Bayesian Methods for Fine Mapping in Genetic Association Studies

Eleni Frangou

Keble College

Hilary Term 2013

Department of Statistics

University of Oxford

A dissertation submitted in partial fulfilment of the degree of Master of Science by Research

Supervisor: Dr Jonathan Marchini

“Τα καλά κόποις κτώνται.”

Bayesian Methods for Fine Mapping in Genetic Association Studies

Eleni Frangou, Keble College

Department of Statistics, University of Oxford

M.Sc(Res) Thesis, Hilary Term 2013

Abstract

Over the past several years genetic variation has been the centre of attention for different branches of genetics. The use of genome wide association studies (GWAS) has facilitated the discovery of a plethora of genetic variants which are associated with a vast range of diseases. These GWAS primarily examine a single variant, mostly a Single Nucleotide Polymorphism (SNP), at a time based on genotype data. Although this practice has proven to be powerful, low percentages of heritability prompt a call for the exploration of more complex structures of association.

As diseases are more likely to be caused by several variants, multi-SNP analyses are used to refine the signal leading to the development of fine mapping methodologies. Further biological information is accounted for by studying the diploid nature of the human genome which identifies phenomena, such as compound heterozygosity. We propose a flexible Bayesian fine mapping model for quantitative traits allowing for the analysis of haplotype data under a variety of genotype and haplotype models. A Metropolis-Hastings algorithm is employed to search the model space efficiently. Simulations show competitive performance over existing methods when compound heterozygous effects are present in the data.

An extension of a single cohort framework is meta-analysis where a collection of studies

is used to increase confidence and boost power in noticeable associations. Traditional approaches include the fixed effects model where the assumption is that genetic effects are homogeneous amongst cohorts while the random effects model accounts for heterogeneity. We propose a Bayesian meta-analysis approach which implements both fixed and random effects with the addition of a parameter measuring the existence of effect in each cohort. Simulations and application to real samples explore the performance of the method under different conditions and show comparable evidence over established methodologies.

Acknowledgements

I would like to thank Jonathan Marchini for his guidance, support and advice on this topic. It has been both challenging and valuable and without his supervision I would not have been capable of completing this process. I would also like to thank my viva examiners, Andrew Morris and Cathryn Lewis, for taking the time to read my thesis and for their valuable comments. I am most grateful to the Engineering and Physical Sciences Research Council and the Department of Statistics for the funding I have received throughout the course of my degree. I would also like to thank the IT team of the department who have been very patient while answering every single computing question I had.

My officemates, past and present, have been helpful and supportive over the past few years and without whom my working place would have been different. Many thanks to Androniki Menelaou, Claire Churchhouse, Valentina Iotchkova, Marie Forest and Melissa Mackza.

I have been lucky enough to be a member of a vibrant Greek society which has always offered an enjoyable environment and provided a balance to my academic work. I would especially like to thank Margarita Satraki, Chris Menelaou, Dimitris Sakellaridis, Lucy Mathioudaki, Christiana Kartsonaki, Andreas Anastasiou, Savina Joseph and all the members of the Nostos choir. I would also like to thank other friends along the way: Styliana Tremouri, Tereza Andreou, Aditi Nafde, Ruth Simmons and Sophie Kershaw.

Last, but not least, I would like to say a huge thank you to my family who have provided me with vital encouragement and support throughout the years. I dedicate this thesis to them. My parents, who live more than two thousand miles away, have been my restless companions on this journey and have never stopped believing in my abilities. My brothers and their families have always been there for me to help me through any difficulties. And my grandparents for their generous gifts and kind words.

Contents

1 Introduction **15**

1.1 Exploration of Genetic Variation 15

1.2 Genome Wide Association Studies 17

1.3 The Detection of Multiple Variants 20

1.3.1 The Importance of Haplotype Information 27

1.3.2 Compound Heterozygosity 30

1.4 Analysis of Multiple Cohorts 34

1.5 Summary 45

I A Novel Model for Fine Mapping **47**

2 Methodology **48**

2.1 The Bayesian Linear Model 49

2.1.1 The Likelihood 50

2.1.2 The Prior Distributions 50

2.1.3 The Posterior Distribution 51

2.1.4 The Design Matrix, $\mathbf{X}$ 52

2.1.5 Model Comparison 54

2.2 The Metropolis-Hastings Algorithm 55

2.2.1 Formulation of the Algorithm 55

2.2.2 The Proposal Density, q . . . 57
2.3 Assessment Measures . . . 59
2.4 Summary . . . 60

3 Simulation Study **61**
3.1 Generation of Simulated Data . . . 61
3.1.1 HAPGEN . . . 61
3.1.2 Generation of Phenotypes . . . 62
3.2 Power Comparisons . . . 65
3.3 Examination of the Speed to Convergence . . . 83
3.3.1 The Kullback Leibler Divergence . . . 84
3.3.2 Results . . . 86
3.4 Sensitivity to Prior Assumptions . . . 91
3.5 Acceptance Rate of the Metropolis-Hastings Algorithm . . . 91
3.6 Summary . . . 94

II A Novel Approach to Meta Analysis **96**

4 Methodology **97**
4.1 The Random Effects Model, All Cohorts Associated . . . 99
4.1.1 The Gibbs Sampler . . . 101
4.1.1.1 Full Conditional Posterior Distributions . . . 101
4.1.1.2 The Gibbs Sampling Algorithm . . . 101
4.2 The Fixed Effects Model, All Cohorts Associated . . . 103
4.3 The Random Effects Model, Subset of Cohorts Associated . . . 104
4.4 The Fixed Effects Model, Subset of Cohorts Associated . . . 107
4.5 The Marginal Likelihood from the Gibbs Output . . . 109
4.5.1 Chib Approach . . . 109

4.5.2 Pritchard Approach . . . 112
4.6 Assessment Measures . . . 113
4.7 Summary . . . 115

5 Simulation Study and Analysis of Real Samples 116
5.1 Generation of Simulated Data . . . 116
5.1.1 Heritability Estimates . . . 118
5.1.2 The Prior Distributions . . . 121
5.2 Probability to Detect the True Model . . . 123
5.3 Power to Detect Association with the Phenotype . . . 127
5.4 Power to Identify an Effect in Associated Cohorts . . . 135
5.5 Convergence Analysis . . . 139
5.5.1 Graphical Analysis . . . 140
5.5.2 The Potential Scale Reduction Factor . . . 140
5.5.3 Stability of Chib's Method . . . 144
5.6 Analysis of Real Samples . . . 146
5.6.1 Data Description . . . 146
5.6.2 Pre-phasing and Imputation . . . 147
5.6.3 Association Analysis . . . 150
5.6.4 Meta Analysis . . . 150
5.7 Summary . . . 167

