2014; Fu, Festen, and Wijmenga 2011). Different genotyping quality and imputation accuracy can also contribute to heterogeneity across studies (de Bakker et al. 2008; Zaitlen and Eskin 2010). Furthermore, non-genetic exposures such as different level of environmental exposure or life style may also differ across study cohorts, and may affect effect size. Thus, clinical, methodological, and/or genetic differences may lead to statistical heterogeneity that can be accounted for using an appropriate statistical model.

Conventional meta-analytic methods, inverse-variance-weighted fixed-effects (FE) and random-effects (RE) models, have widely been used in the meta-analysis. The FE model
provides a weighted average of effect estimates under the assumption of that the underlying effects are the same across studies (Fleiss 1993; Hedges and Olkin 1985). The inverse of the variance of the estimate can be used as a weight to account for difference in sample size, so a larger study is given greater weight compared to a smaller study. In the presence of heterogeneity, the FE model is inappropriate because it underestimates the standard error of the effect estimate. We can formally test for homogeneity of effect sizes using Cochran’s Q statistic (Cochran 1952). We can use the $I^2$ statistic (Higgins and Thompson 2002) to quantify how much of the total variability is due to heterogeneity in the effect sizes. Higgins and Thompson (2002) suggest that a value of $I^2$ greater than 50% might indicate substantial heterogeneity among effect sizes and that such heterogeneity might influence the interpretation of the meta-analysis that we should not ignore. In the presence of heterogeneity, the use of a RE model is more appropriate. The RE model properly estimates the standard error and yields more accurate confidence intervals by incorporating between-study variability into the weights (DerSimonian and Laird 1986; Hedges and Olkin 1985; Fleiss 1993). However, the conventional RE model is less powerful than the FE model even in the presence of heterogeneity (B. Han and Eskin 2011) because of its conservative assumption that the between-study variance is not zero under the null hypothesis. Therefore, the conventional RE model is not optimal for testing pooled effects associated with an outcome via meta-analysis and some methodological development is needed to improve power.
One alternative to the conventional meta-analytic models is meta-regression. Meta-regression is a regression approach in which effect size is used as the dependent variable and study-level covariate measures as explanatory variables to examine the impact of study-level characteristics on treatment effects in clinical trials (Simmonds and Higgins 2007; Abrams, Gillies, and Lambert 2005; Greenland 1987). Similar to the conventional FE and RE models, FE- or RE-meta-regression (FE-metaReg / RE-metaReg) can be used depending on the assumption of homogeneity of effect size. Under the FE-metaReg, for a set of covariate values, we assume that the true underlying effect among studies is the same. Under the RE-metaReg, for any set of covariate values, we assume that the true effects may vary across studies and may follow a certain distribution. When data are pulled from the literature or from studies conducted by different investigators, it is plausible to assume that the covariates in a model would explain some, but not all, of the true heterogeneity in effect sizes across studies, and any remaining heterogeneity should be acknowledged in the analysis (S. Thompson and Higgins 2002; Viechtbauer 2007). In this case, the RE-metaReg is more appropriate to use because it properly accounts for the remaining heterogeneity that has not been explained by the covariates in the model.

However, both metaReg models test the null hypothesis of no linear relationship between the covariates and the effect size. In reality, this linearity assumption is questionable. If the true relationship is non-linear between the covariates and the effect size, it is hard to determine the true relationship with the small number of studies in a typical meta-analysis. In this case, the use of the metaReg models would violate the linearity assumptions.
Another random-effects meta-analytic approach proposed by Han and Eskin (RE-HE) (B. Han and Eskin 2011) improves power compared to the conventional FE and RE models especially in the presence of heterogeneity. The null hypothesis for the RE-HE approach is that the mean effect is 0 and there is no heterogeneity. This is in contrast to the conventional RE model, which tests the null hypothesis that the mean effect is 0 but allows for heterogeneity. The alternative for both models is the same – that the mean effect is not 0, or that there is heterogeneity. The RE-HE model assumes a single aggregated between-study heterogeneity that blends all potential sources. However, the assumption of a single source of heterogeneity may not be realistic and heterogeneity may arise from multiple sources as described earlier.

In this chapter, we incorporate the idea of testing the between-study variances of RE-HE approach under meta-regression framework and propose a new statistical model that accommodates multiple sources of potential heterogeneity which can be measured by study-level characteristics.

We develop our method in section 2. In order to account for different sources of heterogeneity, we assume that study characteristics are pertinent to heterogeneity in the effect size and treat them as random effects. We use a linear mixed effect model to incorporate the multiple random effects and implement the likelihood ratio test (LRT) to perform hypothesis testing. In section 3, we demonstrate the performance of our proposed
approach in extensive simulations in terms of type I error and power, and compare the approach to other existing meta-analytic methods including FE, RE, RE-HE and FE-/RE-metaReg over a range of heterogeneity scenarios. We then apply our new model in section 4 to meta-analyze 15 GWAS studies (N=36,846) of fasting insulin from the Meta-Analyses of Glucose- and Insulin-related traits Consortium (MAGIC) to identify variants associated with fasting insulin considering study-specific mean age and BMI as random effects. We conclude this article with a discussion in section 5.

3.2 Methods

3.2.1 Linear Mixed Effects Model with Multiple Random Effects

3.2.1.1 Notations and Assumptions

We first define notation and assumptions and then derive a test statistic that accounts for heterogeneity from multiple sources. Suppose we would like to determine the association between a treatment and an outcome of interest by synthesizing the association findings from $K$ studies. For each of these $K$ studies, summary statistics measuring the association with the outcome of interest are available and study-level characteristics are measured. We assume that the treatment effect for each study may vary according to the study characteristics and thus has potential heterogeneity. We treat these study-level covariates as random effects to take the heterogeneity into consideration in the statistical modeling using a linear mixed effects model as follows:

$$ b = \beta + Zv + e , $$
where \( \mathbf{b} \) is a \( K \)-dimensional vector of observed effect estimates from \( K \) studies and \( \mathbf{\beta} \) is an unknown true effect size. \( \mathbf{Z} \) is a matrix of size \( K \) by \( r \) for the \( r \) study characteristics and \( \mathbf{v} \) is an \( r \)-dimensional vector of unknown random effects. We assume the random effects, \( \mathbf{v} \), and the residual errors, \( \mathbf{e} \), are uncorrelated and follow multivariate normal distributions with mean \( \mathbf{0} \) and covariance matrices, \( \Sigma_v \) and \( \mathbf{R} \), respectively. That is,

\[
\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \Sigma_v),
\]

\[
\mathbf{e} \sim \mathcal{N}(\mathbf{0}, \mathbf{R}).
\]

The covariance matrix for the random effects is \( \Sigma_v = \text{var}[\mathbf{v}] = \begin{bmatrix} \tau_1^2 & \cdots & \tau_{1r} \\ \vdots & \ddots & \vdots \\ \tau_{r1} & \cdots & \tau_r^2 \end{bmatrix} \). We assume the random effects are uncorrelated and therefore the off-diagonal elements of \( \Sigma_v \) are equal to zero (\( \tau_{ij} = \tau_{ji} = 0 \), for all \( i \neq j \)). The covariance matrix for the residual errors is \( \mathbf{R} = \begin{bmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_K^2 \end{bmatrix} \). We make the conventional assumption that the elements of the within-study covariance matrix \( \mathbf{R} \) are known values (Higgins and Thompson 2002).

The parameters of interest are \( \mathbf{\theta} = (\beta, \tau_1^2, \ldots, \tau_r^2) \). Under this setting, we would like to establish a (log-)likelihood ratio test (LRT) statistic to test \( H_0: \mathbf{\theta} \in \Theta_0 \) vs. \( H_1: \mathbf{\theta} \in \Theta_1 \), where \( \Theta_0 = \{ \mathbf{\theta}_0: \beta = 0, \tau_1^2 = 0, \ldots, \tau_r^2 = 0 \} \) and \( \Theta_1 = \{ \mathbf{\theta}: \beta \neq 0, \tau_1^2 > 0, \ldots, \text{or } \tau_r^2 > 0 \} \). In our composite hypotheses, we assume no association (\( \beta = 0 \)) and no heterogeneity (\( \tau_1^2 = 0, \ldots, \tau_r^2 = 0 \)) under the null hypothesis (B. Han and Eskin 2011; Higgins, Thompson, and Spiegelhalter 2009; Lebrec, Stijnen, and van Houwelingen 2009).
2010). Under these assumptions, the marginal distribution of \( b \) follows a multivariate normal distribution with mean \( \mathbb{E}[b] = \beta \) and covariance matrix \( \text{Var}[b] = W \), i.e.

\[
b \sim N(\beta, W),
\]

where \( W = Z\Sigma_vZ' + R \).

### 3.2.1.2 Likelihood Ratio Test Statistic

Using this marginal distribution of \( b \), we can obtain the likelihoods (L) under the null \( \theta \in \Theta_0 \) and the union of \( \theta \in \Theta_0 \cup \Theta_1 \) as follows:

\[
L_{\theta_0} = \frac{1}{(\sqrt{2\pi})^k} |R|^{-\frac{1}{2}} \ast \exp \left( -\frac{1}{2} b' R^{-1} b \right),
\]

\[
L_{\theta} = \frac{1}{(\sqrt{2\pi})^k} |W|^{-\frac{1}{2}} \ast \exp \left( -\frac{1}{2} (b - \beta)' W^{-1}(b - \beta) \right).
\]

We are interested in estimating the parameters that maximize the likelihood under the alternative hypothesis. There is a closed-form solution for the maximum likelihood estimator of the true effect size \( \beta \); however, there is no closed-form for the variances of the random effects \( \tau_j^2 \), for \( j = 1, ..., r \) (See derivations in Appendix A). Hence we use a Nelder-Mead optimization algorithm implemented in dfoptim R package (Varadhan and Borchers 2011; Kelley 1999) to numerically obtain the parameter estimates which optimize the likelihood. With these estimates, the LRT statistic can be expressed as follows:
\[
T = -2 \ln \left( \frac{L_{\hat{\theta}}}{L_{\theta}} \right) = \left( \ln |R| + b' R^{-1} b \right) - \left( \ln |\bar{W}| + (b - \bar{\beta})' \bar{W}^{-1} (b - \bar{\beta}) \right),
\]

where \( \bar{W} = Z \hat{\Sigma}_v Z' + R, \hat{\Sigma}_v = \text{diag}(\hat{\tau}_1^2, \hat{\tau}_2^2, ..., \hat{\tau}_r^2), R = \text{diag}(\sigma_1^2, \sigma_2^2, ..., \sigma_r^2), \beta \) and \( \hat{\tau}_j^2 \), for \( j = 1, ..., r \), are the estimates from the numerical approach (Kelley 1999; Varadhan and Borchers 2011).

### 3.2.2 Asymptotic Distribution

The statistical significance of the test statistic can be assessed using the asymptotic distribution of the LRT statistic \( T \). \( \beta \) is unrestricted but \( \tau_j^2 \), for \( j = 1, ..., r \), are the boundary parameters restricted to be non-negative. Hence \( \beta \) corresponds to a \( \chi_1^2 \) distribution and each of \( \tau_j^2 \), for \( j = 1, ..., r \), to a 50:50 mixture of \( \chi_0^2 \) and \( \chi_1^2 \) distributions (Self and Liang 1987). The asymptotic distribution of the test statistic \( T \) is determined by the number of random effects we put in a model. For example, if we assume one random effect in the model \((j=1)\) the test statistic \( T \) would asymptotically follow a 50:50 mixture of \( \chi_1^2 \) and \( \chi_2^2 \) distributions (B. Han and Eskin 2011; Self and Liang 1987). When we assume two random effects \((j=2)\) with no correlation, the test statistic would asymptotically follow a mixture of three Chi-square distributions, \( \chi_1^2, \chi_2^2 \) and \( \chi_3^2 \), with mixing probability \( p = \frac{1}{4} \) (Self and Liang 1987). Therefore, the resulting distribution for the test statistic \( T \) with two random effects asymptotically follows a \( \frac{1}{4} \chi_1^2 + \frac{1}{2} \chi_2^2 + \frac{1}{4} \chi_3^2 \).
We can easily expand this for more than two random effects under the assumption of no pair-wise correlation among the random effects. The general form of the asymptotic distribution of the LRT statistic, testing \( H_0: \beta = 0, \tau_1^2 = 0, \ldots, \tau_r^2 = 0 \), can be expressed as a binomial function with a mixing probability \( p = \left( \frac{1}{2} \right)^r \) as follows

\[
\sum_{j=0}^{r} \binom{r}{j} \left( \frac{1}{2} \right)^r \chi_{j+1}^2,
\]

where \( r \) indicates the number of random effects and the mixing probability assuming the random effects are pairwise uncorrelated. For the one random effects case, the mixing probability is \( p = \frac{1}{2} \) and the distribution of the test statistic, testing \( H_0: \beta = 0, \tau^2 = 0 \), would be \( \sum_{j=0}^{1} \frac{1}{2} \chi_{j+1}^2 \). If we are testing two random effects along with the main effect, \( H_0: \beta = 0, \tau_1^2 = 0, \tau_2^2 = 0 \), the distribution would be \( \sum_{j=0}^{2} \binom{2}{j} \left( \frac{1}{2} \right)^2 \chi_{j+1}^2 \).

### 3.2.3 Corrected Null Distribution

Although we derive the asymptotic distribution, a large number of studies are required to reach the asymptotic distribution, and the number of studies in a meta-analysis is typically small in practice. If the asymptotic distribution is not achieved, \( p \)-values tend to be overly conservative (B. Han and Eskin 2011; B. Han, Kang, and Eskin 2009).

There are several ways to better approximate the distribution of the test statistic \( T \) for limited sample sizes. Han and Eskin (2011) evaluated \( p \)-values by tabulated values they calculated for each possible number of studies from 2 to 50 (B. Han and Eskin 2011).
Han and Pan (2010) re-calibrated the null distribution of their test statistic so that the type I error rate would be well controlled (F. Han and Pan 2010). Han and Pan empirically evaluated scaling and shifting parameters of their asymptotic distribution and used the scaled-shifted null distribution to obtain a p-value.

We take a similar approach as described in the Han and Pan (2010) to find a “corrected” null distribution that would correct the conservativeness of the asymptotic distribution. (See the type I error rates using the asymptotic distribution in Table 3.1.) Here we consider the case of two random effects. Because we know that the asymptotic distribution of our test statistic for two random effects follows the mixture of $\chi^2_g$ with $g = 1, 2$ and 3 degrees of freedom, we compare several candidate distributions with $g \leq 3$ from the three general forms:

(a) $a \chi^2_g + b$ where $g=$degree(s) of freedom of 1, 2, or 3.

(b) $a \chi^2_{g_1} + b \chi^2_{g_2} + c$ where $g_1$ and $g_2$ are a combination of degree(s) of freedom of 1, 2, and 3 such that $g_1=1$ or 2, and $g_2=2$ or 3.

(c) $a \chi^2_1 + b \chi^2_2 + c \chi^2_3$.

We want to estimate the scaled-shifted parameters, $a$, $b$ and/or $c$, and compare type I errors evaluated from of the candidate distributions. The proposed implementation to estimate the scaled-shifted parameters for each candidate distribution is as follows:
(1) Generate summary measures, $\sigma_i^2$ and $b_i$, for the $i^{th}$ study, where $i = 1, \ldots, K$, using $\sigma_i^2 \sim \chi_1^2$ and $b_i \sim N(0, \sigma_i^2)$.

(2) Numerically estimate the parameters $\beta$, $\tau_1^2$ and $\tau_2^2$, which optimize the likelihood ratio using a Nelder-Mead optimization algorithm implemented in dfoptim R package (Varadhan and Borchers 2011; Kelley 1999).

(3) Compute the null LRT statistic, $LRT_{emp}$, for $N=50$ as shown in the equation 1 in 3.2.1.2 using the generated summary measures in (1) and parameter estimates in (2).

(4) Repeat the process $B$ times, giving $B$ test statistics $LRT_{emp}$.

(5) Generate random numbers from $\chi_1^2$, $\chi_2^2$ and $\chi_3^2$ and compare with the $B$ test statistics $LRT_{emp}$ in (4) to estimate the scaled-shifted parameters ($a$, $b$ and/or $c$) using a regression, method of moments (F. Han and Pan 2010; S. Li and Cui 2012), or a numerical method (Turlach).

Once the scaled-shifted parameters are estimated, we construct an empirical distribution and calculate p-values of the $LRT_{obs}$ statistics by calculating the proportion of the $B$ test statistics $LRT_{emp}$ that are greater than or equal to $LRT_{obs}$. We evaluate type I error rates by calculating the proportion of the empirical p-values that passes a pre-specified $\alpha$ level considering four study covariate combinations; continuous, dichotomous, proportion variables, and a combination of continuous and dichotomous variables. So we compare the type I error rates for the four study covariates among the candidate distributions and
choose one that corrects the conservativeness of the asymptotic distribution with well-controlled type I error rates.

After carefully following these steps and comparing the candidate distributions (Appendix B), the following mixture distribution, \( a \chi_1^2 + b \chi_2^2 + c \), shows anti-conservative but well-controlled type I error rates and therefore is used as our “corrected” null distribution. The scaled-shifted parameters \( a, b, \) and \( c \) are estimated according to the study covariates following the proposed steps (1)-(4).

### 3.3 Simulation Study

#### 3.3.1 Study Design of Simulation

**3.3.1.1 General Setting**

We conducted a simulation study to evaluate the type I error and power of our test statistic \( T \) in the presence or absence of random effects. For simplicity and without loss of generality, we considered the situation with two random effects for this simulation study.

We compared the performance of our approach to the conventional FE and RE models, Han and Eskin’s modified random-effects, RE-HE, and FE-/RE-meta-regression (metaReg) methods. We considered three scenarios where both of the random effects can be represented by proportion, continuous, or dichotomous characteristics, and an additional scenario having different types of study-level covariates where the two covariates were a combination continuous and dichotomous variables.
3.3.1.2 Type I error Simulation

We evaluated the type I error under the null hypothesis of no association and no heterogeneity by setting all three parameters to zero, \((H_0: \beta = 0, \tau_1^2 = 0, \text{ and } \tau_2^2 = 0)\).

We first generated within-study variances, \(\sigma_i^2\), from a \(\chi_1^2\) distribution and exclude extreme values that are less than 10% or greater than 90% percentiles of the distribution. Among remaining values, we randomly selected \(K\) values for within-study variances, \(\sigma_i^2\), for \(i = 1, ..., K\), and assigned them to studies as the elements of the diagonal covariance matrix \(R\). We generated observed effect estimates, \(b\), from a multivariate normal distribution \(N(0, R)\). Then we numerically estimated parameters, \(\beta, \tau_1^2\) and \(\tau_2^2\) from these data, and computed a null LRT statistic. We vary the number of studies, \(K\), with \(K= 10, 20, 50\) and 100, to investigate the effect of the number of studies on the meta-analysis results.

We can evaluate our test statistic and corresponding p-value from the asymptotic distribution. However, given the small number of studies, the asymptotic distribution yielded overly conservative type I error rates in our preliminary simulation results (Table 3.1) and a corrected null distribution was needed. Thus we established a corrected null distribution as described earlier (in 3.2.3) and computed the p-value from the corrected null distribution. We evaluated the empirical type I error rates at different significance levels \(\alpha\) by computing the proportion of p-values that are less than or equal to the corresponding \(\alpha\) level. We repeated this procedure for each four study-characteristic
scenarios where the characteristics of the random variables are continuous, dichotomous, proportion covariates, and a combination of continuous and dichotomous variables.

### 3.3.1.3 Power Simulation

We evaluated power in three heterogeneity scenarios. In scenario 1, we considered a fixed effects (i.e., no heterogeneity, \( \beta \neq 0, \tau_1^2 = 0, \tau_2^2 = 0 \)) scenario, in which we assumed that the underlying true effect was consistent across studies and so presented no heterogeneity. In scenario 2, we considered situations where main effects and two random effects were present and non-zero among the studies (i.e. \( \beta \neq 0, \tau_1^2 \neq 0, \tau_2^2 \neq 0 \)). In scenario 3, we considered the situation where main effects and only one random effects were non-zero (i.e. \( \beta \neq 0, \tau_1^2 \neq 0, \tau_2^2 = 0 \)). The within-study variances, \( R \), were generated and assigned to the studies in the same way as in the type I error simulation. However, we simulated the observed values of effect estimates, \( b \), from a multivariate normal distribution \( MN(\boldsymbol{\beta}, \boldsymbol{W}) \), where \( \boldsymbol{\beta} \) is a vector of the pre-specified true effect sizes, 

\[
\boldsymbol{W} = \boldsymbol{Z} \boldsymbol{\Sigma}_v \boldsymbol{Z}' + \boldsymbol{R}
\]

with 

\[
\boldsymbol{\Sigma}_v = \begin{bmatrix}
\tau_1^2 & 0 \\
0 & \tau_2^2 
\end{bmatrix}
\]

\( \boldsymbol{Z} \) is a \( K \) by \( r \) matrix containing study-level covariate measures as elements, where \( K \) is the number of studies and \( r \) is the number of unknown random effects.

In all simulation scenarios, we considered two random effects (\( r = 2 \)) and the study-level covariates were generated as follows. For the continuous variables, the values were generated from a standard normal distribution, \( N(0, 1) \). When the variables were
dichotomous, we generated from a binomial distribution with a success probability of $p$, $Bernoulli(p = 0.5)$. We also varied the probability $p$ and showed results in Appendix C.2. If the variables were proportions, we generated from a uniform distribution between 0 and 1, $Uniform(0,1)$. To evaluate power, we varied the values of four parameters: the effect size, $\beta$ from 0.01 to 0.05, 0.1, 0.15 and 0.2, the number of studies, $K$, from 10, 20, 50 and 100, and the two between-study variances, $\tau_1^2$ and $\tau_2^2$, from 0.05, 0.1, 0.2, 0.5, 1 and 2 to reflect the level of heterogeneity. In order to investigate the effect of each of the four parameters ($\beta, K, \tau_1^2$ and $\tau_2^2$) on power, we changed the value of one parameter while keeping the other parameters constant.

3.3.2 Simulation Study Results

3.3.2.1 Type I Error Simulations

Table 3.1 shows type I error rates of our LRT approach using both asymptotic and empirical distributions at different $\alpha$ levels in four different settings of study characteristics. The results suggested that the type I error rates from the asymptotic distribution were overly conservative. In contrast, our corrected distribution retained the correct type I error rates regardless of the number of studies and the study characteristics. Therefore, in subsequent real data analysis we used the corrected distribution to evaluate p-values. The type I error rates of the FE, RE, RE-HE and FE-/RE-metaReg at $\alpha=0.01$ were shown in Table 3.2. The p-values for RE-HE model were computed using metasoft
(B. Han and Eskin 2011). All models had well controlled type I error rates even though RE, RE-HE and RE-metaReg were somewhat conservative.

### 3.3.2.2 Power Simulations

Power of our LRT, FE, RE, RE-HE, and FE-/RE-metaReg models for the three heterogeneity scenarios were evaluated. We computed p-values of our LRT using both corrected (cLRT) and asymptotic (aLRT) distributions. In all scenarios, two study-level covariates were included as the sources of heterogeneity.

**Figure 3.1** shows the power results of the scenario 1 where we assumed no heterogeneity. Power increased in all models as the number of studies or the effect size increased. FE was the most powerful when there was no heterogeneity in effect size followed by cLRT, RE-HE and RE in all four different study characteristic scenarios. aLRT and FE-/RE-metaReg were less powerful than RE.

**Figure 3.2** shows the power results under the scenario 2 where we introduced heterogeneity by two non-zero random effects. In this scenario, each method performed slightly differently according the different study covariates. Overall, power improved in all models as the number of studies, the effect size, or the size of between-study variances increased. When the covariates were continuous, cLRT, RE-HE, FE-metaReg and aLRT performed fairly well, followed by RE-metaReg, FE and RE. When the covariates were a combination of the continuous and dichotomous, FE and RE models remained less
powerful but the power of RE-metaReg became similar to aLRT and RE-HE while cLRT and FE-metaReg remained most powerful. When the covariates were dichotomous or a proportion, the difference in power among the models became less distinctive, because in this scenario, the between-study variance pertinent to the study-level covariates became relatively small. In the dichotomous and proportion scenarios, the cLRT remained most powerful among the models followed by RE-HE and FE-metaReg in general, and aLRT, RE-metaReg and FE models were not far behind even though there were some fluctuations among the remaining models in terms of power.

The conventional FE and RE appeared less powerful than the other models as heterogeneity was introduced, especially when the study covariates were continuous or a combination of continuous and dichotomous. The RE model was least powerful even when heterogeneity was present and even had decreased power in the continuous covariates scenario as the between-study variances increased, as shown in Figure 3.2C. This phenomenon was previously observed (B. Han and Eskin 2011; Hong et al.). The RE model assumes non-zero between-study variance under the null hypothesis. As a result, there is no gain in power in the RE model with larger between-study variances.

We additionally show the scenario of a single non-zero random effect in Figure 3.3. Here, the overall trend of power results appeared similar to that of scenario 2. The cLRT was consistently the most powerful approach among the models when only one of the two random effects was non-zero.
Table 3.1 Type I error rates of our approach at different $\alpha$ levels using the asymptotic distribution with fifty studies and the corrected distribution

<table>
<thead>
<tr>
<th>$\alpha$ levels</th>
<th>Study Characteristics $^a$</th>
<th>Asymptotic</th>
<th>Corrected</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$N^b=50$</td>
<td>$N=10$</td>
</tr>
<tr>
<td>$\alpha=0.01$</td>
<td>CONT+CONT</td>
<td>0.0045</td>
<td>0.0100</td>
</tr>
<tr>
<td></td>
<td>DICH+DICH</td>
<td>0.0035</td>
<td>0.0099</td>
</tr>
<tr>
<td></td>
<td>PROP+PROP</td>
<td>0.0028</td>
<td>0.0094</td>
</tr>
<tr>
<td></td>
<td>CONT+DICH</td>
<td>0.0040</td>
<td>0.0102</td>
</tr>
<tr>
<td>$\alpha=0.001$</td>
<td>CONT+CONT</td>
<td>0.0004</td>
<td>0.0008</td>
</tr>
<tr>
<td></td>
<td>DICH+DICH</td>
<td>0.0003</td>
<td>0.0009</td>
</tr>
<tr>
<td></td>
<td>PROP+PROP</td>
<td>0.0003</td>
<td>0.0010</td>
</tr>
<tr>
<td></td>
<td>CONT+DICH</td>
<td>0.0002</td>
<td>0.0009</td>
</tr>
</tbody>
</table>

We vary the number of studies from 10 to 20, 50 and 100. Continuous, dichotomous, proportion and a combination of continuous and dichotomous covariates are used for the sources of heterogeneity. The p-values are assessed under the null hypothesis of no association and no heterogeneity using 100,000 replicates.

$^a$ Types of study characteristics as the sources of heterogeneity: CONT-Continuous, DICH-dichotomous, PROP-proportion, CONT+DICH-a combination of continuous and dichotomous covariates.

$^b$ The number of studies in the simulation study.
Table 3.2 Type I error rates of FE, RE, RE-HE and FE-meta-regression at $\alpha=0.01$

<table>
<thead>
<tr>
<th>$\alpha$ levels</th>
<th>Different Models $^a$</th>
<th>N=10</th>
<th>20</th>
<th>50</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha=0.01$</td>
<td>FE</td>
<td>0.0101</td>
<td>0.0101</td>
<td>0.0098</td>
<td>0.0101</td>
</tr>
<tr>
<td></td>
<td>RE</td>
<td>0.0065</td>
<td>0.0067</td>
<td>0.0071</td>
<td>0.0074</td>
</tr>
<tr>
<td></td>
<td>RE-HE</td>
<td>0.0088</td>
<td>0.0085</td>
<td>0.0086</td>
<td>0.0072</td>
</tr>
<tr>
<td></td>
<td>CONT $^b$</td>
<td>0.0101</td>
<td>0.0103</td>
<td>0.0103</td>
<td>0.0100</td>
</tr>
<tr>
<td></td>
<td>DICH $^b$</td>
<td>0.0103</td>
<td>0.0102</td>
<td>0.0101</td>
<td>0.0100</td>
</tr>
<tr>
<td></td>
<td>PROP $^b$</td>
<td>0.0098</td>
<td>0.0097</td>
<td>0.0100</td>
<td>0.0098</td>
</tr>
<tr>
<td></td>
<td>CONT+DICH $^b$</td>
<td>0.0099</td>
<td>0.0097</td>
<td>0.0102</td>
<td>0.0094</td>
</tr>
<tr>
<td></td>
<td>FE-metaReg $^b$</td>
<td>0.0069</td>
<td>0.0073</td>
<td>0.0077</td>
<td>0.0078</td>
</tr>
<tr>
<td></td>
<td>DICH $^b$</td>
<td>0.0071</td>
<td>0.0074</td>
<td>0.0075</td>
<td>0.0076</td>
</tr>
<tr>
<td></td>
<td>PROP $^b$</td>
<td>0.0067</td>
<td>0.0065</td>
<td>0.0074</td>
<td>0.0074</td>
</tr>
<tr>
<td></td>
<td>CONT+DICH $^b$</td>
<td>0.0068</td>
<td>0.0068</td>
<td>0.0079</td>
<td>0.0070</td>
</tr>
</tbody>
</table>

We vary the number of studies from 10 to 20, 50 and 100. Continuous, dichotomous, proportion and a combination of continuous and dichotomous covariates are used for the sources of heterogeneity. The p-values are assessed under the null hypothesis of no association and no heterogeneity using 100,000 replicates.

$^a$ Different models used; FE, Fixed effects; RE, Random effects; RE-HE, Han and Eskin random effects; FE-metaReg, Fixed effects meta-regression; RE-metaReg, Random effects meta-regression models.

$^b$ Types of study characteristics used only for the meta-regression model; CONT-Continuous, DICH-dichotomous, PROP-proportion, CONT+DICH-a combination of continuous and dichotomous covariates.
Figure 3.1 Power results of Scenario1: Effect-size Homogeneity scenario where random effects are zero.

Continuous, dichotomous, proportion and a combination of continuous and dichotomous covariates are used for the sources of heterogeneity.

Five meta-analytic methods are compared; our new LRT model using Corrected (C) and Asymptotic (A) distributions, modified RE by Han and Eskin (H), FE(M)/RE(m)-meta-regression global test, FE (F) and RE (R) models.
Figure 3.2 Power results of Scenario 2: Effect-size Heterogeneity scenario with two non-zero random effects

Continuous, dichotomous, proportion and a combination of continuous and dichotomous covariates are used for the sources of heterogeneity.
Five meta-analytic methods are compared: our new LRT model using Corrected (C) and Asymptotic (A) distributions, modified RE by Han and Eskin (H), FE(M)-/RE(m)-meta-regression global test, FE (F) and RE (R) models.
Figure 3.3 Power results of Scenario 3: Effect-size Heterogeneity scenario with one random effect where only one random effect is present and non-zero

Continuous, dichotomous, proportion and a combination of continuous and dichotomous covariates are used for the sources of heterogeneity.