6 Conclusions and Future Prospects 169

Appendices 175

A Bayesian Linear Model Derivations 176

B Naming system for genetic models 178

C Sum of two squared forms 180

D Fully Conditional Posterior Distributions of the Meta Analysis Models 181

D.1 Random Effects Model, All Cohorts Associated 181

D.2 Fixed Effects Model, All Cohorts Associated 185

D.3 Random Effect Model, Subset of Cohorts Associated 188

D.4 Fixed Effects Model, Subset of Cohorts Associated 194

E Additional Simulation Results for the Fine Mapping Model 199

F Probability to Detect the True Model using the Pritchard Approach 218

List of Figures

1.1 Compound heterozygosity effects 31

3.1 Power to detect marginal effects, low flanking LD 70
3.2 Power to detect joint effects, low caus/low flanking LD 71
3.3 Power to detect joint effects, med caus/low flanking LD 72
3.4 Power to detect joint effects, high caus/low flanking LD 73
3.5 Power to detect marginal effects, med flanking LD 74
3.6 Power to detect joint effects, low caus/med flanking LD 75
3.7 Power to detect joint effects, med caus/med flanking LD 76
3.8 Power to detect joint effects, high caus/med flanking LD 77
3.9 Power to detect marginal effects, high flanking LD 78
3.10 Power to detect joint effects, low caus/high flanking LD 79
3.11 Power to detect joint effects, med caus/high flanking LD 80
3.12 Power to detect joint effects, high caus/high flanking LD 81
3.13 Power to detect the compound heterozygosity model 82
3.14 KL divergence measure to assess speed to convergence, low caus/low flanking LD 87
3.15 KL divergence measure to assess speed to convergence, low caus/med flanking LD 88
3.16 KL divergence measure to assess speed to convergence, low caus/high flanking LD 89

3.17 KL Divergence measure to assess sensitivity to prior assumptions 92

5.1 Prior Distributions of the Model Parameters 122
5.2 Power to detect association, $\lambda = 1000$ 130
5.3 Power to detect association, $\lambda = 100$ 131
5.4 Power to detect association, $\lambda = 15$ 132
5.5 Power to detect association, $\lambda = 5$ 133
5.6 Power to detect association, $\lambda^{-1} = 0$ 134
5.7 Power to detect effect in each cohorts,$\lambda = 100$ 136
5.8 Power to detect effect in each cohorts,$\lambda = 15$ 137
5.9 Power to detect effect in each cohorts,$\lambda^{-1} = 0$ 138
5.10 Traceplots for $\boldsymbol{\beta}$ 141
5.11 Post-imputation information measure from two independent imputation runs for the cohort LOLIPOP 149
5.12 Meta-analysis p-values for the SNPs and indels of the 15q25 region obtained from **META** via fitting a fixed effects model 153
5.13 Meta-analysis p-values for the SNPs and indels of the 15q25 region obtained from **METASOFT**'s fixed effects model 154
5.14 Meta-analysis Bayes factors for the SNPs and indels of the 15q25 region obtained under the fixed effects model, all cohorts associated 155
5.15 Meta-analysis p-values for the SNPs and indels of the 15q25 region obtained from **METASOFT**'s binary effects model 156
5.16 Meta-analysis Bayes factors for the SNPs and indels of the 15q25 region obtained under the fixed effects model, subset of cohorts associated 157
5.17 Meta-analysis p-values for the SNPs and indels of the 15q25 region obtained from **METASOFT**'s random effects model 158
5.18 Meta-analysis Bayes factors for the SNPs and indels of the 15q25 region obtained under the random effects model, all cohorts associated 159

5.19 Meta-analysis Bayes factors for the SNPs and indels of the 15q25 region obtained under the random effects model, subset of cohorts associated . . . 160

5.20 Prior distributions of the heterogeneity parameter λ for different combinations of the hyperparameters c and d 164

5.21 Meta-analysis Bayes factors for the SNPs and indels of the 15q25 region obtained under the random effects model, subset of cohorts associated, $c = 500$, $d = 1$ 165

E.1 Power to detect marginal effects, low flanking LD and $\beta_1 = 0.20$ 201

E.2 Power to detect joint effects, low caus/low flanking LD and $\beta_1 = 0.20$ . . . 202

E.3 Power to detect joint effects, med caus/low flanking LD and $\beta_1 = 0.20$ . . . 203

E.4 Power to detect joint effects, high caus/low flanking LD and $\beta_1 = 0.20$ . . . 204

E.5 Power to detect marginal effects, med flanking LD and $\beta_1 = 0.20$ 205

E.6 Power to detect joint effects, low caus/med flanking LD and $\beta_1 = 0.20$ . . . 206

E.7 Power to detect joint effects, med caus/med flanking LD and $\beta_1 = 0.20$ . . 207

E.8 Power to detect joint effects, high caus/med flanking LD and $\beta_1 = 0.20$ . . 208

E.9 Power to detect marginal effects, high flanking LD and $\beta_1 = 0.20$ 209

E.10 Power to detect joint effects, low caus/high flanking LD and $\beta_1 = 0.20$ . . . 210

E.11 Power to detect joint effects, med caus/high flanking LD and $\beta_1 = 0.20$ . . 211

E.12 Power to detect joint effects, high caus/high flanking LD and $\beta_1 = 0.20$ . . 212

E.13 Power to detect the compound heterozygosity model, $\beta_1 = 0.20$ 213

E.14 KL divergence measure to assess speed to convergence, low caus/low flanking LD, $\beta_1 = 0.20$ 214

E.15 KL divergence measure to assess speed to convergence, low caus/med flanking LD, $\beta_1 = 0.20$ 215

E.16 KL divergence measure to assess speed to convergence, low caus/high flanking LD, $\beta_1 = 0.20$ 216

E.17 KL Divergence measure to assess sensitivity to prior assumptions, $\beta_1 = 0.20$ 217

List of Tables

2.1 Haplotype Combinations for 2 SNPs 53

2.2 Single Haplotype Models 53

2.3 Probability distribution of the genetic models given $|S|$ 57

3.1 LD Intervals 63

3.2 Probabilities assigned to the different actions in order to assess the speed of the model to converge to the truth 85

3.3 Acceptance rates of the Metropolis Hastings algorithm 93

5.1 Parameter details for the simulation of phenotypes 118

5.2 Hyperparameter values 122

5.3 Probability to detect the correct meta analysis model 126

5.4 Posterior Summaries of $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_F$ 142

5.5 $\hat{R}_c$ for notable parameters of each model 144

5.6 Stability of Chib's method 145

5.7 Summary information for the cohorts in the meta-analysis 147

5.8 Genomic control inflation factor for each cohort 151

5.9 Meta analysis results for the 15-78874842 deletion 162

5.10 Posterior Summaries of top deletion for different values of c and d 166

5.11 Posterior Probability of effect of top deletion for different models 167

B.1 Naming System of Models for the Fine Mapping Model 179

F.1 Probability to detect the correct meta analysis model using Pritchard's method219

F.2 Stability of Pritchard's method 220

Chapter 1

Introduction

1.1 Exploration of Genetic Variation

Genetic variation describes naturally occurring genetic differences among individuals. It can occur through mutation, recombination and migration of genes. Following the completion of the Human Genome Project in 2003, researchers were able to locate areas along the genome that varied between individuals. This international breakthrough project aimed to determine the sequence of the approximately 3 billion chemical base pairs which compose our DNA. As a result of its success another major, multi-ethnic effort was initiated in 2005, The International HapMap Project, which aimed to describe the common patterns of human genetic variation. The International HapMap resource reports patterns of genome-wide variation and correlation (linkage disequilibrium) in different population samples, which greatly aids the design and analysis of genetic studies. The most common type of genetic variation is the difference in individual bases, commonly known as the Single Nucleotide Polymorphisms (SNPs). Approximately 10 million SNPs were identified in a limited set of DNA samples consisting of 270 individuals from four different origins: 30 trios from Yoruba in Nigeria, 45 unrelated individuals from Beijing in China, 45 unrelated individuals from Tokyo in Japan and 30 trios of American residents with northern and western

European ancestry [1]. The knowledge of these SNPs and their linkage-disequilibrium patterns enabled studies, however understanding of human genetic variation remained limited with respect to variant type, frequency and population diversity [96]. In 2008, The 1000 Genomes Project was launched with the aim of establishing by far the most detailed catalogue of human genetic variation. The aim of this project was to discover, genotype and provide accurate haplotype information on all types of human DNA polymorphism in multiple populations. By 2010, over 95% of the accessible variants found in any individual had been catalogued. This included 15 million SNPs, 1 million short insertions and deletions and 20,000 structural variants [94]. Two years later, the genomes of 1,092 individuals from 14 populations gave rise to a validated haplotype map of 38 million SNPs, 1.4 million short insertions and deletions and more than 14,000 larger deletions. It was showed that individuals from different populations carry different profiles of rare ($< 5\%$ frequency) and common variants and that low-frequency variants show significant geographic differentiation. The percentage of accessible genome, as reported by The 1000 Genomes Project, raised by 3% and enabled further the analysis of common and low-frequency variants in individuals from diverse, including admixed, populations [95].

The information of The HapMap and The 1000 Genomes Projects, which is available to the public, has facilitated a new type of research effort: the genome-wide association study (GWAS). In such a study, the distribution of SNPs is determined in hundreds or even thousands of people with and without a particular disease. It involves rapidly scanning markers across the complete sets of DNA, or genomes, of many people to find genetic variations associated with a particular disease. Once new genetic associations are identified, researchers can use the information to develop better strategies to detect, treat and prevent the disease. Since 2005, enormous progress in mapping complex traits in humans has been made with more than 1,350 GWAS reported until the end of the first half of 2012 [2].

1.2 Genome Wide Association Studies

The genetic architecture of many complex traits has been undeniable and for numerous diseases there has been a rapid expansion in the numbers of loci implicated in predisposition including type 2 diabetes [78] where more than 300,000 SNPs were tested in a French case-control cohort, prostate cancer [22] where seven loci were detected to be associated with the most common cancer affecting males in developed countries and breast cancer [21] where loci were detected from conducting a two-stage genome wide association study. An enormous breakthrough in the history of these studies was the Wellcome Trust Case Control Consortium (WTCCC) where seven major diseases were studied and highly significant associations were reported. More specifically, case-control comparisons identified 24 independent association signals using the frequentist measure of a p-value for assessment. These conditions include Crohn's disease, type 1 and 2 diabetes, rheumatoid arthritis and bipolar disorder [97].

Mathematical and statistical tools have evolved to deal with the accomplishment of such large scale studies. The default approach to determining evidence for the association between genetic variants and phenotypes is to calculate the p-value for the null hypothesis of no association. Frameworks based on frequentist methods have flourished since the start of the GWAS era and numerous models have been developed for this purpose. An enlightening survey which discusses the key methods and addresses the challenges involved is given by Balding in 2006 [6]. It is stressed that preliminary analyses is of great importance and data should be thoroughly checked, for example, for allele frequency inconsistencies, missing data patterns as well as LD and recombination patterns and possible population stratification. Testing for association has focused on single-SNP analysis, dealing with both case-control and quantitative phenotypes. Single SNP association studies make the assumption that disease variants will be detectable based on their marginal effects. In a case-control study, one can choose between the Pearson test or the Fisher exact test and logistic regression while the natural statistical tool for continuous traits is linear regression which requires

the trait to be approximately normally distributed for each genotype. For complex traits, it is widely applicable that contributions to disease risk will often be additive where the heterozygote risk will be intermediate between the two homozygote risks. Different models can be incorporated such as the dominant and the recessive.

Alternative approaches for detecting associations are provided by Bayesian methods which require some additional modelling assumptions. Computational limitations prevented these approaches from being widely used, however in the last 20 years, they have become more and more attractive and a reason for this is the ability to incorporate prior information into the specification of the model. Bayesian inference has been applied to different areas of genetic analysis: population genetics, genomics and human genetics. Standard techniques on model testing are based on the Bayes factor which is used to directly test hypotheses about association [8]. This quantity has been viewed as an alternative to the p-value which, as has been described in [85], has a fundamental limitation since it is difficult to quantify the confidence that a given SNP is truly associated with a phenotype depending on the factors that affect the power of the test. Balding and Stephens in [85] discuss the advantages of Bayesian methodology over classical tests and provide guidelines for appropriate model specifications. This review can be viewed as the Bayesian equivalent of Balding's review of frequentist approaches for association studies [6].

Multiple software packages have been developed to facilitate genome wide scans, some of them performing association tests as a part of a broader genetic analysis package. For example, **PLINK** is a whole genome association analysis tool kit designed to perform a range of basic, large-scale analyses in a computationally tractable way. Tests are based on comparing allele frequencies between cases and controls, the Cochran-Armitage trend test, Fisher's exact test, different genetic models, tests for stratified samples and tests for quantitative traits [68]. For the determination of causal SNPs in the aforementioned WTCCC studies, **SNPTEST** was used. This is a program for the analysis of single SNP associations in genome-wide studies. The tests employed consist of those for binary and quantitative

phenotypes both under a frequentist and Bayesian framework, conditional analysis on sets of covariates or SNPs and a set of methods to account for imputed data. The program is designed to work with the output of **IMPUTE** which carries out genotype imputation [52]. This is an important topic in the analysis of genetic samples and it is of utmost importance to account for uncertainty in the imputed genotypes when searching for phenotypic associations for various reasons such as boosting power, fine-mapping associations and combining results in a meta-analysis. Along the same lines Servin and Stephens introduced a framework for the analysis of association studies designed to allow untyped variants to be more efficiently and directly assessed for association with a phenotype. This software is called **BIMBAM** [75]. Although, the approach focuses on associations in a restricted region, it is computationally feasible for genome-wide studies.

The genome-wide association method has certainly been one of the major breakthroughs in the area of genetics where the identification of susceptibility variants leads to novel biological insights and personalised medicine [54]. The underlying rationale for GWAS is the "common disease, common variant" hypothesis (CDCV hypothesis) which states that complex diseases are largely attributable to a moderate number of common variants. On the other hand, most common variants explain only a small proportion of the phenotypic variance. As an example, about 200 loci have been found to be associated with human height but they explain only about 10% of phenotypic variance [53] while low percentages of heritability have been explained for type 2 diabetes despite the large number of contributing loci identified [109]. A coincise summary of more conditions with their respective heritability estimates is included in [49]. This problem of missing heritability has replaced the CDCV hypothesis and states that the loci detected by genome wide association studies explain without exception a small minority of the inferred genetic variance [31]. Many explanations for this "missing heritability" have been suggested, for example the analysis of rare variants which are present in less than 5% of the population. Rare variants appear quite frequently along the genome as the latest 1000 Genomes Project data releases suggest.