Five meta-analytic methods are compared; our new LRT model using Corrected (C) and Asymptotic (A) distributions, modified RE by Han and Eskin (H), FE(M)/RE(m)-meta-regression global test, FE (F) and RE (R) models.
3.4 Application to Meta-Analysis of GWAS on Fasting Insulin

3.4.1 Background

The prevalence and burden of type 2 diabetes (T2D) has increased over time (Cowie et al. 2009). T2D accounts for 90-95% of all diagnosed diabetic cases, and its complications are major risk factors for morbidity and mortality (Engelgau 2004). In addition, T2D usually begins with insulin resistance in which insulin that is produced in the body is not properly used. As a result, glucose remains in the blood instead of being absorbed by cells and blood glucose level rises above the normal range. Hyperglycemia, an excess of glucose in the bloodstream, may involve complications, including visual impairment, kidney failure, lower-extremity disease, and increased risk of mortality (Engelgau 2004).

There is great interest in explaining how genes confer genetic risk in different environmental contexts and in better understanding the etiology of T2D (Permutt, Wasson, and Cox 2005).

3.4.2 Statistical Analysis

We assembled fifteen cohorts from the Meta-Analyses of Glucose- and Insulin-related traits Consortium (MAGIC), totaling up to 36,846 non-diabetic individuals of European ancestry from fifteen studies. These studies were part of a previous study that performed genome-wide joint meta-analysis analysis looking for genetic variants associated with glycemic traits simultaneously adjusting for body mass index (BMI) and allowing for genetic interaction with BMI (Manning et al. 2012). Specifically, each participating
cohort provided results from genome-wide association analysis of single nucleotide polymorphisms (SNPs) with fasting insulin adjusting for age, sex, BMI, and principal components of genetic information to account for potential population stratification (Price et al. 2006). Fasting insulin level was measured from whole blood, plasma, or serum, and was log-transformed for analysis. Both genotyped and imputed SNPs were used. We include genotyped SNPs that passed quality control inclusion criteria, (i) Hardy-Weinberg equilibrium P-value of >1e-6; (ii) minor allele frequency of > 1%; (iii) variant call rate > 95%. Imputation was performed based on the International HapMap Project (HapMap) Phase 2 Utah residents of Northern and Western European ancestry (CEU) population using MACH (Y. Li et al. 2009; Y. Li et al. 2010) (r² >0.3) or IMPUTE (Howie, Donnelly, and Marchini 2009; Marchini et al. 2007) (proper_info>0.4).

It has been shown that BMI influences the heritability of insulin resistance (X Wang et al. 2009) and modifies the association of genotypes with insulin resistance (Mayer et al. 1996; Herbert et al. 2006). Thus, there is interest in finding genetic variants associated with fasting insulin accounting for heterogeneity attributed to BMI. Three study characteristics, mean age, mean BMI and the proportion of the sample of each gender, were reported in supplementary material of the original paper (Manning et al. 2012). For ease of comparison with our simulation study, we selected the two covariates showing the most variability across studies, mean BMI and mean age, to include as random effects to accommodate additional heterogeneity that may not have been accounted for in the
original GWAS. Because our approach showed competitive power in the simulation study, we evaluated our approach and compared it to the conventional FE model.

As described in the simulation study, we used an empirical distribution to calculate p-values instead of the asymptotic distribution due to its conservativeness. We randomly selected 100,000 SNPs that were present across all of the fifteen cohorts and constructed the empirical distribution using the corresponding standard errors of the effects of randomly selected SNPs as follows. First, we simulated null effect sizes from $MN(0, R)$ where the diagonal elements of the covariance matrix $R$ were the variances of the SNP from the fifteen studies while the off-diagonal elements were zero. Second, we numerically estimated $\beta$ and $\tau^2$'s and computed the null LRT statistics. Third, we estimated the scaled-shifted parameters ($a$, $b$, and $c$) of the corrected null distribution, $a \chi_1^2 + b \chi_2^2 + c$, that we selected based on its anti-conservativeness and controlled type I error rates by comparing 100,000 random $\chi_1^2$ and $\chi_2^2$ values to the computed null LRT statistics. With the estimated scaled-shifted parameters, we constructed an empirical distribution, $0.749 \chi_1^2 + 0.318 \chi_2^2 - 0.08$, with which we evaluated test statistics and computed p-values.

### 3.4.3 Meta-analysis Results

The Quantile-Quantile (Q-Q) plot with p-values computed from the empirical distribution against expected quantiles from 0 and 1, and the genome-wide association plot using the meta-analysis results from the fifteen studies in the MAGIC consortium are shown in
Figures 3.4 and 3.5. Table 3.3 lists all of variants identified in the meta-analysis with p-value $< 1 \times 10^{-6}$ using either our proposed LRT approach or the conventional FE approach. We restricted the meta-analysis results to variants present in at least five participating studies and with sample size greater than one-third of the total possible sample size. We found thirteen SNPs that had p-value $< 1 \times 10^{-6}$ for either our LRT or the FE approach. Among the thirteen variants, we found three SNPs located within genes, NMT2 and COBLL1, and one SNP 69kb downstream of TSHZ3 and the other SNPs in intergenic regions with IRS1 or PDGFC as the closest gene.

COBLL1, Cordon-Bleu WH2 Repeat Protein-Like 1, is a protein coding gene and is known to be associated with Type 2 diabetes in European-descent population (Albrechtsen et al. 2013), and fasting insulin and HOMA-IR levels in overweight and obese children (Mancina et al. 2013). Two SNPs, rs10490694 and rs6738627, in COBLL1 were identified that were different from the reported SNP in the Manning et al. study (Manning et al. 2012) with LD $R^2=1$ and 0.282, respectively. NMT2 (N-myristoyltransferase) is a novel locus that has not been previously found and is the only SNP that reaches genome-wide significance (rs11592551, p-value=$1.28 \times 10^{-8}$). NMT2 is a protein coding gene and Gene Ontology (GO) annotation related to this gene may include glycylpeptide N-tetradecanoyltransferase activity (Thinon et al. 2014).

In order to explore the performance of our approach compared to FE under relatively heterogeneous situations, we compared p-values of our approach and corresponding p-
values of FE for heterogeneous SNPs, defined by Q-statistic heterogeneity p-value < 0.05 or $I^2 > 50\%$. Although it is known to be low power, a significant Q-statistic (p-value < 0.05) may imply the presence of heterogeneity. $I^2$ measures the extent of heterogeneity and its value > 50% may indicate substantial heterogeneity (Higgins and Thompson 2002). **Figure 3.6** suggests that our approach generally gives lower p-values than FE model when substantial heterogeneity is present.
Table 3.3 Variants associated with Fasting Insulin adjusting for mean age and mean BMI at p-value less than $1 \times 10^{-6}$ of either our new LRT or FE models

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>rs number</th>
<th>Position (build 36)</th>
<th># Sample size</th>
<th>New LRT P-value $^a$</th>
<th>FE P-value $^a$</th>
<th>$I^2$ $^b$</th>
<th>Heterogeneity P-value $^b$</th>
<th>Gene $^c$</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>rs11592551</td>
<td>15243953</td>
<td>31694</td>
<td>1.3E-08</td>
<td>4.3E-09</td>
<td>51.33</td>
<td>0.02</td>
<td>NMT2$^*$</td>
</tr>
<tr>
<td>2</td>
<td>rs10490694</td>
<td>165256815</td>
<td>28629</td>
<td>3.7E-07</td>
<td>1.3E-07</td>
<td>27.38</td>
<td>0.18</td>
<td>COBLL1$^*$</td>
</tr>
<tr>
<td>19</td>
<td>rs1874492</td>
<td>36388218</td>
<td>35616</td>
<td>4.0E-07</td>
<td>3.4E-07</td>
<td>31.07</td>
<td>0.12</td>
<td>TSHZ3$^§$</td>
</tr>
<tr>
<td>2</td>
<td>rs6738627</td>
<td>165252696</td>
<td>28629</td>
<td>4.5E-07</td>
<td>1.6E-07</td>
<td>19.55</td>
<td>0.26</td>
<td>COBLL1$^*$</td>
</tr>
<tr>
<td>2</td>
<td>rs2972146</td>
<td>226808942</td>
<td>35616</td>
<td><strong>6.2E-07</strong></td>
<td>4.6E-06</td>
<td>54.13</td>
<td><strong>6.5E-03</strong></td>
<td>IRS1</td>
</tr>
<tr>
<td>2</td>
<td>rs2943645</td>
<td>226807424</td>
<td>35616</td>
<td><strong>6.4E-07</strong></td>
<td>4.6E-06</td>
<td>53.87</td>
<td><strong>6.8E-03</strong></td>
<td>IRS1</td>
</tr>
<tr>
<td>2</td>
<td>rs2943652</td>
<td>226816690</td>
<td>35616</td>
<td><strong>6.9E-07</strong></td>
<td>3.7E-06</td>
<td>53.01</td>
<td><strong>8.2E-03</strong></td>
<td>IRS1</td>
</tr>
<tr>
<td>2</td>
<td>rs2943650</td>
<td>226814165</td>
<td>35616</td>
<td><strong>7.0E-07</strong></td>
<td>3.2E-06</td>
<td>53.14</td>
<td><strong>7.9E-03</strong></td>
<td>IRS1</td>
</tr>
<tr>
<td>2</td>
<td>rs2972144</td>
<td>226809655</td>
<td>35615</td>
<td><strong>7.1E-07</strong></td>
<td>4.0E-06</td>
<td>54.43</td>
<td><strong>6.1E-03</strong></td>
<td>IRS1</td>
</tr>
<tr>
<td>2</td>
<td>rs2943646</td>
<td>226807778</td>
<td>35579</td>
<td><strong>8.0E-07</strong></td>
<td>5.3E-06</td>
<td>53.64</td>
<td><strong>7.2E-03</strong></td>
<td>IRS1</td>
</tr>
<tr>
<td>2</td>
<td>rs2138157</td>
<td>226811961</td>
<td>35616</td>
<td><strong>8.2E-07</strong></td>
<td>6.0E-06</td>
<td>51.24</td>
<td>0.01</td>
<td>IRS1</td>
</tr>
<tr>
<td>14</td>
<td>rs1642875</td>
<td>76114390</td>
<td>35616</td>
<td>1.1E-06</td>
<td>3.9E-07</td>
<td>0.00</td>
<td>0.85</td>
<td>ESRRB$^§$</td>
</tr>
<tr>
<td>4</td>
<td>rs13141804</td>
<td>157582999</td>
<td>35616</td>
<td>2.1E-06</td>
<td>7.7E-07</td>
<td>0.00</td>
<td>0.61</td>
<td>PDGFC</td>
</tr>
</tbody>
</table>

$^a$ New LRT, our new LRT method; FE, Fixed effects.

$^b$ $I^2$, the percentage of the variability in effect estimates that is due to heterogeneity rather than sampling error; Heterogeneity P-value, Q (Chi-square)-statistic p-value.

$^c$ Variants are located $^*$within gene, $^§$within 100kb of gene, and the rest of them are around 500kb in distance.
Figure 3.4 Quantile-Quantile (Q-Q) plot with p-values from the empirical distribution

![Quantile-Quantile (Q-Q) plot](image)

Figure 3.5 Genome-wide association plot from the fifteen MAGIC GWAS results using our new LRT method

![Genome-wide association plot](image)

The red dotted lines indicate $\alpha = 10^{-6}$ and $5 \times 10^{-8}$ significance levels
Figure 3.6 Comparison of our new LRT and FE models where substantial heterogeneity is present

(A) Heterogeneity, Q-statistic p-value is less than 0.05 which indicates heterogeneity in effect size among studies is present; (B) when $I^2$, the percentage of the variability in effect estimates that is due to heterogeneity rather than sampling error, is greater than 50%.
3.5 Discussion

We proposed a new statistical approach to accommodate multiple random effects introduced by study-level characteristics in the context of meta-analysis of potentially heterogeneous studies. As shown in previous simulation studies (B. Han and Eskin 2011; Hong et al.), the conventional RE model was less powerful than the FE model even when heterogeneity was present. Further, our proposed model improved power by assuming no association and no heterogeneity under the null hypothesis. Under the assumption that the sources of heterogeneity were related to study characteristics, our approach jointly tested the main effects and study-level characteristics that may influence the main effects. Our simulation study showed that our approach had greater power than the conventional FE and RE models and can be useful in identifying underlying effects associated with an outcome when heterogeneity in effects among studies in the meta-analysis exists and is considered in the analysis.

In the application to the genome-wide association analysis of fasting insulin from the fifteen studies from MAGIC consortium, our approach yielded eleven SNPs that had p-values < 10e-6, while FE found six SNPs. Two SNPs were found only using the FE model; these two SNPs showed no heterogeneity ($I^2=0$) among studies. When there is no heterogeneity, the FE model tends to perform better than any other random-effects model (Hong et al.).
As previously mentioned, the examination of the effect of study-level characteristics is of high interest to investigators in many fields including clinical trials and statistical genetics (Simmonds and Higgins 2007; Sterne et al. 2002; Greenland 1987; Manning et al. 2012). Among several approaches to explore interaction effects of the study characteristics, our approach jointly tests the main effect and the study-level covariate effects that may influence the main effect. This may, in practice, be similar to a joint meta-analysis of the main and interaction effects. However, one should remember that our approach considers only study-level characteristics rather than individual-level characteristics, and thus there is a possibility of aggregation bias (Greenland 1987). Aggregation (or ecological) bias arises when the relation between effect size and study-level characteristics is not the same as that between individual values of outcome and covariates. For example, the accountability of the variability of BMI among studies in a meta-analysis may not be the same as that of the variability of all of individuals in all of the studies. The study-level covariates are likely to misrepresent the individual measures. However, our approach is still useful especially when only summary measures are available. The joint meta-analysis (jointly testing main effects and interaction effects) or the meta-analysis of only interaction effects require the interaction effect estimates from each of the included studies. For meta-analysis of results extracted from literature or collected from studies conducted by different researchers, it may either be infeasible or take considerable time and effort to collect such data and the unavailability of the data restricts the study design to perform such meta-analyses. One advantage of our approach is that it does not require the access to individual data and can use summary data.
extracted from literature or a GWAS result repository as long as study-level measures of variables are available.

Although our approach is proposed in the general context of meta-analysis, it can be applied to a meta-analysis of ethnically diverse studies such as trans-ethnic meta-analysis of GWAS. Meta-analysis is commonly used in genetic studies and has been a highly successful tool for identifying common genetic variants associated with diseases and disease-associated quantitative traits. Because the FE model provides a pooled effect size and p-value, it has predominantly been used among European-descent studies. As the number of GWAS increases both within the European ancestry and across other ethnic ancestries, trans-ethnic meta-analysis, meta-analyzing over multiple GWAS across genetically diverse populations, suggests both opportunities and challenges. Because genetic architecture is distinctive in different ancestral populations, it is ambiguous how to better leverage different linkage disequilibrium, different spectra of causal variants and their allele frequencies. However, properly analyzing the evidence from diverse populations accounting for heterogeneity would improve power (Y. Li and Keating 2014; Morris 2011; Hong et al.) and would also provide valuable opportunities to explore genetic variants that are either transferable across populations or unique to a certain ethnicity (Liu et al. 2013; Liu et al. 2012) and enhance the ability to fine-map causal signals (Boerwinkle and Heckbert 2014; Liu et al. 2012; Franceschini et al. 2012). Because of this genetic heterogeneity in a trans-ethnic meta-analysis, the FE model may violate its homogeneity assumption and so should be avoided in application (Cantor,
In contrast, our approach takes potential heterogeneity into account in the form of study covariates and thus can be used in trans-ethnic meta-analysis of GWAS.