These are extremely difficult to identify since they carry small genetic effects. Interest is also shown in detecting interactions between genetic or environmental factors as most common diseases are influenced by several factors. This introduces a whole new class of models which aim to detect effects that, due to their interaction with other genetic or environmental causes, are not identified by standard single SNP tests. Hence, one end of the spectrum contains diseases which are completely explained by genetics. The other end, which is more complicated, describes diseases which are affected by our genes but also as importantly by lifestyle, dietary preferences and other so called environmental factors. Hence the lack of a complete picture of the phenotypic variation is addressed by considering interactions between loci [49, 16, 99].

1.3 The Detection of Multiple Variants

The notion of "missing heritability" suggests that additional genetic factors remain to be discovered. Certain explanations of this notion have been suggested: the true causal variants are rare, the power of GWAS to detect loci of small effect is low and the SNP by SNP testing strategy is underpowered to detect loci that interact with other genetic or environmental factors since such effects may not be detectable unless the interacting factors are also included in the analysis. The elevated power of such statistical models motivates the development of enhanced methods for identifying and modelling statistical interaction in the context of GWAS. For example, a recent study investigating body mass index (BMI) detected 32 SNPs which together explained only 1.45% of the inter-individuals variation in BMI where the strongest SNP accounted for only 0.34% of the variance [81]. As an attempt to explain unseen variation, gene-gene interactions were surveyed in four European populations via exhaustive pair-wise genome scans. Interactions were indeed reported although it was supported that additional efforts are needed in order to confirm such findings [105]. Two-locus interactions raise a multiple testing issue as well as computational

time constraints. An approach which uses three submodels of the general two-locus case is used to compare different strategies for the detection of interactions [51]. These three models consider interlocus interactions where the first one includes multiplicative effects within and between loci while the second and third include a baseline value. Then, for model 2 the odds increase multiplicatively within and between genotypes while model 3 specifies a threshold of disease effects rather than multiplicative gene action. This framework compares three search strategies for interactions: single locus, which requires either at least one or both of the loci to achieve significance, full interaction, which fits the full model with at most nine parameters and two-stage, where all loci meeting a threshold in a single-locus analysis are then examined for interactions. The two-stage strategy also facilitates the search of higher order interactions but the time taken to perform the analysis increases exponentially with the order of interaction. The drawback of this approach is that loci will only be filtered into the second or subsequent stages and the procedure will not be helpful for detecting interactions that occur in the absence of marginal effects [16].

An alternative approach for selecting the best predictor variables and their interactions is Bayesian model selection techniques. The distinction between classical approaches is the specification of prior distributions for the unknown regression parameters and/or the possible models available to fit [16]. Bayesian methods additionally incorporate the uncertainty in model form and in the estimated parameters, which allows the averaging across all models taking into account their posterior probabilities. It is also advantageous that Bayesian analysis results in probability statements about the parameters as opposed to frequentist p-values. Finally, the use of stochastic procedures such as Markov Chain Monte Carlo (MCMC) allow one to search the model space for the most probable models and variables without fitting them all [26].

Over the past few years there has been a remarkable progress in the development of methods, especially based on the Bayesian framework, which serve the purpose of replicating associations that have been detected through genome-wide association studies. Fine map-

ping is a methodology used to resolve locus boundaries. As complex diseases are more likely to be caused by several genes and variants, multi-SNP and candidate gene analyses can help to refine signal. Candidate gene studies focus on genes that are selected due to prior findings about their role in disease. They take advantage of the escalated statistical efficiency of association analysis and the biological understanding of the disease or phenotype [90]. Furthermore, imputation provides a high-resolution picture of an associated region and increases the probability that a causal SNP can be precisely identified. Imputed SNPs which are strongly associated can be preferred candidates for replication studies. A limitation that affects the accuracy of fine-mapping is the amount of LD in the region [52]. For example, if SNPs are in strong LD and each is strongly associated with the phenotype, it is possible that one is causal but it is difficult to distinguish which one. In this case, high-LD SNPs are likely to "share" the total posterior probability of association while each SNP has only a modest posterior probability [85]. We proceed by describing different approaches which have notably contributed to build the literature around these aspects of genetic studies.

A Bayesian model averaging approach is proposed by Ferreira and Marchini in [24]. Different models are presented which allow for interactions of each SNP tested with established loci. They span the range of possible ways of association in a parsimonious way and they include from only marginal effects to simple and more complex interactions. Averaging over plausible models overcomes the need to fix on a particular model or to fit varied models of interaction. In the calculation of a single Bayes factor, the marginal likelihood under the alternative hypothesis is a sum of each individual model's marginal likelihood weighted by its prior. The marginal likelihood is obtained using a Bayesian linear model.

In a recent follow-up to the WTCCC studies [98], 14 associated regions in 3 diseases (type 2 diabetes, coronary artery disease and Graves' disease) are investigated further to refine associations. A Bayesian approach is adopted and under the assumptions that there is a single causal SNP in the region and that it is typed in the study, two measures are derived.

Based upon a uniform prior on the SNP's causality the region Bayes factor is calculated as the sum of each single-SNP Bayes factor. This is then used to derive the posterior probability of a given SNP being causal. Deviation from additivity is also tested by comparing the Bayes factor of the general model versus the additive model. Evidence for the involvement of multiple SNPs is detected in 3 regions regarding type 2 diabetes with the strength of the genetic effect and the heritability explained by the locus increased.

Multiple mutations are tested through analyzing the genealogy of the case-control sample using **GENECLUSTER** [98]. **GENECLUSTER** estimates the genealogy of a case-control sample in an attempt to extract signals of association and to estimate allelic heterogeneity at a locus. On each branch of the tree a mutation is considered and a population genetics model is employed to predict the likely genotypes at this SNP in each individual. Evidence for association with disease status is subsequently checked using a Bayes factor. The uncertainty of the predicted genotypes is accounted for by averaging over the Bayes factor for each branch. Further, two mutations (or more) can be placed on the genealogical tree to assess association for a pair of SNPs. By comparing the 2-mutation and 1-mutation Bayes factors, we can test for the presence of allelic heterogeneity [88].

A Bayesian variable selection procedure proposed by Guan and Stephens is proven to produce reliable results using relatively simple MCMC techniques for problems even as large as a GWAS. Particular emphasis is placed on the proportion of variance explained in the hope that it will address the issue of "missing" heritability. The goal is to identify genomic regions containing SNPs affecting the phenotype by examining regions where SNPs are most likely to have non-zero coefficients. A fairly standard regression framework is employed where the genotype of an individual is coded in an additive manner. The standard approach on prior specifications which imposes the same prior information on the regression coefficients is bypassed and stronger shrinkage is applied on more complex models. This avoids the assumption that models with many covariates have larger proportion of variance explained. Hence, care is taken to control the sparsity of the model and the typical

size of the non-zero regression coefficients aiming to impose a uniform prior on the proportion of variance explained. A Metropolis-Hastings algorithm is used with a simple local proposal density to obtain samples from the posterior distribution leading to the estimation of the posterior inclusion probability for each covariate using Rao-Blackwellization techniques. The proportion of variance explained is also estimated using the sampled values from the MCMC scheme [32].