In summary, we proposed a statistical framework for the meta-analysis of potentially heterogeneous studies. Our approach accommodates multiple sources of heterogeneity related to study characteristics and generally achieves higher power than the FE model approach when heterogeneity in effect size across studies is substantial. Although it is somewhat limited in examining the cause of heterogeneity and challenging in interpretation because we only consider study-level measures rather than individual-level measures, our method can extend the ability of currently used FE and RE meta-analytic methods. We have also demonstrated that our method is computationally feasible to apply to data much larger than conventional clinical datasets such as genetics data. Because our approach allows for heterogeneity attributed to study characteristics, the application of our method can be extended beyond the typical synthesis of summary information of published data from literature and has the ability to meta-analyze over studies with big data such as genome-wide association studies with outcomes measured using different techniques, diverse populations, or multiple diseases with similar etiology.
Chapter 4  Score Test Approach to Statistical Model with Multiple Random Effects to Account for Heterogeneity in the Meta-Analysis

4.1  Introduction

Genome-wide association studies (GWAS) have achieved great success, however, the effect size of genetic variants is usually relatively modest and many studies with small to moderate sample size have lacked power to identify genetic determinants that are weakly associated with complex disease. The variants with weaker associations can be identified by increasing the sample size, and thus, power (McCarthy et al. 2008; Manolio et al. 2009). Meta-analysis is a practical approach to explore those weaker signals. The meta-analysis approach effectively improves power by increasing total sample size and is as efficient as combining individual participant data under most conditions (Lin and Zeng 2010). Hence, investigators provide summary statistics of the individual studies and meta-analyze them to detect those weaker signals with better power.

Although the meta-analysis approach is popular in genetics studies, it has been mostly restricted to studies of samples of European-ancestry. Because the genetic architecture is likely to be similar across samples having the same ancestral population, studies have commonly used inverse-variance-weighted fixed-effect (FE) meta-analysis. This model makes the assumption that there is no heterogeneity of effect size across contributing studies. However, as the number of GWAS in populations with a wide range of ancestries increases, investigators are interested in combining the summary statistics from as many
available studies as possible regardless of their ancestry. One advantage of combining studies with different genetic backgrounds is an increase in total sample size and hence a boost in power for novel discovery of associations between genetic variants and complex traits (Y. Li and Keating 2014). Combining studies with diverse ancestries also enhances fine-mapping of functional genetic variants by leveraging LD over different genetic structures (Liu et al. 2012; Liu et al. 2014; Boerwinkle and Heckbert 2014; Franceschini et al. 2012; Y. Li and Keating 2014).

When studies with diverse ancestries are meta-analyzed, the genetic architecture of the studies may not be homogeneous and the use of the FE meta-analysis is inappropriate. The random effect (RE) model accounts for the heterogeneity in the effect sizes, correctly estimates the between-study variance, and provides more appropriate confidence intervals for the pooled effect size estimate (Fleiss 1993; Hedges and Olkin 1985). However, the standard RE model has low power due to its conservative null hypothesis (Hedges and Olkin 1985; B. Han and Eskin 2011; Hong et al.) and its assumption of a single source of heterogeneity. Because variation across studies may arise due to multiple sources from complex study designs, populations, or different environmental exposures, assuming a single source of heterogeneity may not be realistic.

In chapter 3 we developed a new likelihood ratio test (LRT) that can accommodate multiple sources of heterogeneity using a linear mixed effects model. The LRT is formed by maximizing the likelihood function over both the restricted and unrestricted parameter
spaces (Casella and Berger 2002). If we consider many possible restrictions in the parameter space, the evaluation of the maximum likelihood may be inefficient and possibly intractable. These issues can be avoided by using a score test. The main advantage of the score test is that it requires the estimation of model parameters only under the null hypothesis, and it is computationally more efficient and asymptotically equivalent to the likelihood ratio test.

In this chapter, we propose a score test in the context of a meta-analysis evaluating the beta coefficients for genetic variants and the between-study variances pertinent to study covariates. This method provides a joint test of significance and a test of heterogeneity. We present a simulation study to compare this approach to the LRT developed in chapter 3.

4.2 Methods

4.2.1 Linear Mixed Effects Model with Multiple Random Effects

4.2.1.1 Notations and Assumptions

We consider the linear mixed-effect model and derive a score test statistic that accounts for heterogeneity. We assume that association results are gathered from \( N \) independent studies where genetic variants are either genotyped or imputed, and a trait of interest is also measured. Summary statistics assessing association between the trait and genetic variants, including the effect estimate \( (b_i) \) and its standard error \( (se(b_i)) \), are available
from these studies as well as study-level variables \((z_{ij}, i = 1, ..., N \text{ and } j = 1, ..., r; N \text{ is the number of studies and } r \text{ is the number of variables})\). We assume that the effect size across the studies may vary according to the study-level variables and heterogeneity in effect size may be introduced by these variables. We incorporate the heterogeneity by treating the variables as random effects using a linear mixed effects model. The model is

\[
b = \beta + Zv + e.
\]

The notations and model assumptions are the same as described in chapter 3. Note that \(b\) and \(\beta\) are vectors of observed effect estimates and a mean vector of the unknown true effect parameters, respectively. The \(Z\) is a matrix of study-level variable measures that are related to the random effects, \(v\).

\[
b = \begin{bmatrix} b_1 \\ \vdots \\ b_N \end{bmatrix}, \quad \beta = \beta \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}, \text{ and } Z = \begin{bmatrix} z_{11} & \ldots & z_{1r} \\ \vdots & \ddots & \vdots \\ z_{N1} & \ldots & z_{Nr} \end{bmatrix}.
\]

Generally, the random effects, \(v\), and the residual errors, \(e\), are assumed to be uncorrelated and to follow multivariate normal distributions with mean 0 and covariance matrices, \(\Sigma_v\) and \(R\), respectively,

\[
v \sim N(0, \Sigma_v),
\]

\[
e \sim N(0, R),
\]

where

\[
\Sigma_v = \text{var}[v] = \begin{bmatrix} \tau_1^2 & \ldots & \tau_{1r} \\ \vdots & \ddots & \vdots \\ \tau_{r1} & \ldots & \tau_r^2 \end{bmatrix} \text{ and } R = \begin{bmatrix} \sigma_1^2 & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \sigma_r^2 \end{bmatrix}.
\]

Because we assume that the random effects, elements of \(v\), are uncorrelated, the off-diagonal elements of the covariance matrix, \(\Sigma_v\), are zero \((\tau_{mn} = \tau_{nm} = 0, \text{ for all } m \neq n)\).
Also we assume that the estimated standard errors, \( se(b_i) \), of the effect estimates are true parameter values (Higgins and Thompson 2002) and so replace the within-study variances, \( \sigma_i^2 \) of the covariance matrix of the residual errors, \( e \). Under these assumptions, the notations can be simplified as

\[
\Sigma_v = \text{var}[\nu] = \begin{bmatrix}
\tau_1^2 & \cdots & 0 \\
\vdots & \ddots & \vdots \\
0 & \cdots & \tau_r^2
\end{bmatrix}
\quad \text{and} \quad
\hat{R} = \begin{bmatrix}
se(b_1)^2 & \cdots & 0 \\
\vdots & \ddots & \vdots \\
0 & \cdots & se(b_r)^2
\end{bmatrix},
\]

and the marginal distribution of \( b \) follows a multivariate normal distribution with mean \( \beta \) and covariance matrix \( W \),

\[
b \sim MN(\beta, W),
\]

where \( W = Z \Sigma_v Z' + R \). We then can obtain a likelihood function under the marginal distribution of \( b \) as follows,

\[
L(b, Z, R|\beta, \Sigma_v) = \frac{1}{(\sqrt{2\pi})^k |W|^{1/2}} \exp \left( - \frac{1}{2} (b - \beta)' W^{-1} (b - \beta) \right).
\]

We use this likelihood function to derive score functions, the information matrix and then the score test statistic.

4.2.1.2 Score Test Statistic

The usual form of the score function \( U(\theta) \) is the derivative of the log-likelihood with respect to the parameters of interest, that is, \( \beta, \tau_1^2, ..., \tau_r^2 \). In addition, the Fisher information \( I(\theta) \) provides an estimate of the covariance matrix of the score function \( U(\theta) \). The log-likelihood function is
\[
\begin{align*}
l(b, Z, R | \beta, \Sigma_v) &= C - \frac{1}{2} \ln |W| - \frac{1}{2} (b - \beta)' W^{-1} (b - \beta),
\end{align*}
\]

where \( C = -\frac{1}{k} \ln(\sqrt{2\pi}) \).

The score function can be obtained from the log-likelihood function by the first derivative of the log-likelihood function with respect to the parameters of interest

\[
U(\theta) = \frac{\partial l}{\partial \theta},
\]

where \( \theta = (\beta, \tau_1^2, ..., \tau_r^2) \). The score function has mean \( E[U(\theta)] = 0 \) and variance \( Var[U(\theta)] = I(\theta) \). The information matrix can also be obtained by the negative expectation of the second derivatives of the log-likelihood function with respect to the parameters of interest, \( (\beta, \tau_1^2, ..., \tau_r^2) \), as follows

\[
I(\theta_{pq}) = -E \left[ \frac{\partial^2 l}{\partial \theta_p \partial \theta_q} \right],
\]

where \( \theta_p \in \theta \) and \( \theta_q \in \theta \). We can obtain the first and second derivative of log-likelihood function as follows. We present detailed derivations in Appendix D. \( J_1 \) is a N-dimensional vector of 1’s.

\[
\frac{\partial^2 l}{\partial \beta^2} = -J_1' W^{-1} J_1,
\]

\[
\frac{\partial^2 l}{\partial \beta \partial \tau_i^2} = \frac{\partial^2 l}{\partial \tau_i^2 \partial \beta} = -(b - \beta)' (W^{-1} z_i z_i' W^{-1}) J_1, \quad i = 1, ..., r,
\]

\[
\frac{\partial^2 l}{\partial \tau_i^2 \partial \tau_j^2} = \frac{1}{2} tr \left( W^{-1} z_j z_j' W^{-1} z_i z_i' \right)
\]

\[
- \frac{1}{2} (b - \beta)' (W^{-1} z_j z_j' W^{-1} z_i z_i' W^{-1} + W^{-1} z_i z_i' W^{-1} z_j z_j' W^{-1})(b - \beta),
\]
\[ i = 1, \ldots, r \text{ and } j = 1, \ldots, r.\]

Hence the vector of score functions and the information matrix for our parameters are

\[ U(\theta) = \begin{bmatrix} U(\beta) \\ U(\tau_i^2) \\ \vdots \\ U(\tau_r^2) \end{bmatrix}, \]

where

\[ U(\beta) = (b - \beta)'W^{-1}f, \]

\[ U(\tau_i^2) = -\frac{1}{2} tr(W^{-1}z_i'z_i) + \frac{1}{2} (b - \beta)'(W^{-1}z_i'W^{-1})(b - \beta), \quad i = 1, \ldots, r, \]

and

\[ I(\theta) = -E \begin{bmatrix} \frac{\partial^2 l}{\partial \beta^2} & \frac{\partial^2 l}{\partial \beta \partial \tau_i^2} & \cdots & \frac{\partial^2 l}{\partial \beta \partial \tau_r^2} \\ \frac{\partial^2 l}{\partial \tau_i^2 \partial \beta} & \frac{\partial^2 l}{\partial \tau_i^2 \partial \tau_1^2} & \cdots & \frac{\partial^2 l}{\partial \tau_i^2 \partial \tau_r^2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 l}{\partial \tau_r^2 \partial \beta} & \frac{\partial^2 l}{\partial \tau_r^2 \partial \tau_1^2} & \cdots & \frac{\partial^2 l}{\partial \tau_r^2 \partial \tau_r^2} \end{bmatrix}, \]

We are interested in testing the null hypothesis \( H_0 : \beta = 0, \tau_1^2 = 0, \ldots, \text{and } \tau_r^2 = 0 \), versus the alternative hypothesis \( H_1 : \beta \neq 0, \tau_1^2 > 0, \ldots, \text{or } \tau_r^2 > 0 \). We construct a score test statistic using the derived score vector and the information matrix. The main advantages of the score test are that it requires the estimation of the parameters only under the null
hypothesis and it is asymptotically equivalent to the likelihood ratio test. Under the null hypothesis, we can simplify our score function and information matrix as follows

\[
U(\theta_0) = \begin{bmatrix} U(\beta_0) \\ U(\tau_{i0}^2) \\ \vdots \\ U(\tau_{r0}^2) \end{bmatrix},
\]

where

\[
U(\beta_0) = b' R^{-1} J_1,
\]

\[
U(\tau_{i0}^2) = -\frac{1}{2} tr(R^{-1} z_i z_i') + \frac{1}{2} b'(R^{-1} z_i z_i' R^{-1}) b, \quad i = 1, ..., r,
\]

and

\[
I(\theta_0) = \begin{bmatrix} -E \left[ \frac{\partial^2 l}{\partial \beta^2} \right]_{\theta_0} & -E \left[ \frac{\partial^2 l}{\partial \beta \partial \tau_1^2} \right]_{\theta_0} & \cdots & -E \left[ \frac{\partial^2 l}{\partial \beta \partial \tau_r^2} \right]_{\theta_0} \\
-E \left[ \frac{\partial^2 l}{\partial \tau_1^2 \partial \beta} \right]_{\theta_0} & -E \left[ \frac{\partial^2 l}{\partial \tau_1^2 \partial \tau_1^2} \right]_{\theta_0} & \cdots & -E \left[ \frac{\partial^2 l}{\partial \tau_1^2 \partial \tau_r^2} \right]_{\theta_0} \\
-E \left[ \frac{\partial^2 l}{\partial \tau_2^2 \partial \beta} \right]_{\theta_0} & -E \left[ \frac{\partial^2 l}{\partial \tau_2^2 \partial \tau_1^2} \right]_{\theta_0} & \cdots & -E \left[ \frac{\partial^2 l}{\partial \tau_2^2 \partial \tau_r^2} \right]_{\theta_0} \\
\cdots & \cdots & \ddots & \cdots \\
-E \left[ \frac{\partial^2 l}{\partial \tau_r^2 \partial \beta} \right]_{\theta_0} & -E \left[ \frac{\partial^2 l}{\partial \tau_r^2 \partial \tau_1^2} \right]_{\theta_0} & \cdots & -E \left[ \frac{\partial^2 l}{\partial \tau_r^2 \partial \tau_r^2} \right]_{\theta_0},
\end{bmatrix}
\]

where

\[
-E \left[ \frac{\partial^2 l}{\partial \beta^2} \right]_{\theta_0} = J_1' R^{-1} J_1,
\]

\[
-E \left[ \frac{\partial^2 l}{\partial \beta \partial \tau_i^2} \right]_{\theta_0} = -E \left[ \frac{\partial^2 l}{\partial \tau_i^2 \partial \beta} \right]_{\theta_0} = 0, \quad i = 1, ..., r.
\]
\[-E \left[ \frac{\partial^2 l}{\partial \tau_i^2 \partial \tau_j^2} \right]_{\theta_0} = \frac{1}{2} tr \left( R^{-1} z_i z_j' R^{-1} z_i' \right), \quad i = 1, \ldots, r \text{ and } j = 1, \ldots, r. \]

The vector $b$ contains the observed effect estimates, the vectors $z_i$ and $z_j$ have study-level measures, and the within-study covariance matrix $R$ has the squared standard errors of the effect estimates for the diagonal elements and zero for the off-diagonal values. All these values are assumed to be known as described in 4.2.1.1. The score test statistic is thus

\[ T_s = U'(\theta_0) I(\theta_0)^{-1} U(\theta_0), \]

and the score vector $U(\theta_0)$ and the information matrix $I(\theta_0)$ are defined in the previous paragraph.

**4.2.2 Asymptotic Distribution**

The statistical significance of the score test $T_s$ could be assessed using its asymptotic distribution. For testing the null hypothesis, $H_0: \beta = 0, \tau_1^2 = 0, \ldots, \text{ and } \tau_r^2 = 0$, the parameter $\beta$ is unbounded and the asymptotic distribution of a score test for $\beta$ is a $\chi_1^2$ distribution. The parameter $\tau^2$ is bounded to be non-negative in the parameter space; however, $\tau$ is unbounded and thus the asymptotic distribution of a score test for $\tau$ is a $\chi_1^2$ distribution (Boos 1992). If the score test $T_s$ is constructed with respect to all parameters that are unbounded, its asymptotic distribution follows a $\chi_g^2$ distribution with $g$ degrees
of freedom determined by the number of parameters specified under the null hypothesis (Boos 1992). We derive the score functions and information matrix with respect to both $\tau^2$ and $\tau$ in Appendix E and show that the test statistics derived by both parameters are the same (Dupuis, Siegmund, and Yakir 2007b; Kumar 1973; Misra 1972). This result implies that our score test $T_s$ can also test $H_0: \beta = 0, \tau_1 = 0, \ldots, \tau_r = 0$, and therefore, we may reasonably assume that our score test statistic $T_s$ asymptotically follows a $\chi^2_g$ distribution with $g = 1 + r$ degrees of freedom.

4.3 Simulation Study

4.3.1 Study Design of Simulation

In the simulation study, we evaluated the type I error rates and power of the proposed score test statistic. We considered the case of two random effects throughout the simulation study. The two random effects were represented by two study-level covariates. We considered four different combinations of the covariates in a model; two continuous, two dichotomous, two proportional variables, and a combination of a continuous and a dichotomous variables. We varied the number of studies to evaluate the type I error rates and additionally varied the effect size and the size of between-study variances to evaluate power. We compared the performance of the score test to the LRT approach from chapter 3.
4.3.1.1 Type I error Simulation

We evaluated the type I error rates under the null hypothesis by setting all of the parameters of interest to zero, that is, \( H_0: \beta = 0, \tau_1^2 = 0, \tau_2^2 = 0 \). The overall steps of data simulation were similar to those in chapter 3. We first simulated the error variance \( \sigma_i^2 \) from a \( \chi^2_1 \) distribution and generated observed effect sizes from \( b_i \sim N(0, \sigma_i^2) \) for the \( i^\text{th} \) study. We excluded extreme values of \( \sigma_i^2 \) that were less than 20\textsuperscript{th} percentile or greater than 80\textsuperscript{th} percentile in order to mimic the practical range of variances in genetic studies such that the sample size ratio of small sample sized study (300) to relatively large sample sized study (10,000) is around 30. The study-level covariates \( z_{ij} \) were generated according to the variable types. The continuous variables were generated from \( N\left(0, \frac{1}{12}\right) \). Here, we made the distribution for continuous covariates similar to that of the proportion covariates, Uniform(0,1), with mean 0.5 and variance \( \frac{1}{12} \). The dichotomous variables were generated from Bernoulli(1, \( p = 0.5 \)). The proportion variables were generated from Uniform(0, 1). We varied the number of studies \( N \) from 10, 20, 50 and 100 to evaluate the type I error rates. We computed the score test statistics \( T_s \) and evaluated p-values from the asymptotic distribution (Table 4.1). However, we observed highly inflated type I error rates. Our attempts to obtain controlled type I error rates are described in Appendix F. We decided ultimately to use an empirical distribution.
Table 4.1 Type I error rates evaluated from a $\chi^2_3$ distribution

<table>
<thead>
<tr>
<th></th>
<th>Continuous</th>
<th>Continuous+Dichotomous</th>
<th>Dichotomous</th>
<th>Proportional</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha=0.05$</td>
<td>0.073</td>
<td>0.079</td>
<td>0.078</td>
<td>0.079</td>
</tr>
<tr>
<td>$\alpha=0.01$</td>
<td>0.042</td>
<td>0.047</td>
<td>0.050</td>
<td>0.052</td>
</tr>
</tbody>
</table>

4.3.1.2 Empirical distribution

We constructed an empirical distribution from additional null simulations. We generated a hundred null effect sizes from each replicate, $b_{emp} \sim N(0, \sigma_i^2)$. So we obtained a hundred null score statistics $T_{emp}$ for each simulation replicate (or SNP) and aggregated the computed null score statistics to construct the empirical null distribution. The p-value is computed as the proportion of empirical score values that are greater or equal to the observed score test.

4.3.1.3 Power Simulation

We simulated data for power in a similar manner to the type I error simulation. We generated the error variance $\sigma_i^2$ from a $\chi^2_1$ distribution for the elements of the error covariance matrix $R$, and pre-specified true effect sizes $\beta_i, i = 1, \ldots, N$, between-study variances $\tau_j^2, j = 1, 2$ for each $N$. We generated the design matrix $Z$ for the study-level covariates where each element $z_{ij}$ of the design matrix $Z$ was generated according to the variable types described in the type I error simulation. Using these values, we generated the observed effect sizes from a multivariate normal distribution $b \sim MN(\beta, W)$ where $\beta$ was a vector of the pre-specified true effect sizes and $W$ was calculated from $W =$
\[ \mathbf{Z} \mathbf{\Sigma}_{v} \mathbf{Z}' + \mathbf{R}, \text{ and } \mathbf{\Sigma}_{v} = \begin{bmatrix} \tau_{1}^{2} & 0 \\ 0 & \tau_{2}^{2} \end{bmatrix} \]

We computed the score test statistic \( T_{s} \) and simultaneously computed the null score statistics \( T_{emp} \) to construct the empirical null distribution as described in the previous section. We calculated the p-value from the corresponding empirical null distribution.

We evaluated power in three heterogeneity scenarios testing \( H_{0}: \beta = 0, \tau_{1}^{2} = 0, \tau_{2}^{2} = 0 \) versus \( H_{1}: \beta \neq 0, \tau_{1}^{2} \neq 0, \text{ or } \tau_{2}^{2} \neq 0 \). In scenario 1 (Effect-size Homogeneity scenario), there is no heterogeneity between studies: the true between-study variances \( \tau_{1}^{2} \) and \( \tau_{2}^{2} \) are zero while \( \beta \) is non-zero. In this scenario, we generated the observed effect sizes from \( \mathbf{b} \sim \mathcal{MN}(\mathbf{\beta}, \mathbf{R}) \). In scenario 2 (Two Random Effects scenario), the between-study variances \( \tau_{1}^{2} \) and \( \tau_{2}^{2} \) are non-zero. Hence we generated the observed effect sizes from \( \mathbf{b} \sim \mathcal{MN}(\mathbf{\beta}, \mathbf{W}) \). In scenario 3 (One Random Effects scenario), the between-study variance \( \tau_{1}^{2} \) was non-zero while \( \tau_{2}^{2} \) was zero. In this scenario, we generated the observed effect sizes from \( \mathbf{b} \sim \mathcal{MN}(\mathbf{\beta}, \mathbf{W}) \), where \( \mathbf{W} = \mathbf{Z} \mathbf{\Sigma}_{v} \mathbf{Z}' + \mathbf{R}, \text{ and } \mathbf{\Sigma}_{v} = \begin{bmatrix} \tau_{1}^{2} & 0 \\ 0 & \tau_{2}^{2} \end{bmatrix} \). Under each scenario, we generated study-level covariates as described in 4.3.1.1. We additionally varied the effect sizes \( \beta \) from 0.01 to 0.05, 0.1, 0.15 and 0.2, the size of between-study variance \( \tau_{1}^{2} \) from 0.05 to 0.1, 0.2, 0.5, 1 and 2, and the number of studies \( N \) as 10, 20, 50 and 100.
4.3.2 Simulation Results

4.3.2.1 Type I error Simulation

Figure 4.1 shows the type I error rates of the score test at $\alpha = 0.001$ and 95% confidence intervals for the four different combinations of the study covariates varying the number of studies. The type I error rates were calculated by the proportion of simulations that passed the p-value less than or equal to the $\alpha$ level. For all scenarios, the 95% confidence intervals covered the $\alpha$ level.

Figure 4.1 Empirical type I error rates and its 95% confidence intervals for the four different combinations of the study covariates at $\alpha = 0.001$

CC, two continuous variables; CD, a combination of continuous and dichotomous variables; DD, two dichotomous variables; PP, two proportions
4.3.2.2 Power Simulation

We evaluated power of the score test and compared it to that of the LRT approach in which p-values were calculated using both asymptotic (LRT only) and corrected null distributions from chapter 3. Figure 4.2 shows the Effect-size Homogeneity scenario (Scenario 1) results. In this scenario, power of the score test improved when the number of studies or the effect size increased. When the covariates were proportions or continuous variables power of the score test was similar to the LRT approach using the correct null distribution. When the covariates were dichotomous or the combination of the continuous and dichotomous variables, power of the score test was a little bit below the LRT with corrected distribution. When the covariates were dichotomous variables, the power of the score test was greater than that of LRT with the asymptotic distribution and less than that of LRT with the corrected distribution when $N = 50$ or $\beta = 0.1$ but the power of the score test became similar to that of LRT with the asymptotic distribution when $N = 100$ or $\beta = 0.2$.

Figure 4.3 shows the power results under the Two Random Effects scenario (Scenario 2) where we introduce heterogeneity in effect size across studies through the study covariates. When the covariates were proportions, the score test showed similar power to the LRT approach using the correct null distribution. When the covariates were dichotomous the power gain with increasing parameter values ($N, \beta$, or $\tau_1^2$ and $\tau_2^2$) was lower for the score test than LRT approach. When the covariates involved continuous variables, either two continuous variables or the combination of continuous and
dichotomous variables, the score test was less powerful than either LRT approach. The One Random Effects scenario (Scenario 3) showed similar results as in scenario 2. The results of scenario 3 are shown in Figure 4.4.
Figure 4.2 Power of the score test and LRT tests under the Effect-size Homogeneity scenario (Scenario 1)

(A) Power Varying NUMBER OF STUDY (Set $\beta = 0.1$)

(B) Power Varying EFFECT SIZE (Set $N = 50$)
Figure 4.3 Power of score test and LRT tests under the Two Random Effects scenario (Scenario 2)
Figure 4.4 Power of score test and LRT tests under the One Random Effects scenario (Scenario 3)

(A) Power Varying NUMBER OF STUDY (set $\beta = 0.1$, $\tau_1^2 = 0.05$, $\tau_2^2 = 0$)

(B) Power Varying EFFECT SIZE (set $N = 50$, $\tau_1^2 = 0.05$, $\tau_2^2 = 0$)

(C) Power Varying BETWEEN-STUDY VARIANCE (set $N = 50$, $\beta = 0.1$)
4.4 Discussion

In this chapter, we proposed a score test approach for meta-analysis of heterogeneous studies when the heterogeneity in effect size is explained by study-level characteristics. Because we assume no main effect and no heterogeneity under the null hypothesis, the proposed test statistic jointly tests the main effect and the heterogeneity pertinent to study covariates. The proposed approach only requires summary statistics of SNPs and study-level covariate measures from each study. We have explored different types of study covariates on the performance of the score test in the simulation study.

Compared with the LRT approach using the corrected null distribution, the score test approach has lower power under most of the scenarios. The power of the score test is similar to the LRT when both covariates are continuous or proportions under the Effect-size Homogeneity scenario (Scenario 1), or when the covariates are proportions under the heterogeneity scenarios (Scenarios 2 and 3). The score test achieves lower power than the LRT for the other scenarios and the difference in power increases with increasing parameters. Note that in the One Random Effects scenario (Scenario 3) we generate data assuming only one of the between-study variances is non-zero but test the null hypothesis that both between-study variances and the SNP main effect is zero. Hence, our analysis model is not the same as we generate the data. However, we observe similar results in the scenario 3 as in the Two Random Effects scenario (Scenario 2), demonstrating our proposed model is robust to such a situation.
We derived the score functions and information matrix with respect to both $\tau$ and $\tau^2$ (Appendix E), and showed that both score tests follow a $\chi^2$ distribution with degrees of freedom determined by the number of parameters we specify to be equal to 0 under the null hypothesis. We showed that the type I error rates are inflated due to some influential points with small variances and relatively large effect sizes that cause inflation in the score statistics. This inflation led us to use the empirical null distribution that yields well controlled type I error rates (Appendix F).

Future work may identify alternative ways to correct the null distribution. One possible alternative to the empirical null distribution approach is a shrinkage estimation for dispersions. An Empirical Bayes shrinkage method has been proposed and implemented to shrink a dispersion parameter in the context of RNA-seq data analysis (Wu, Wang, and Wu 2013; Love, Huber, and Anders 2014). If the estimated variances shrink towards a mean estimate we might be able to avoid the inflation in the score statistics due to the influential points with small variances.

Although the proposed test has slightly lower power than the LRT approach developed in chapter 3, the computational advantage (Table 4.2) is substantial despite the need for the empirical null distribution. With this evident computational efficiency, one might suggest a two-stage approach such that one uses the score test for genome-wide scan and computes the LRT for any detected significant SNPs. This can be an alternative to utilize both score and LRT statistics for large data analysis such as genome-wide studies.
In summary, we proposed a score test approach for the meta-analysis of potentially heterogeneous studies. Because it estimates the parameters only under the null hypothesis, the computational advantage of the score test is evident and the estimation of the test statistic is rapid. Given the power of the score test is similar to that of the LRT for some of the scenarios especially when the covariates are proportions and with its computational advantage, the score test with some modification can be an alternative approach in jointly testing the effect size and the heterogeneity for the meta-analysis of heterogeneous studies.

Table 4.2 Average computational time (in minutes) for the calculation of the LRT and the score test statistics for 10,000 simulations

<table>
<thead>
<tr>
<th>Number of Studies (N)</th>
<th>10</th>
<th>20</th>
<th>50</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>LRT</td>
<td>17.1</td>
<td>20.4</td>
<td>44.7</td>
<td>171.8</td>
</tr>
<tr>
<td>Score</td>
<td>1.3</td>
<td>1.2</td>
<td>1.4</td>
<td>2.5</td>
</tr>
</tbody>
</table>

We used Intel® Xeon® Processor E5-2650 v2 (20M Cache, 2.60 GHz) for this comparative computational time analysis.
Chapter 5 Conclusion and Future Work

Meta-analysis is an effective statistical approach in genetic association studies. It has been successful in identifying novel variants. However, until recently, the lack of diversity of ancestry in GWA studies has led to the application of meta-analysis primarily for single ancestry, predominantly European-descent population studies. With increasing number of GWAS in non-European-ancestry populations, we believe that properly aggregating evidence from multiple studies with ethnically diverse populations would provide a great opportunity to expand our understanding of disease etiology. New statistical approaches are required to tackle challenges in such meta-analyses.

In this dissertation, we proposed a statistical strategy and two new meta-analytic models in the context of meta-analysis of potentially heterogeneous studies. In chapter 2, we proposed a two-stage approach to account for heterogeneity in trans-ethnic meta-analysis in which we clustered studies with cohort-specific ancestry information prior to meta-analysis using a hierarchical clustering algorithm. In the extensive simulation study, we compared this approach to a no clustering (one-stage), evaluating type I error and power of these two strategies. We showed that the two-stage approach could be an effective and efficient intermediate step in meta-analysis to account for heterogeneity due to studies with varying ancestries. From our simulation studies, the two-stage approach is robust to incorrect clustering by taking a sub-optimal dissimilarity value for the threshold. However, the currently implemented fixed-threshold approach is subjective and might be
potentially prone to incorrect clustering. Therefore, one potential future direction for this work is to develop a better, more objective mechanism or measurement, such as a data-driven approach to find an optimal threshold for clustering.