A Bayesian method called *Sparse Partitioning* for identifying causal predictors without placing any restrictions on how they may influence the response, is presented by Speed and Tavaré [80]. The method is designed for problems involving binary or tertiary predictors and allows the number of predictors to exceed the size of the sample. Let us assume that $\boldsymbol{Y}$ and $\boldsymbol{X}$ denote the response vector and predictors' matrix respectively. If we specify a link function l then the regression model can be written as $l(\mathbb{E}(\boldsymbol{Y})) = f(\boldsymbol{X})$ where $f(\boldsymbol{X})$ is influenced by additive contributions from groups of interacting predictors. Sparse partitioning is described as a search for high scoring partitions. The underlying relationship can be written as:

$$\begin{aligned} f(\boldsymbol{X}) &= f_1(X_{G_{1,1}}, ..., X_{G_{1,s_1}}) + ... + f_K(X_{G_{K,1}}, ..., X_{G_{K,s_K}}) \\ &= f_1(X_{\boldsymbol{G}_1}) + ... + f_K(X_{\boldsymbol{G}_K}), \end{aligned}$$

where f_k describes the contribution of predictors $G_{k,1}, ..., G_{k,s_k}$ to $f(\boldsymbol{X})$. The sets $\boldsymbol{G}_1, ..., \boldsymbol{G}_K$ represent groups of associated predictors and $\boldsymbol{G}_0$ includes the non-associated predictors defining $\mathbb{G} = \{\boldsymbol{G}_0, \boldsymbol{G}_1, ..., \boldsymbol{G}_K\}$. The predictors in each group are said to interact with each other but not to interact with a predictor in a different group. To overcome the restriction of not allowing interactions between predictor groups, each predictor is allowed to appear multiple times. Another way of describing a partition is by the vector $\mathbf{I} = (I_1, I_2, ..., I_N)$, where I_g denotes to which group predictor g belongs. The method focuses on identifying which predictors belong in any of the groups and whether predictors appear in the same group indicating an interaction. Each model is defined by a partition $\mathbb{G}$ and a correspond-

ing set of functions $\boldsymbol{f}$. Nonetheless, $\boldsymbol{f}$ is considered a nuisance parameter (which can be saved for follow-up analysis) and more interest is placed on $\mathbb{G}$. The posterior probability $P(\mathbb{G}|Data)$ is required to be calculated but this would require an exhaustive search of the space of partitions hence an MCMC technique is employed. A Gibbs sampling algorithm is used to first propose a change to each component of $\mathbf{I}$ and then a change to one element of $\mathbb{G}$.

BIMBAM by Servin and Stephens, which was briefly introduced in the previous section where we described GWASs, is also applicable for candidate gene studies which can contain multiple causal variants. We now take a closer look at how this framework is constructed. As an imputation-based method, knowledge of reference data and genotype data at tag SNPs collected on a phenotype study sample are used to estimate unmeasured genotypes. This is feasible by several imputation algorithms publicly available and the authors propose **PHASE** [86] and **fastPHASE** [73]. The phenotypes are then modelled by a standard linear regression:

$$y_i = \mu + \sum_j \boldsymbol{x}_{ij}\beta_j + \epsilon_i,$$

where y_i is the phenotype for individual i, μ is the phenotype mean of individuals carrying the reference genotype, the x_{ij}s are the elements of a design matrix which depends on the genotype data, the β_js are the corresponding regression coefficients and ϵ_i is a residual. The general genetic model is fitted across SNPs. Two separate prior specifications are described for the model parameters denoted as D_1 and D_2. The priors on μ and τ, the phenotypic variance are on the one hand Jeffrey's and on the other conjugate priors. The prior distribution on the SNP effects has two components: a prior on which SNPs are causal and a prior on the effects of these causal SNPs. The priors D_1 and D_2 regarding the effect sizes differ in the way they handle the dominance effect. Prior D_1 assumes that the prior probability of overdominance is independent from the additive effect. However, an MCMC scheme needs to be employed for the calculation of posterior quantities. The convenience of D_2 is that posterior probabilities can be computed analytically. With regards to the prior

on the number of multiple causal SNPs it is suggested to concentrate the mass on models with relatively few regressors. To measure the evidence of association, a Bayes factor is obtained by:

$$BF = \frac{P(\boldsymbol{y}|\boldsymbol{G}, H_1)}{P(\boldsymbol{y}|\boldsymbol{G}, H_0)},$$

where H_0 is the null hypothesis of no associations and H_1 is the complementary event. In the special case where only a single association is allowed, the Bayes factor takes the form of a mean of the single-SNP measures:

$$BF = \frac{1}{n}\sum_{j=1}^{n} \frac{P(\boldsymbol{y}|\boldsymbol{G}, H_j)}{P(\boldsymbol{y}|\boldsymbol{G}, H_0)},$$

where in this case H_j forms the hypothesis that the j^{th} SNP is associated, where the total number of SNPs is n [75].

The final multifactorial approach we examine is called **BAMSE** and is based on an MCMC algorithm and allows for the identification of sets of SNPs and environmental factors that when combined increase the risk of a disease [3]. The method explores sets of effects called risk sets that increase the risk for individuals who fulfill the characteristics defined by the sets while combining information from different models by evaluating the effect of all possible interactions. The j^{th} risk set, T_j, is a subset of the parameter space that partitions the individuals on the basis of their genotypes and environmental factors. Individuals whose characteristics comply with T_j belong to the risk group G_j. The quantitative trait for individual i, y_i, is modelled as:

$$y_i \sim \sum_{j=0}^{n_m} I_{i \in G_j} N(\alpha_j, \sigma^2),$$

where n_m is the number of risk sets, α_j is the mean of risk set, j and σ^2 is the variance, equal across risk sets. The state space contains the risk sets, their means and σ. Each risk set defines a genetic and environmental profile and for each continuous risk parameter a

threshold can be defined that excludes individuals with observations above this value. An additional factor to be considered is the situation where an individual may fit several risk sets. In this case, an individual is assigned to the risk group with the highest mean. The notion of active components in a risk set is introduced since it is thought that not all characteristics are bound to affect the trait in a causative combination. The specification of priors is as follows. The means, α_j, are normally distributed with mean equal to $\bar{y}$ and variance proportional to R^2. The prior on the variance is uniformly distributed on the positive real line. Furthermore, the priors for the number of active risk components and the number of risk sets are geometrically distributed. A Metropolis-Hastings algorithm is formulated to search the model space and the parameters are updated one at a time. However, for the number of active components or the number of risk sets, several parameters must be proposed simultaneously. The posterior probabilities of a certain risk set and the number of risk sets can be approximated and used for inference [3].