The source of heterogeneity in meta-analysis is not limited to the difference in genetic backgrounds of samples. Heterogeneity could arise from different study designs or environmental exposures. Incorporating such heterogeneity in meta-analysis is crucial. In chapters 3 and 4, we developed new meta-analytic models that could be used in the meta-analysis of heterogeneous studies. In chapter 3, we extended the Han and Eskin (B. Han and Eskin 2011) random-effects approach under a meta-regression framework and developed a new likelihood ratio test (LRT) statistic to accommodate multiple random effects. We showed that our new LRT was comparable to the other methods in the simulation study, especially when heterogeneity in effect sizes was present. The new LRT was successfully applied in a meta-analysis of genome-wide association studies from the Meta-Analyses of Glucose- and Insulin-related traits Consortium with a log-transformed fasting insulin with study-level mean body mass index and age as covariates. The new LRT was developed for common genetic variant analyses. Because of increasing availability of rare variant analyses as described in chapter 1.1 and the advantage of multiple ancestries for meta-analysis of rare variants association tests (Mensah-Ablorh et al. 2016), future work may involve the extension of our approach to rare variant meta-analyses of potentially heterogeneous studies.
In chapter 4, we developed a score test statistic in the same context as in chapter 3, hoping to improve computational efficiency and numerical stability. We observed that the type I error rates using an asymptotic distribution were inflated due to some influential points with small variances and relatively large effect sizes that caused inflated score statistics. The empirical null distribution yielded controlled type I error rates; however, power for the score test was lower than for the corrected LRT in chapter 3. A possible alternative to the empirical null distribution we implemented is a shrinkage approach. Empirical Bayes shrinkage methods have been proposed and implemented to shrink a dispersion parameter in the context of RNA-seq data analysis (Wu, Wang, and Wu 2013; Love, Huber, and Anders 2014). If the estimated variances shrink towards a mean estimate we might be able to avoid the inflation in the score statistics due to the influential points with small variances.

We have described three projects related to the meta-analysis of heterogeneous studies and demonstrated that the proposed approach and models could effectively and efficiently aggregate the evidence from heterogeneous studies and improve the ability to detect genetic variants with weaker associations that could have been missed in the traditional genetic association studies.
Appendices

Appendix A Derivation of Maximum Likelihood Estimators for $\beta$ and $\tau^2_i$

The log-likelihood function of the likelihood we obtained in chapter 3.2.1.2 under the union of $\theta \in \theta_0 \cup \theta_1$ is

$$l = C - \frac{1}{2} ln|W| - \frac{1}{2} (b - \beta)'W^{-1}(b - \beta),$$

(1)

where $C = -\frac{1}{k} ln(\sqrt{2\pi})$ and $W = Z\Sigma_vZ' + R$ and

$b = N$ by 1 vector of observed effects from $N$ studies,

$\beta = N$ by 1 vector of the unknown true effect size of $N$ studies,

$Z = N$ by $r$ design matrix for covariates for unknown random effects,

$\Sigma_v = r$ by $r$ diagonal covariance matrix for unknown random effects,

$R = N$ by $N$ of a diagonal residual covariance matrix for $N$ studies.

The maximum likelihood estimators (MLEs) of the true effect size $\beta$ and the variances of the random effects $\tau^2_j$, for $j = 1, ..., r$, can be derived by setting the first derivative of the log-likelihood function with respect to the parameters to zero.

The MLE of the $\beta$ is
\[
\frac{\partial l}{\partial \beta} = (b - \beta)'W^{-1}j_1 \equiv 0
\]

\[\iff b'W^{-1}j_1 = \beta'j_1W^{-1}j_1\]

\[\iff \hat{\beta} = b'W^{-1}j_1(j_1'W^{-1}j_1)^{-1}\]

where \(j_1\) is a \(N\) dimensional vector of 1’s.

The MLE of the \(\tau_j^2\) is

\[
\frac{\partial l}{\partial \tau_j^2} = -\frac{1}{2} \left( tr(W^{-1} \frac{\partial W}{\partial \tau_j^2}) - \frac{1}{2} \left( (b - \hat{\beta})' \left( -W^{-1} \frac{\partial W}{\partial \tau_j^2} W^{-1} \right) (b - \hat{\beta}) \right) \right) \equiv 0
\]

\[\iff -\frac{1}{2} \left( tr(W^{-1}z_jz_j') \right) - \frac{1}{2} \left( (b - \hat{\beta})' \left( -W^{-1}z_jz_j'W^{-1} \right) (b - \hat{\beta}) \right) = 0\]

\[\iff -\frac{1}{2} \left( tr(W^{-1}z_jz_j') \right) - \frac{1}{2} \left( (b - \hat{\beta})' (-W^{-1}z_jz_j'W^{-1})(b - \hat{\beta}) \right) = 0\]

\[\iff tr(W^{-1}z_jz_j') - (b - \hat{\beta})'(W^{-1}z_jz_j'W^{-1})(b - \hat{\beta}) = 0\]

A closed-form solution for the maximum likelihood estimator of the true effect size \(\beta\) can be obtained, however, there is no closed-form for the variances of the random effects \(\tau_j^2\), for \(j = 1, \ldots, r\).
### Appendix B Candidate “Correct” Null Distribution Comparison

Table B.1 Type I error comparison for candidate “corrected” null distribution for N=10

<table>
<thead>
<tr>
<th>α levels</th>
<th>Study Characteristics</th>
<th>N=10</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>α=0.01</td>
<td>CONT+</td>
<td>0.008</td>
<td>0.008</td>
<td>0.011</td>
<td>0.011</td>
<td>0.012</td>
<td>0.012</td>
<td>0.011</td>
<td>0.009</td>
</tr>
<tr>
<td>CONT</td>
<td></td>
<td>0.009</td>
<td>0.009</td>
<td>0.012</td>
<td>0.012</td>
<td>0.014</td>
<td>0.013</td>
<td>0.012</td>
<td>0.010</td>
</tr>
<tr>
<td>DICH+</td>
<td>0.009</td>
<td>0.009</td>
<td>0.013</td>
<td>0.012</td>
<td>0.014</td>
<td>0.013</td>
<td>0.012</td>
<td>0.010</td>
<td>0.010</td>
</tr>
<tr>
<td>DICH</td>
<td>0.009</td>
<td>0.009</td>
<td>0.012</td>
<td>0.012</td>
<td>0.014</td>
<td>0.013</td>
<td>0.010</td>
<td>0.009</td>
<td>0.009</td>
</tr>
<tr>
<td>PROP+</td>
<td>0.008</td>
<td>0.008</td>
<td>0.0014</td>
<td>0.0013</td>
<td>0.0020</td>
<td>0.0019</td>
<td>0.0011</td>
<td>0.0008</td>
<td>0.0007</td>
</tr>
<tr>
<td>PROP</td>
<td>0.0006</td>
<td>0.0006</td>
<td>0.0015</td>
<td>0.0015</td>
<td>0.0023</td>
<td>0.0021</td>
<td>0.0012</td>
<td>0.0009</td>
<td>0.0009</td>
</tr>
<tr>
<td>CONT+</td>
<td>0.0008</td>
<td>0.0008</td>
<td>0.0017</td>
<td>0.0017</td>
<td>0.0025</td>
<td>0.0022</td>
<td>0.0012</td>
<td>0.0010</td>
<td>0.0010</td>
</tr>
<tr>
<td>DICH</td>
<td>0.0008</td>
<td>0.0008</td>
<td>0.0016</td>
<td>0.0016</td>
<td>0.0024</td>
<td>0.0022</td>
<td>0.0011</td>
<td>0.0009</td>
<td>0.0009</td>
</tr>
</tbody>
</table>

* Types of study characteristics as the sources of heterogeneity: CONT-Continuous, DICH-dichotomous, PROP-proportion, CONT+DICH-a combination of continuous and dichotomous covariates.

Scaled-shifted parameters ($a$, $b$ and/or $c$) are estimated using $^b$ a regression, $^c$ method of moments, and $^d$ a numerical method.
Table B.2 Type I error comparison for candidate “corrected” null distribution for N=20

<table>
<thead>
<tr>
<th>α levels</th>
<th>Study Characteristics</th>
<th>N=20</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>α=0.01</td>
<td>CONT+CONT</td>
<td>0.008</td>
<td>0.010</td>
</tr>
<tr>
<td></td>
<td>DICH+DICH</td>
<td>0.008</td>
<td>0.009</td>
</tr>
<tr>
<td></td>
<td>PROP+PROP</td>
<td>0.009</td>
<td>0.010</td>
</tr>
<tr>
<td></td>
<td>CONT+DICH</td>
<td>0.010</td>
<td>0.010</td>
</tr>
<tr>
<td>α=0.001</td>
<td>CONT+CONT</td>
<td>0.0005</td>
<td>0.0009</td>
</tr>
<tr>
<td></td>
<td>DICH+DICH</td>
<td>0.0006</td>
<td>0.0009</td>
</tr>
<tr>
<td></td>
<td>PROP+PROP</td>
<td>0.0007</td>
<td>0.0011</td>
</tr>
<tr>
<td></td>
<td>CONT+DICH</td>
<td>0.0009</td>
<td>0.0010</td>
</tr>
</tbody>
</table>

Types of study characteristics as the sources of heterogeneity; CONT-Continuous, DICH-dichotomous, PROP-proportion, CONT+DICH-a combination of continuous and dichotomous covariates.

Scaled-shifted parameters \((a, b)\) and/or \((c)\) are estimated using \(^b\) a regression, \(^c\) method of moments, and \(^d\) a numerical method.
Table B.3 Type I error comparison for candidate “corrected” null distribution for \( N=50 \)

<table>
<thead>
<tr>
<th>( \alpha ) levels</th>
<th>Study Characteristics</th>
<th>( a \chi^2_1 + b )</th>
<th>( a \chi^2_1 + b )</th>
<th>( a \chi^2_2 + b )</th>
<th>( a \chi^2_3 + b )</th>
<th>( a \chi^2_2 + b )</th>
<th>( b \chi^2_2 + c )</th>
<th>( a \chi^2_1 + b \chi^2_2 + c )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \alpha=0.01 )</td>
<td>CONT+</td>
<td>0.008</td>
<td>0.008</td>
<td>0.012</td>
<td>0.012</td>
<td>0.013</td>
<td>0.013</td>
<td>0.012</td>
</tr>
<tr>
<td></td>
<td>CONT</td>
<td>0.008</td>
<td>0.008</td>
<td>0.012</td>
<td>0.012</td>
<td>0.013</td>
<td>0.013</td>
<td>0.011</td>
</tr>
<tr>
<td></td>
<td>DICH+</td>
<td>0.009</td>
<td>0.009</td>
<td>0.012</td>
<td>0.012</td>
<td>0.013</td>
<td>0.013</td>
<td>0.011</td>
</tr>
<tr>
<td></td>
<td>DICH</td>
<td>0.009</td>
<td>0.009</td>
<td>0.013</td>
<td>0.012</td>
<td>0.014</td>
<td>0.013</td>
<td>0.011</td>
</tr>
<tr>
<td></td>
<td>CONT+DICH</td>
<td>0.0006</td>
<td>0.0006</td>
<td>0.0014</td>
<td>0.0014</td>
<td>0.0022</td>
<td>0.0021</td>
<td>0.0012</td>
</tr>
<tr>
<td></td>
<td>CONT+</td>
<td>0.0006</td>
<td>0.0006</td>
<td>0.0016</td>
<td>0.0016</td>
<td>0.0024</td>
<td>0.0022</td>
<td>0.0013</td>
</tr>
<tr>
<td></td>
<td>DICH+</td>
<td>0.0007</td>
<td>0.0007</td>
<td>0.0017</td>
<td>0.0017</td>
<td>0.0026</td>
<td>0.0024</td>
<td>0.0012</td>
</tr>
<tr>
<td></td>
<td>DICH</td>
<td>0.0007</td>
<td>0.0007</td>
<td>0.0019</td>
<td>0.0018</td>
<td>0.0028</td>
<td>0.0025</td>
<td>0.0010</td>
</tr>
</tbody>
</table>

\( ^a \) Types of study characteristics as the sources of heterogeneity: CONT-Continuous, DICH-dichotomous, PROP-proportion, CONT+DICH-a combination of continuous and dichotomous covariates.

Scaled-shifted parameters (\( a, b \) and/or \( c \)) are estimated using \( ^b \) a regression, \( ^c \) method of moments, and \( ^d \) a numerical method.
Table B.4 Type I error comparison for candidate “corrected” null distribution for N=100

<table>
<thead>
<tr>
<th>$\alpha$ levels</th>
<th>Study Characteristics a</th>
<th>$a \chi_1^2 + b$ b</th>
<th>$a \chi_1^2 + b$ c</th>
<th>$a \chi_2^2 + b$ b</th>
<th>$a \chi_3^2 + b$ c</th>
<th>$a \chi_3^2$ + $b \chi_2^2$ + $c \chi_3^2$ d</th>
<th>$a \chi_1^2$ + $b \chi_2^2$ + $c$ b</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha=0.01$</td>
<td>CONT+</td>
<td>0.008</td>
<td>0.008</td>
<td>0.012</td>
<td>0.011</td>
<td>0.013</td>
<td>0.013</td>
</tr>
<tr>
<td></td>
<td>CONT</td>
<td>0.009</td>
<td>0.009</td>
<td>0.012</td>
<td>0.012</td>
<td>0.014</td>
<td>0.013</td>
</tr>
<tr>
<td></td>
<td>DICH+</td>
<td>0.009</td>
<td>0.009</td>
<td>0.012</td>
<td>0.012</td>
<td>0.014</td>
<td>0.012</td>
</tr>
<tr>
<td></td>
<td>DICH</td>
<td>0.010</td>
<td>0.010</td>
<td>0.013</td>
<td>0.013</td>
<td>0.014</td>
<td>0.011</td>
</tr>
<tr>
<td>$\alpha=0.001$</td>
<td>CONT+</td>
<td>0.0005</td>
<td>0.0005</td>
<td>0.0012</td>
<td>0.0012</td>
<td>0.0019</td>
<td>0.0018</td>
</tr>
<tr>
<td></td>
<td>CONT</td>
<td>0.0005</td>
<td>0.0005</td>
<td>0.0013</td>
<td>0.0013</td>
<td>0.0020</td>
<td>0.0019</td>
</tr>
<tr>
<td></td>
<td>DICH+</td>
<td>0.0007</td>
<td>0.0007</td>
<td>0.0016</td>
<td>0.0016</td>
<td>0.0024</td>
<td>0.0022</td>
</tr>
<tr>
<td></td>
<td>DICH</td>
<td>0.0008</td>
<td>0.0008</td>
<td>0.0018</td>
<td>0.0018</td>
<td>0.0028</td>
<td>0.0026</td>
</tr>
</tbody>
</table>

a Types of study characteristics as the sources of heterogeneity; CONT-Continuous, DICH-dichotomous, PROP-proportion, CONT+DICH-a combination of continuous and dichotomous covariates.

Scaled-shifted parameters ($a$, $b$ and/or $c$) are estimated using b a regression, c method of moments, and d a numerical method.
Appendix C Type I Error Rates from the Corrected Null Distribution in Two Additional Simulation Situations

After carefully following these steps and comparing the candidate distributions (Appendix B), we select the following mixture distribution, \( a\chi^2_1 + b\chi^2_2 + c \), that shows anti-conservative but well-controlled type I error rates. Now we want to show that how it performs in two different simulation situations.