1.3.1 The Importance of Haplotype Information

Thus far we have discussed several methodologies which are used to discover associations either at single or multiple SNPs based on genotype data which are either typed directly via a genotyping platform or imputed using any of the many imputation algorithms available in the literature such as **IMPUTE** and **PHASE**. The diploid nature of the human genome is often ignored since sequencing studies do not routinely separate or "phase" maternally and paternally derived sequence information. Large scale projects such as The HapMap Project and The 1000 Genomes Project have utilized the importance of haplotyping. Haplotypes are used to assess, with the aid of linkage disequilibrium measures, the likelihood that variants at one genomic position indicate the presence of variants at flanking positions. The absence of phased data is mostly due to the computational complexity associated with phase information and the lack of cost-effective approaches [93].

The ancestral structure captured in the distribution of haplotypes sometimes provides larger

power than single-marker tests for genetic disease associations. For example, in case-control studies, haplotype-based methods can increase power especially when the marker alleles are not in strong LD with each other but in strong LD with the causative alleles. Haplotypes can be very helpful especially for fine mapping because the ancestral recombination trims haplotypes to informative segments that include a disease locus. Such methods take advantage of the fact that in the neighbourhood of a causal locus, haplotypes of affected subjects share more ancestry than haplotypes of unaffected individuals and this decreases with distance from the causative locus. The goal is essentially to pinpoint the disease susceptibility locus [72]. Aside from the gain in power, Clark in [14] discusses two more reasons for which haplotypes should be considered an improvement. At each protein coding gene an individual will only produce at most two polypeptide chains, one corresponding to the maternal and one to the paternal haplotype. Physical properties of the protein can depend on simple or higher-order interactions of amino-acid sites. If these interactions are important, haplotypes have biological importance. Secondly, population genetic theory shows us that variation is inherently structured in haplotypes. A new mutation on a haplotype can increase the frequency by random genetic drift and it may then be divided into segments by recombination. The blend of mutation, drift, selection, migration/population mixing and recombination results in genetic variation that has a strong piecewise haplotype structure.

A method developed to analyze SNPs and assess disease associations in a candidate region and estimate corresponding haplotype relative risks under a case-control sample is presented by Morris in [56]. It is designed for use in candidate genes or small candidate regions and the likelihood is defined within a logistic regression framework. Unphased SNP genotype data are directly analyzed and given these, maximum likelihood estimates are obtained for the relative frequencies of the haplotypes. Using the Bayesian partition model, cluster centres are specified with each cluster given a disease odds parameter. The haplotypes are then assigned to the "nearest" centre, where similarity is calculated in terms of marker allele matches. Uncertainty in phase information is accounted for by all the pos-

sible haplotype configurations consistent with each genotype. A reversible jump MCMC algorithm is developed to sample from the space of haplotype clusters and corresponding odds and its output is used to estimate relative-risks of disease, identify clusters of haplotypes with similar such risks, identify groups of cases carrying the same high-risk variants and estimate the posterior probability of haplotype association.

This approach is later generalized in [57] to allow for more flexible modelling of marker-SNP haplotype association with disease and to allow for dominance effects. Thus, the logistic regression model includes the additive effect and the dominance effect of causal variants over other alleles at the functional polymorphisms. Results over simulations provide gain in power over existing haplotype-based tests which ignore dominance effects and clustering. The specification of the model follows the lines of [56] with the addition of a parameter, which is unconstrained with respect to the values it can take, in the Bayesian partition model. Furthermore, to evaluate the evidence against the model of no association, a Bayes factor is calculated. A Metropolis-Hastings algorithm is employed to obtain samples of the posterior probability of the parameters, where the proposal density proposes small changes to the current state at each iteration. The marginal likelihood can be obtained by integrating over the model parameters is approximated by the inverse of the average of the likelihood function recorded on each output of the MCMC chain. The Bayes factor is also used to estimate the posterior probability of association.

Another approach which performs fine mapping of disease genes via haplotype clustering is presented in [102]. Haplotype clusters are defned using a similarity score of an observed haplotype with a hypothetical ancestral haplotype. The values of this measure differ according to the allele match between the ancestral and observed haplotype. The number of an individual's haplotypes in the cluster produces a prediction of the individual's number of copies of the disease predisposing allele. An MCMC algorithm is then implemented to search over locations and ancestral haplotypes to identify large, case-rich clusters.

Numerous studies have used haplotype-based approaches to fine-map associations for dif-

ferent diseases. For example, a comprehensive haplotype analysis of the CYP19 gene is performed to assess whether common variation in this gene is associated with risk of hormone-related diseases and breast cancer. A haplotype-based analysis was employed to search for common disease associated variants among women in the Multiethnic Cohort Study. Twenty five haplotype-tagging SNPs were selected to predict the common haplotypes with high probability and genotyped in a breast cancer case–control study. The findings suggested that women with the long-range CYP19 haplotype can be carriers of a predisposing breast cancer susceptibility allele [33]. Haplotype associations in the VEGF gene were assessed with lung carcinoma. Analysis revealed that haplotypes of the VEGF gene conveys the risk for lung cancer. SNPs examined appear to influence lung cancer susceptibility while associations were not significant in expression levels [61]. Other studies which carry out fine mapping in order to identify susceptibility loci include diseases such as type 2 diabetes and myocardial infraction [76], primary biliary cirrhosis [45] and Huntington disease [41].

1.3.2 Compound Heterozygosity

The importance of phase information is emphasized for established phenomena for relating genotype to phenotype. Such a phenomenon is *compound heterozygosity*. Numerous studies which investigate the genetic mechanisms that underlie variation, mostly in the context of monogenic Mendelian disorders, have implicated the event of compound heterozygosity [77, 106, 55]. It occurs when the two homologous copies of a genomic region harbour distinct sequence variants respectively but at different positions in the region (Figure 1.1). This variation is assumed to alter the function of the two homologous copies of a gene in different ways while their joint molecular effects result in a different phenotype from that observed if one homologous gene carries both harfmul variants. Therefore, knowing that an individual is heterozygous for variants at specific loci is not enough to detect such an effect; the knowledge about the specific diplotype is crucial. It is not obvious how often compound

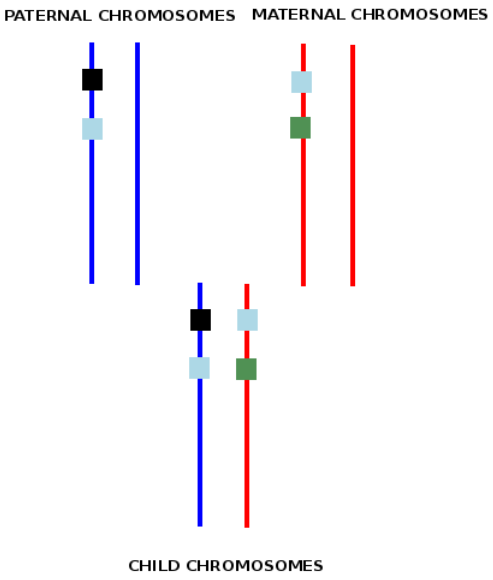

Figure 1.1: The diagram illustrates the concept of compound heterozygote effects. The black and green squares represent two deleterious variants in different positions in a gene. The black variant is passed on to the child from the father and the green from the mother. Each deleterious variant is heterozygous. In the child, both copies of the gene are defective due to the two deleterious variants.

heterozygosity is likely to influence different diseases and conditions. Nonetheless, the fact that there are many instances in which such an effect is suggested means that it should be taken into account, and the only way to do so is by considering haplotype information [93].

There are many examples of diseases or conditions influenced by compound heterozygosity in single genes. We take a look at some of them. A study on epidermolysis bullosa simplex associated muscular dystrophy (EBS-MD) was performed on ten patients. This is an autosomal recessive disorder caused by genetic defects in the plectin gene. It is a heterogeneous group of genodermatoses in which blistering of the skin occurs as a result of mechanical trauma of varying degree. Among the ten unrelated patients with EBS-MD, six were male and four were female with an age range of 3 to 46 years old. The results showed that 70% were products of consanguineous marriage and homozygous for the mutations, whereas 30% were compound heterozygous for plectin gene mutations. All patients had generalized blistering at birth and experienced severe late-onset MD, apart from one patient who did

not have blisters at birth and whose muscle strength was only slightly affected at 46 years old. This patient had a homozygous deletion mutation in the plectin gene [77].

Another study considered fructose-1,6-bisphosphatase deficiency caused by a mutation of the FBP1 gene which resulted in impaired gluconeogenesis. A patient with hypoglycemia was examined and sequencing was performed on the entire FBP1 coding region of the patient and his family members. Direct and allele-specific sequence analysis of the FBP1 gene showed that the patient had a compound heterozygote for the G164S and 838delT mutations, which were inherited from his carrier parents. His father and mother had heterozygous 838delT and G164S mutations respectively without any symptoms of hypoglycemia [55].

The most common cause of genetic deafness are mutations in GJB2 gene and this has been indentified in many populations. One of them, W44C, was initially detected in two families from France and subsequently a family in the United States. Two siblings were found to be compound heterozygous for the W44C and K15T mutations. Their father, the original proband, was heterozygous for the dominant W44C mutation and their mother was compound heterozygous for two recessively inherited mutations, K15T and 35delG. The parents would, essentially, have a 50% chance of passing on both a dominant and a recessive mutation in GJB2 which would result in a deaf child. Both children were deaf and use manual communication while their parents could communicate orally since their hearing loss was not as severe [106].

Two mutations, C282Y and H63D, in the HFE gene on chromosome 6 play an important role in iron metabolism. In a population-based study of the elderly, iron parameters were significantly increased in subjects homozygous, heterozygous and compound heterozygous. Linear regression models were fitted for each gender separately. The effect of compound heterozygotes was included as interaction between the two mutations. Compound heterozygosity was associated with a significant effect on all serum iron indices except from one of them in women [64].

Retinitis pigmentosa (RP) is a heterogeneous group of progressive retinal degenerations. It affects 1 in 3,500 in the general population and the symptoms include blindness, tunnel vision and bone-spicule pigmentation in retina. A study reported twenty two subjects from a four-generation Chinese family with RP and used exome sequencing to analyze the exomes of three patients and their mother. It identified 32,216 SNPs and 2,477 indels that were shared by the three patients. As variants satisfying a recessive homozygous inheritance model were not detected, recessive compound heterozygous inheritance was investigated. One heterozygous mutation, CYP4V2 c.1091-2A>G, was identified in all affected individuals but was not heterozygous in the mother's exome, hence inherited from the father. The other pathogenic heterozygous variant, c.802-8_810del17insGC, inherited from the mother was identified by re-filtering the exome sequencing data for variations present in all patients and the mother. These mutations were previously identified also to be associated with Bietti crystalline corneoretinal dystrophy, an autosomal recessive eye disease [103].

Additional studies include $\beta-$thalassemia [28, 91] and erythermalgia [71] while a further list of conditions driven by compound heterozygote effects can be found in [93].

1.4 Analysis of Multiple Cohorts

Following the identification of a causal SNP or combination of SNPs affecting a phenotype in a single study, a natural development is the further investigation of the findings in different populations or study designs. A useful method for combining the results of different studies on the same research problem and for quantifying and explaining the extent of potential inconsistencies between studies is *meta-analysis*. It can be used to increase confidence and boost power in a noticeable association as well as to investigate heterogeneity of the effect sizes and genetic models. Given that most genetic effects are small and require the collection of many groups to generate large-scale evidence, the development of meta-analyses of data from GWA studies has increased remarkably after the development of different ways to share data and results.

There are two commonly used procedures to execute a meta-analysis and these regard the assumption made for the cohort specific effect sizes. On the one hand, one assumes that the cohort specific effect sizes are homogeneous across studies and all differences are due to chance. This will lead to the use of a *fixed effects model*. On the other hand, genetic effect sizes may vary across studies because of authentic differences in environmental exposures and genetic backgrounds or, possibly, differential biases hence directing to the use of a *random effects model* which incorporates heterogeneity amongst studies. Beyond genuine diversity in the genetic effects, Kavvoura and Ioannidis in [38] describe sources of between-study heterogeneity. Amongst others, one possible reason of heterogeneity is that a polymorphism derived from a GWAS might not be the truly associated one, but simply highly correlated to the actual association. This is somewhat of a rule in association studies where tag polymorphisms are usually selected for testing based on frequency, redundancy or coverage considerations. Along the same lines, the phenotype under study may be linked with the truly associated phenotype and this correlation can vary across studies. Diversity in effects across populations may also reflect population-specific gene-gene or gene-environment interactions.

The traditional approach to performing a fixed effects model is the inverse-variance-weighted effect size estimate [34]. Let us denote $X_1, ..., X_N$ the effect size estimates in N independent studies. Usually, these follow normal distributions if the sample sizes in each study are large enough. Let $SE(X_i)$ be the standard error of X_i, $V_i = SE(X_i)^2$ and $W_i = V_i^{-1}$, the inverse variance. The inverse-variance-weighted effect-size estimator is:

$$\bar{X} = \frac{\sum W_i X_i}{\sum W_i}$$

and its associated standard error is $SE(\bar{X}) = \left(\sqrt{\sum W_i}\right)^{-1}$. Since $\bar{X}$ follows a Normal distribution the statistic Z_{FE} is as follows:

$$Z_{FE} = \frac{\bar{X}}{SE(\bar{X})} = \frac{\sum W_i X_i}{\sqrt{\sum W_i}},$$

which follows a standard normal distribution under the null hypothesis of no association and the p-value can be calculated accordingly. The inverse-variance-weighted effect-size estimator is the best to use as it has minimum variance when the weights are chosen to be proportional to the inverse of the variance of the cohort-specific effect sizes. We prove this as follows. Let

$$Var[\bar{X}] = F(W_1, W_2, ..., W_N) = \frac{\sum W_i^2 Var(X_i)}{\left(\sum W_i\right)^2}.$$

We apply the constraint $\sum W_i = 1$ and define $G(W_1, W_2, ..., W_N) = \sum W_i - 1$. We make use of the Lagrange multiplier method and define the function H as follows:

$$H(W_1, W_2, ..., W_N) = F(W_1, W_2, ..., W_N) + \lambda G(W_1, W_2, ..., W_N).$$

The minimum of F subject to the constraint G can be found by equating the partial deriva-

tives of H with respect to each W_i to zero. For a study j

$$\frac{\partial H}{\partial W_j} = \frac{\partial}{\partial W_j} \sum W_i^2 Var(X_i) + \lambda \frac{\partial}{\partial W_j} \left(\sum W_i - 1 \right).$$

This partial derivative is equal to zero when $W_j \propto 1/Var(X_j)$. The second derivative of H with respect to W_j is positive or zero hence the solution gives the minimum variance estimate.