C.1 Varying Within-Study Variance

In simulation study, we generate data from \( b_i \sim N(0, \sigma_i^2) \) and \( \sigma_i^2 \sim \chi^2_1 \). In GWAS, different MAFs may affect the standard errors of effect size estimates of SNPs. In order to mimic and to explore how the corrected null distribution would work in such a situation, we vary \( \sigma_i^2 \) values when we generate data. We generate data using 10*\( \sigma_i^2 \) and 0.1*\( \sigma_i^2 \), and evaluate type I error rates from the corrected null distribution.
Table C.1 Type I error rates varying within-study variance for N=50

<table>
<thead>
<tr>
<th>α levels</th>
<th>Study Characteristics</th>
<th>1*σ_i²</th>
<th>10*σ_i²</th>
<th>0.1*σ_i²</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Type I error</td>
<td>95% C.I. LL</td>
<td>95% C.I. UL</td>
</tr>
<tr>
<td>α=0.01</td>
<td>CONT+</td>
<td>0.010</td>
<td>0.009</td>
<td>0.010</td>
</tr>
<tr>
<td></td>
<td>CONT</td>
<td>0.010</td>
<td>0.009</td>
<td>0.010</td>
</tr>
<tr>
<td></td>
<td>DIC+</td>
<td>0.009</td>
<td>0.009</td>
<td>0.010</td>
</tr>
<tr>
<td></td>
<td>DIC</td>
<td>0.010</td>
<td>0.009</td>
<td>0.010</td>
</tr>
<tr>
<td>α=0.001</td>
<td>CONT+</td>
<td>0.0009</td>
<td>0.0007</td>
<td>0.0010</td>
</tr>
<tr>
<td></td>
<td>CONT</td>
<td>0.0009</td>
<td>0.0007</td>
<td>0.0011</td>
</tr>
<tr>
<td></td>
<td>DIC+</td>
<td>0.0009</td>
<td>0.0007</td>
<td>0.0011</td>
</tr>
<tr>
<td></td>
<td>DIC</td>
<td>0.0008</td>
<td>0.0006</td>
<td>0.0010</td>
</tr>
</tbody>
</table>

* Types of study characteristics as the sources of heterogeneity: CONT-Continuous, DICH-dichotomous, PROP-proportion, CONT+DIC-a combination of continuous and dichotomous covariates.

b 95% C.I. LL-95% confidence interval lower limit; 95% C.I. UL-95% confidence interval upper limit.
C.2 Varying Z distribution

We explore how the corrected null distribution would work when we evaluate the test statistic using study characteristic measures simulated from different distribution. We examine when the study characteristics are continuous+continuous and dichotomous+dichotomous. The original continuous covariate measures are generated from $N(0, \sigma_z^2)$, $\sigma_z^2=1$. We vary the variance $\sigma_z^2$ to 0.09 and 0.25. The original dichotomous covariate measures are generated from Bernoulli(0.5). We vary the probability for the two dichotomous covariates to (0.3,0.6), and (0.1,0.9).

Table C.2 Type I error rates varying Z distribution when covariates are continuous

<table>
<thead>
<tr>
<th>$\alpha$ levels</th>
<th>$\sigma_z^2$</th>
<th>N=10</th>
<th>95% CI</th>
<th>95% CI</th>
<th>N=20</th>
<th>95% CI</th>
<th>95% CI</th>
<th>N=50</th>
<th>95% CI</th>
<th>95% CI</th>
<th>N=100</th>
<th>95% CI</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha=$ 0.01</td>
<td>0.09</td>
<td>0.008</td>
<td>0.008</td>
<td>0.009</td>
<td>0.009</td>
<td>0.010</td>
<td>0.009</td>
<td>0.010</td>
<td>0.009</td>
<td>0.010</td>
<td>0.009</td>
<td>0.010</td>
<td>0.010</td>
</tr>
<tr>
<td>$\alpha=$ 0.001</td>
<td>0.25</td>
<td>0.009</td>
<td>0.009</td>
<td>0.009</td>
<td>0.009</td>
<td>0.010</td>
<td>0.009</td>
<td>0.010</td>
<td>0.009</td>
<td>0.010</td>
<td>0.009</td>
<td>0.010</td>
<td>0.010</td>
</tr>
</tbody>
</table>

$a$ The variance of normal distribution for continuous covariates.

$b$ 95% C.I. LL-95% confidence interval lower limit; 95% C.I. UL-95% confidence interval upper limit.
<table>
<thead>
<tr>
<th>α levels</th>
<th>$p_1, p_2$</th>
<th>N=10 Type I error</th>
<th>95% CI LL</th>
<th>95% CI UL</th>
<th>N=20 Type I error</th>
<th>95% CI LL</th>
<th>95% CI UL</th>
<th>N=50 Type I error</th>
<th>95% CI LL</th>
<th>95% CI UL</th>
<th>N=100 Type I error</th>
<th>95% CI LL</th>
<th>95% CI UL</th>
</tr>
</thead>
<tbody>
<tr>
<td>α= 0.3,0.6</td>
<td>0.009</td>
<td>0.009</td>
<td>0.010</td>
<td>0.010</td>
<td>0.011</td>
<td>0.010</td>
<td>0.010</td>
<td>0.011</td>
<td>0.010</td>
<td>0.009</td>
<td>0.011</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.01</td>
<td>0.1,0.9</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>0.009</td>
<td>0.009</td>
<td>0.010</td>
<td>0.010</td>
<td>0.009</td>
<td>0.010</td>
<td></td>
</tr>
<tr>
<td>α= 0.3,0.6</td>
<td>0.0009</td>
<td>0.0007</td>
<td>0.0011</td>
<td>0.0010</td>
<td>0.0008</td>
<td>0.0012</td>
<td>0.0009</td>
<td>0.0007</td>
<td>0.0011</td>
<td>0.0008</td>
<td>0.0006</td>
<td>0.0010</td>
<td></td>
</tr>
<tr>
<td>0.001</td>
<td>0.1,0.9</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>0.0008</td>
<td>0.0006</td>
<td>0.0009</td>
<td>0.0008</td>
<td>0.0006</td>
<td>0.0009</td>
<td></td>
</tr>
</tbody>
</table>

The probability of Bernoulli distribution for two dichotomous covariates.

95% C.I. LL-95% confidence interval lower limit; 95% C.I. UL-95% confidence interval upper limit.
Appendix D Derivation of Score Functions and Information Matrix for Score Test

We derive the score function of our parameters of interest, \( \beta, \tau_1^2, \ldots, \tau_r^2 \), and their variance and covariance by the first and second derivatives of the log-likelihood in the equation (1) in Appendix A with respect to the parameters. For our convenience, we replace \((b - \beta)\) by \(\tilde{b}\). The score function, and its variance and covariance are obtained as follows

\[
u(\theta_p) = \frac{\partial l}{\partial \theta_p},
\]

\[
I(\theta_{pq}) = -E \left[ \frac{\partial^2 l}{\partial \theta_p \partial \theta_q} \right].
\]

where \(\theta_p, \theta_q \in \theta = (\beta, \tau_1^2, \ldots, \tau_r^2)\). Note that because \(R\) and \(\Sigma_v\) are diagonal matrices, the transpose of the covariance matrices are the same as the corresponding covariance matrices; \(R' = R, \Sigma_v' = \Sigma_v, W' = [Z\Sigma_vZ' + R]' = Z\Sigma_vZ' + R = Z\Sigma_vZ' + R = W\).

Now the score functions of our parameters are

\[
u(\beta) = \frac{\partial l}{\partial \beta} = \frac{\partial}{\partial \beta} \left[ C - \frac{1}{2} \ln|W| \right] + \frac{\partial}{\partial \beta} \left[ -\frac{1}{2} \tilde{b}'W^{-1}\tilde{b} \right]
\]

\[= 0 + \frac{\partial}{\partial \beta} \left[ -\frac{1}{2} \tilde{b}'W^{-1}\tilde{b} \right]
\]

\[= -\frac{1}{2} \tilde{b}' \left[ (-J_1W^{-1})' W^{-1}(-J_1) \right] \]

\[= \tilde{b}'W^{-1}J_1 \]
\[ u(\tau_j^2) = \frac{\partial l}{\partial \tau_j^2} = \frac{\partial}{\partial \tau_j^2} \left[ C - \frac{1}{2} \ln |W| \right] + \frac{\partial}{\partial \tau_j^2} \left[ -\frac{1}{2} \tilde{b}'W^{-1}\tilde{b} \right] \]

\[ = 0 - \frac{1}{2} \left[ \text{tr} \left( W^{-1}\frac{\partial w}{\partial \tau_j^2} \right) \right] - \frac{1}{2} \left[ \tilde{b}' \frac{\partial w^{-1}}{\partial \tau_j^2} \tilde{b} \right] \]

by \( \partial (\ln |X|) = \text{tr}(X^{-1} \partial X)' \)

\[ = -\frac{1}{2} \left[ \text{tr} \left( W^{-1}z_jz_j' \right) \right] - \frac{1}{2} \left[ \tilde{b}' \left( -W^{-1} \frac{\partial w}{\partial \tau_j^2} W^{-1} \right) \tilde{b} \right] \]

by \( \partial X^{-1} = -X^{-1}(\partial X)X^{-1} \)

\[ = -\frac{1}{2} \text{tr} \left( W^{-1}z_jz_j' \right) + \frac{1}{2} \tilde{b}'(W^{-1}z_jz_j'W^{-1})\tilde{b} \]

And the diagonal and off-diagonal elements of the information matrix are the second derivatives of the log-likelihood function as follows

\[-E \left[ \frac{\partial^2 l}{\partial \beta^2} \right] = -E \left[ \frac{\partial u(\beta)}{\partial \beta} \right] = -E[-J'_1 W^{-1}J_1] = J'_1 W^{-1}J_1 \]

\[-E \left[ \frac{\partial^2 l}{\partial \tau_i^2 \partial \tau_j^2} \right] = -E \left[ \frac{\partial u(\tau_j^2)}{\partial \tau_i^2} \right] = -E \left[ \frac{\partial u(\tau_j^2)}{\partial \tau_j^2} \right] \]

\[ = -E \left[ \frac{\partial}{\partial \tau_j^2} \left[ -\frac{1}{2} \text{tr} \left( W^{-1}z_i z_i' \right) + \frac{1}{2} \tilde{b}'(W^{-1}z_i z_i'W^{-1})\tilde{b} \right] \right] \]

\[ = -E \left[ -\frac{1}{2} \frac{\partial}{\partial \tau_j^2} \text{tr} \left( W^{-1}z_i z_i' \right) + \frac{1}{2} \tilde{b}' \left( \frac{\partial W^{-1}}{\partial \tau_j^2} z_i z_i'W^{-1} + W^{-1}z_i z_i' \frac{\partial W^{-1}}{\partial \tau_j^2} \right) \tilde{b} \right] \]

\[ = -E \left[ -\frac{1}{2} \text{tr} \left( \frac{\partial W^{-1}}{\partial \tau_j^2} z_i z_i' \right) + \frac{1}{2} \tilde{b}' \left( \frac{\partial W^{-1}}{\partial \tau_j^2} z_i z_i'W^{-1} + W^{-1}z_i z_i' \frac{\partial W^{-1}}{\partial \tau_j^2} \right) \tilde{b} \right] \]

\[ = -E \left[ -\frac{1}{2} \text{tr} \left( -W^{-1}z_jz_j'W^{-1}z_jz_j' \right) + \frac{1}{2} \tilde{b}' \left( -W^{-1}z_jz_j'W^{-1}z_i z_i'W^{-1} - W^{-1}z_i z_i'z_jz_j'W^{-1} \right) \tilde{b} \right] \]
\[\begin{align*}
&= -\frac{1}{2} E[\text{tr}(W^{-1}z_jz'_jW^{-1}z_i z'_i)] + \frac{1}{2} E[\bar{b}'(W^{-1}z_jz'_jW^{-1}z_i z'_i)\bar{b}] \\
&\quad + \frac{1}{2} E[\bar{b}'(W^{-1}z_i z'_iW^{-1}z_j z'_j)\bar{b}] \\
&= -\frac{1}{2} \text{tr}(W^{-1}z_jz'_jW^{-1}z_i z'_i) + \frac{1}{2} E \left[ \text{tr}\left((W^{-1}z_jz'_jW^{-1}z_i z'_i)\bar{b}\bar{b}'\right) \right] \\
&\quad + \frac{1}{2} E \left[ \text{tr}\left((W^{-1}z_i z'_iW^{-1}z_j z'_j)\bar{b}\bar{b}'\right) \right] \\
&= -\frac{1}{2} \text{tr}(W^{-1}z_jz'_jW^{-1}z_i z'_i) + \frac{1}{2} \text{tr}(W^{-1}z_jz'_jW^{-1}z_i z'_i)W \\
&\quad + \frac{1}{2} \text{tr}(W^{-1}z_i z'_iW^{-1}z_j z'_j)W \\
&= -\frac{1}{2} \text{tr}(W^{-1}z_jz'_jW^{-1}z_i z'_i) + \frac{1}{2} \text{tr}(W^{-1}z_jz'_jW^{-1}z_i z'_i) \\
&\quad + \frac{1}{2} \text{tr}(W^{-1}z_i z'_iW^{-1}z_j z'_j) \\
&= \frac{1}{2} \text{tr}(W^{-1}z_i z'_iW^{-1}z_j z'_j)
\end{align*}\]

If \( i = j \) then

\[-E \left[ \frac{\partial^2 l}{\partial (\tau_j^2)^2} \right] = -E \left[ \frac{\partial u(\tau_j^2)}{\partial \tau_j^2} \right] = \frac{1}{2} \text{tr}(W^{-1}z_jz'_jW^{-1}z_j z'_j)\]

\[-E \left[ \frac{\partial u(\beta)}{\partial \tau_j^2} \right] = -E \left[ \frac{\partial u(\tau_j^2)}{\partial \beta} \right] = -E\left[-\bar{b}'(W^{-1}z_jz'_jW^{-1})J_1\right] = 0\]
Appendix E Derivation of Score Function of $\tau_j$ and its Variance

We use the same log-likelihood function as in Appendix A. In order to obtain the information matrix $I(\theta)$, we show the first and second derivatives of the log-likelihood with respect to $\tau_j$. For our convenience, we replace $(b - \beta)$ by $\bar{b}$. We can obtain a score function by the first derivative of $l$ with respect to $\tau_j$ is

$$\frac{\partial l}{\partial \tau_j} = -\frac{1}{2} \left( tr(W^{-1} \frac{\partial W}{\partial \tau_j}) \right) - \frac{1}{2} \left( \bar{b}' \frac{\partial W^{-1}}{\partial \tau_j} \bar{b} \right)$$

$$= -\frac{1}{2} \left( tr(W^{-1} \frac{\partial W}{\partial \tau_j}) \right) - \frac{1}{2} \left( \bar{b}' \left( -W^{-1} \frac{\partial W}{\partial \tau_j} W^{-1} \right) \bar{b} \right)$$

$$= -\frac{1}{2} \left( tr(W^{-1}2z_j \tau_j z_j') \right) - \frac{1}{2} \left( \bar{b}' \left( -W^{-1}2z_j \tau_j z_j' W^{-1} \right) \bar{b} \right)$$

$$= -tr(W^{-1}z_j z_j') \tau_j + \bar{b}'(W^{-1}z_j z_j' W^{-1}) \bar{b} \tau_j.$$

The first derivative of $l$ with respect to $\tau_j$ evaluated at $\tau_j = 0$ contains no information: $\frac{\partial l}{\partial \tau_j} \bigg|_{\tau_j=0} = 0$. However, the second derivative of $l$ with respect to $\tau_j$ evaluated at $\tau_j = 0$ is not 0 and is proportional to the first derivative of $l$ with respect to $\tau_j^2$ evaluated at $\tau_j = 0$ (Dupuis, Siegmund, and Yakir 2007a).

The second derivative of $l$ with respect to $\tau_j$ is
\[
\frac{\partial^2 l}{\partial (\tau_j)^2} = -\frac{\partial}{\partial \tau_j} \left[ tr(W^{-1}z_j'\tau_j) \right] + \frac{\partial}{\partial \tau_j} \left[ \tilde{b}'(W^{-1}z_j'W^{-1}) \tilde{b} \tau_j \right]
\]

\[
= -tr \left( \frac{\partial W^{-1}}{\partial \tau_j} z_j' \right) \tau_j - tr(W^{-1}z_j' \frac{\partial \tau_j}{\partial \tau_j})
+ \tilde{b}' \left( \frac{\partial W^{-1}}{\partial \tau_j} z_j' W^{-1} + W^{-1}z_j' \frac{\partial W^{-1}}{\partial \tau_j} \right) \tilde{b} \tau_j + \tilde{b}'(W^{-1}z_j'W^{-1}) \tilde{b}
\]

\[
= -tr(-W^{-1}2z_j' \tau_j z_j' W^{-1} \tau_j) - tr(W^{-1}z_j')
+ \tilde{b}'(-W^{-1}2z_j' \tau_j z_j' W^{-1} + W^{-1}z_j' W^{-1} z_j' - W^{-1}2z_j' \tau_j z_j' W^{-1}) \tilde{b} \tau_j

+ \tilde{b}'(W^{-1}z_j'W^{-1}) \tilde{b}
\]

\[
= 2tr(W^{-1}z_j'W^{-1}z_j') \tau_j^2 - 4\tilde{b}'(W^{-1}z_j'W^{-1}z_j' W^{-1}) \tilde{b} \tau_j^2 - tr(W^{-1}z_j')
+ \tilde{b}'(W^{-1}z_j'W^{-1}) \tilde{b}
\]

This second derivative preserves the property that \( E \left[ \frac{\partial^2 l}{\partial (\tau_j)^2} \right] = 0 \) and carries information evaluated at \( \tau_j = 0 \) such that

\[
\left. \frac{\partial^2 l}{\partial (\tau_j)^2} \right|_{\tau_j=0} = -tr(W_0^{-1}z_i z_i') + \tilde{b}'(W_0^{-1}z_i z_i'W_0^{-1}) \tilde{b}.
\]
where $W_0 = Z\Sigma v_0 Z' + R$, $\Sigma v_0 = \Sigma v|_{\tau_j = 0}$, $\Sigma v = \text{var}[v] = \begin{bmatrix} \tau_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \tau_r^2 \end{bmatrix}$. Therefore, we take the second derivative of $l$ with respect to $\tau_j$ for our score function $u_{\tau_j} = \frac{\partial^2 l}{\partial (\tau_j)^2}|_{\tau_j = 0}$ and derive its variance by $E \left[u_{\tau_j} u_{\tau_j}'\right]$.

The derivation of the variance of the score function $\text{Var} \left[u_{\tau_j}\right]$ follows

$$\text{Var} \left[u_{\tau_j}\right] = E \left[u_{\tau_j} u_{\tau_j}'\right]$$
$$= E \left[\left(tr(W_0^{-1} z_j z_j')\right)^2 - 2tr(W_0^{-1} z_j z_j')tr(W_0^{-1} z_j z_j' W_0^{-1}) \bar{b}ight.$$ 
$$+ \bar{b}'(W_0^{-1} z_j z_j' W_0^{-1}) \bar{b} + (W_0^{-1} z_j z_j' W_0^{-1}) \bar{b}'$$
$$= \left(tr(W_0^{-1} z_j z_j')\right)^2 - 2tr(W_0^{-1} z_j z_j') tr(W_0^{-1} z_j z_j' W_0^{-1} \text{E}[\bar{b} \bar{b}'])$$
$$+ \text{E}[\bar{b}'(W_0^{-1} z_j z_j' W_0^{-1}) \bar{b} + (W_0^{-1} z_j z_j' W_0^{-1}) \bar{b}']$$
$$= \left(tr(W_0^{-1} z_j z_j')\right)^2 - 2tr(W_0^{-1} z_j z_j') tr(W_0^{-1} z_j z_j' W_0^{-1})$$
$$ + \text{E}[\bar{b}'(W_0^{-1} z_j z_j' W_0^{-1}) \bar{b} + (W_0^{-1} z_j z_j' W_0^{-1}) \bar{b}']$$
$$= -\left(tr(W_0^{-1} z_j z_j')\right)^2 + \text{E}[\bar{b}'(W_0^{-1} z_j z_j' W_0^{-1}) \bar{b} + (W_0^{-1} z_j z_j' W_0^{-1}) \bar{b}']$$

Now we have two terms for the variance of the score function and the derivation of the second term is not trivial. Kumar (Kumar 1973) and Misra (Misra 1972) derived the expectation of products of quadratic forms including the expectation of a product of two quadratic forms such that $E[\bar{c}' A_1 \bar{c} \bar{c}' A_2 \bar{c}] = 2tr(A_1 A_2) + tr(A_1)tr(A_2)$.
The second term of the variance of the score function can be re-written as follows

\[
E\left[ \tilde{b}'(W_0^{-1}z_jW_0^{-1})\tilde{b}\left(W_0^{-1}z_jW_0^{-1}\right)\tilde{b} \right] = E\left[ \tilde{b}'W_0^{-2}z_jW_0^{-2}b\tilde{b}'W_0^{-2}z_jW_0^{-2}b \right]
\]

Let \( \tilde{c} = \tilde{b}W_0^{-\frac{1}{2}} \sim N(0, I) \) and \( A = W_0^{-\frac{1}{2}}z_jW_0^{-\frac{1}{2}} \) then it can be re-written in the form of

\[
E[\tilde{c}'A\tilde{c}'A\tilde{c}].
\]

Therefore, as \( E[\tilde{c}'A\tilde{c}'A\tilde{c}] = 2tr(AA) + tr(A)tr(A), \)

\[
E\left[ \tilde{b}'W_0^{-2}z_jW_0^{-2}b\tilde{b}'W_0^{-2}z_jW_0^{-2}b \right] = 2tr\left( W_0^{-2}z_jW_0^{-2}z_jW_0^{-2} \right) + \left\{ tr\left( W_0^{-2}z_jW_0^{-2} \right) \right\}^2
\]

\[
= 2tr\left( W_0^{-1}z_jW_0^{-1}z_jW_0^{-1} \right) + \left\{ tr\left( W_0^{-1}z_jW_0^{-1} \right) \right\}^2
\]

Finally, the variance of the score function \( u_{\tau j_0} \) is

\[
Var\left[u_{\tau j_0}\right] = E\left[u_{\tau j_0}u_{\tau j_0}'\right] = -\left\{ tr\left( W_0^{-1}z_jW_0^{-1}z_jW_0^{-1} \right) \right\}^2 + E\left[ \tilde{b}'(W_0^{-1}z_jW_0^{-1})\tilde{b}\left(W_0^{-1}z_jW_0^{-1}\right)\tilde{b} \right]
\]

\[
= -\left\{ tr\left( W_0^{-1}z_jW_0^{-1}z_jW_0^{-1} \right) \right\}^2 + 2tr\left( W_0^{-1}z_jW_0^{-1}z_jW_0^{-1} \right) + \left\{ tr\left( W_0^{-1}z_jW_0^{-1}z_jW_0^{-1} \right) \right\}^2
\]

\[
= 2tr\left( W_0^{-1}z_jW_0^{-1}z_jW_0^{-1} \right)
\]
Appendix F Investigation on Inflated Score Test Statistic and Search for a Null Distribution that control Type I Error Rates

F.1 Univariate Score Statistics and Type I error rates

In order to track potential issues more effectively, we broke down the 3-df score statistic into three univariate score statistics. As shown in Table F.1, the type I error rates for $\tau^2$'s suggest inflation while the type I error rate for $\beta$ is controlled at the pre-specified $\alpha$ levels. This led us to examine the between-study variances more closely.

Table F.1 Type I error rates for univariate score test

<table>
<thead>
<tr>
<th>Type I error</th>
<th>$H_0: \beta = 0$</th>
<th>$H_0: \tau^2_1 = 0$</th>
<th>$H_0: \tau^2_2 = 0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha = 0.05$</td>
<td>0.0481</td>
<td>0.0565</td>
<td>0.0537</td>
</tr>
<tr>
<td>$\alpha = 0.01$</td>
<td>0.0099</td>
<td>0.0354</td>
<td>0.0352</td>
</tr>
</tbody>
</table>

F.2 Empirical Information Matrix vs. Correlation of Score Functions

We wondered if the expected information matrix was correctly calculated. We simulated data as described in the simulation study and calculated score functions and the information matrix under the null hypothesis. We also computed the covariance matrix using the simulated score functions and compared to the average of the simulated information matrix under the null. We observed similar values for the two matrices (Table F.2) and concluded that the information matrix did not cause the inflation in
calculating the score test statistics. The matrices were calculated using 10,000 null simulations with 50 studies and dichotomous covariates.

Table F.2 Average of simulated information matrix and computed covariance matrix using score functions

<table>
<thead>
<tr>
<th>Average of simulated $I(\theta_0)$</th>
<th>Covariance of simulated score functions</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\bar{I}(\theta_0) = \begin{bmatrix} 181 &amp; 0 &amp; 0 \ 0 &amp; 4258 &amp; 1015 \ 0 &amp; 1015 &amp; 5295 \end{bmatrix}$</td>
<td>$\text{Cov}(U(\theta_0)) = \begin{bmatrix} 182 &amp; 8 &amp; 37 \ 8 &amp; 4211 &amp; 1099 \ 37 &amp; 1099 &amp; 5308 \end{bmatrix}$</td>
</tr>
</tbody>
</table>

F.3 Forest Plot

We first plotted the scatterplot of score statistics (N=50, continuous covariates) versus LRT values (from chapter 3) computed using the same data. In Figure F.1, the top panel shows all Score-LRT pairs and the bottom panel shows the data points only between LRT=9 and LRT=11. Red triangles indicate extreme score statistics where the ratio of score values to LRT are greater than 4. We took three examples of extreme score and non-extreme score values with similar LRT from the bottom panel and explored in forest plots (Figure F.2a-c). In the forest plots, we wanted to examine if there were any possible influential points (in red rectangular box) that might cause inflation in computing test statistics and p-values. Those points had 95% confidence intervals that did not contain zero and had relatively large effect sizes compared to small variances. Because these points were simulated under the null, we expected that effect sizes with
small variances were close to zero or the 95% confidence intervals included zero. If the variance was not small and the 95% confidence did not include zero, the inflation in score statistics did not occur (See Figure F.2b, Non-Extreme Score for study 43 in blue oval box). We computed score statistics with and without the influential points and showed that those points caused inflated score statistics in Table F.3.

Table F.3 Score test statistics ($T_s$) computed including and excluding the influential points

<table>
<thead>
<tr>
<th>Influential Points</th>
<th>Included</th>
<th>Excluded</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T_s$ for Case 1</td>
<td>63.96</td>
<td>1.22</td>
</tr>
<tr>
<td>$T_s$ for Case 2</td>
<td>54.49</td>
<td>0.53</td>
</tr>
<tr>
<td>$T_s$ for Case 3</td>
<td>51.79</td>
<td>2.97</td>
</tr>
</tbody>
</table>

Since we identified the issue with inflated score statistics, we have tested several approaches to obtain controlled type I error rates. Re-calibration of the null distribution by estimating scaled-/shifted-parameters as we did in chapter 3 did not control type I error rates for the score test. Another approach was to add a small value to the error variances in order to avoid getting small variances and to stabilize the variances. We determined small values by the minimum variance divided by 2, 5, 10 or 20 for each simulation. However, this approach did not yield a uniform distribution in the histograms or follow the diagonal line in the Quantile-Quantile (Q-Q) plots as we expected (See Figure F.3). Lastly, we computed null score statistics (N=50) and constructed an
empirical null distribution by aggregating the null score values from all of the simulations. We computed \( p \)-values from the empirical null distribution. As shown in Figure F.4, we obtained well-controlled type I error rates using the empirical null distribution. A detailed description of the construction of the empirical null distribution is described in chapter 4.
Figure F.1 Scatter plot of score statistics vs. LRT
Figure F.2a Forest plots for extreme and non-extreme score statistics; Extreme $T_s = 64$ and Non-Extreme $T_s = 12$
Figure F.2b Forest plots for extreme and non-extreme score statistics; Extreme $T_s = 54$ and Non-Extreme $T_s = 11$
Figure F.2c: Forest plots for extreme and non-extreme score statistics: Extreme $T_s = 52$ and Non-Extreme $T_s = 11$.
Figure F.3 Histogram and Quantile-Quantile (Q-Q) plots for score test p-values adding small variances
Figure F.4 Histogram and Quantile-Quantile (Q-Q) plots for score test p-values calculated from the empirical null distribution.
BIBLIOGRAPHY


Colditz, G A, T F Brewer, C S Berkey, M E Wilson, E Burdick, H V Fineberg, and F


Han, Fang, and Wei Pan. 2010. “A Data-Adaptive Sum Test for Disease Association with Multiple Common or Rare Variants.” *Human Heredity* 70 (1): 42–54. doi:10.1159/000288704.


CURRICULUM VITAE

JAEOYUNG HONG

Department of Biostatistics
Boston University School of Public Health
801 Massachusetts Ave, 3rd floor
Boston, MA 02118
Email: jhong@bu.edu

EDUCATION
Boston University Ph.D in Biostatistics (2016) Boston, MA
University of Illinois at Chicago M.S. in Biostatistics (2011) Chicago, IL
Kookmin University B.S. in Electronic Engineering (2002) Seoul, South Korea

PROFESSIONAL EXPERIENCE

Research Experience:
Department of Biostatistics, Boston University Research Assistant (2011-2016) Boston, MA
Center for Sleep and Circadian Biology & Transportation Center, Northwestern University Research Programmer (2011-2012) Chicago, IL
University of Illinois at Chicago Research Analyst (2011) Chicago, IL

Teaching Experience:
Department of Biostatistics, Boston University Teaching Assistant, Introduction to Biostatistics (Core MPH course) (2012-2014) Boston, MA
Department of Preventive Medicine, Northwestern University Graduate Laboratory Assistant, Advanced Biostatistics (2011) Chicago, IL
Division of Epidemiology and Biostatistics, School of Public Health, University of Illinois at Chicago Chicago, IL
Teaching Assistant, Biostatistics I, II (2010, 2011)

Other Experience:
Statistical Consultant Design and Analysis Core, UIC Center for Clinical and Translational Science Chicago, IL
Statistical Consultant (2011)

Community Services:
April, 2014 UPWARD BOUND program for high school students from underprivileged backgrounds and/or first generation college students, Boston University, Boston, MA
November, 2014 Judge for MPH practicum poster session, School of Public Health, Boston University, Boston, MA

Honors and Awards:
1. CHARGE Meeting Jackson, MS, 2015
2. Best Poster Award at 2013 Genome Science Institute Research Symposium

Publications:


4. Chen, Han, Seung Hoan Choi, Jaeyoung Hong, Chen Lu, Jacqueline N. Milton, Catherine Allard, Sean M. Lacey, Honghuang Lin, and Josée Dupuis. "Rare


**Submitted:**

**Poster Presentations:**
1. “Statistical Model with Multiple Random Effects to Account for Heterogeneity in GWAS Meta-analysis” **Jaeyoung Hong** on behalf of CHARGE and MAGIC Consortium Investigators. Jackson CHARGE Investigator Meeting 2015, Jackson, MS
2. “Evaluation of two-stage approach in the trans-ethnic meta-analysis” Jaeyoung Hong, Josée Dupuis, Ching-Ti Liu. 2014 Genome Science Institute Research Symposium, Boston University School of Medicine, Boston, MA

3. “Comparison of Transethnic Meta-Analysis Methods in Genome-Wide Association Studies” Jaeyoung Hong, Josée Dupuis, Ching-Ti Liu. Scholars Day 2014. Boston University, Boston, MA

4. “Genome-Wide Association with Glycemic Traits in African Americans Suggests Allelic Heterogeneity at Known Loci: the African American Glucose and Insulin Genetic Epidemiology (AAGILE) Consortium.” Jaeyoung Hong, Ching-Ti Liu, Josée Dupuis, James B. Meigs on behalf of AAGILE consortium. 2013 Genome Science Institute Research Symposium, Boston University School of Medicine, Boston, MA

5. “Obstructive sleep apnea risk and obesity prevalence in tow vessel captains and pilots on American waterways.” Reid KJ, Bochkarev M, Rowan E, Hong J, Kang J, Turek FW. Jun3 2013. 27th Annual Meeting of the Associated Professional Sleep Societies, LLC.