Before initializing a fixed effects meta-analysis, one should indeed test the assumption of homogeneity between study specific effect sizes. Commonly used heterogeneity metrics are summarized in [38]. The most popular is *Cochran's Q* statistic. It has been said, though, that the statistic is not able to detect heterogeneity when the number of studies under examination is small but it may over-estimate unimportant heterogeneity when the number of studies is large [36]. It can be estimated as follows:

$$Q = \sum W_i (X_i - \bar{X})^2,$$

where $\bar{X}$ is the summary effect size as obtained above, X_i for $i = 1, ..., N$ are the study-specific effect sizes and W_i is the weight of study i defined as the inverse variance of X_i. Under the null hypothesis of homogeneity the statistic approximately follows a χ^2 distribution with $N - 1$ degrees of freedom.

Another common heterogeneity metric is the *between-study variace* τ^2. It depends on the respective effect size metric used (for example standardized mean difference or odds ratio), therefore it is not comparable among meta-analyses which utilize different metrics. It can be estimated as:

$$\tau^2 = \frac{Q - (N - 1)}{\sum W_i - \frac{\sum W_i^2}{\sum W_i}},$$

where Q is the Cochran's statistic. This metric essentially reflects the difference between the estimates of the true effect sizes in each study.

Finally, Higgins and Thompson [36] proposed, amongst others, the I^2 metric. It is independent of the number of studies and can be compared across meta-analyses with different number of studies and metrics. It describes the percentage of total variation across studies due to heterogeneity. Nonetheless, I^2 can be uncertain when the number of studies in a meta-analysis is small. It is calculated as:

$$I^2 = \frac{Q - (N-1)}{Q}.$$

For a value of 50%, it can be said that large heterogeneity is present while 75% implies very large heterogeneity. A potential disadvantage is that the confidence intervals are very large, especially when the number of studies is small [38, 60].

Although in practice, researchers often apply the fixed effects model in all cases irrespective of homogeneity or not, the commonly applied method of analyzing a collection of studies in the presesnce of heterogeneity is the *random effects model*. The random effects assumption states that the true value of the effect size is sampled from a probability distribution having variance τ^2 as defined above. This parameter can be estimated by the method of moments, the method of maximum likelihood and the method or restricted maximum likelihood [34]. Once the estimated between-study variance $\hat{\tau}^2$ is obtained, the combined effect size with respect to the random effects model is:

$$\bar{X}^* = \frac{\sum (W_i + \hat{\tau}^2)^{-1} X_i}{\sum (W_i + \hat{\tau}^2)^{-1}}$$

where $W_i = 1/SE(X_i)^2$. Then the associated standard error is $SE(\bar{X}^*) = \sqrt{\sum (W_i + \hat{\tau}^2)^{-1}}^{-1}$. The test statistic follows:

$$Z_{RE} = \frac{\bar{X}^*}{SE(\bar{X}^*)}.$$

Neupane et al. [62] describe alternatives to the traditional tests which are named global fixed and random effects respectively. These were previously given by Lebrec et al. [40].

They result in higher power to detect very small to small associations of common variants when large heterogeneity is present across cohorts. Advantage over the traditional approaches is also obtained for less common variants only when heterogeneity is substantial. The global fixed effects model tests whether an association exists in any cohort hence the two hypotheses are:

$$H_0 : \beta_1 = \beta_2 = ... = \beta_N = 0;$$
$$H_1 : \beta_i \neq 0 \quad \text{in at least one cohort } i.$$

A score test is then used to assess the hypothesis (X_i is the estimate of the parameter β_i):

$$T = \sum_{i=1}^{N} \left(\frac{X_i}{SE(X_i)} \right)^2,$$

which asymptotically follows a χ^2_N under the null hypothesis. The new random effects model tests for a non-null average association (μ) or the between-cohort variance (τ^2) is non-zero, that is:

$$H_0 : \mu = 0 \quad \text{and} \quad \tau^2 = 0;$$
$$H_1 : \mu \neq 0 \quad \text{or} \quad \tau^2 > 0.$$

The likelihood ratio test can be used to assess the hypotheses:

$$T = 2 \times [l(\hat{\mu}, \hat{\tau}^2) - l(0, 0)] \sim (\chi^2_1 + \chi^2_2)/2 \quad \text{under } H_0.$$

For large cohorts we assume that $X_i|\beta_i \sim N(\beta_i, SE(X_i)^2)$. As $\beta_i \sim N(\mu, \tau^2)$, the approximate marginal distribution of the estimate in the i^{th} cohort is $X_i \sim N(\mu, SE(X_i)^2 + \tau^2)$ with the corresponding log-likelihood $l_i(\mu, \tau^2)$ and the total log-likelihood $l(\mu, \tau^2) = \sum_{i=1}^{N} l_i(\mu, \tau^2)$.

Han and Eskin in [34] propose a random effects model that relaxes the conservative assumption in hypothesis testing. It proceeds with the traditional approach in calculating the effect size and its confidence interval but instead of calculating a z-score, a likelihood ratio test is applied where each statistic is a single observation and no heterogeneity is assumed under the null hypothesis. We use the same notation as above therefore $X_1, X_2, ..., X_N$ are the effect-size estimates in N studies and V_i is the variance of X_i. Let μ be the unknown true mean effect size and τ the heterogeneity parameter. Under the null hypothesis of no association both of these parameters are set to zero and the likelihoods under the null and alternative hypotheses are:

$$L_0 = \prod_i \frac{1}{\sqrt{2\pi V_i}} \exp\left(-\frac{X_i^2}{2V_i}\right);$$
$$L_1 = \prod_i \frac{1}{\sqrt{2\pi(V_i+\tau^2)}} \exp\left(-\frac{(X_i-\mu)^2}{2(V_i+\tau^2)}\right).$$

The maximum likelihood estimates $\hat{\mu}$ and $\hat{\tau}^2$ are obtained via the following iterative procedure. Given, say, the current estimates $\hat{\mu}_{(n)}$ and $\hat{\tau}^2_{(n)}$ then the $(n+1)^{th}$ estimates are calculated via:

$$\hat{\mu}_{(n+1)} = \frac{\sum \frac{X_i}{V_i+\hat{\tau}^2_{(n)}}}{\sum \frac{1}{V_i+\hat{\tau}^2_{(n)}}}, \quad \hat{\tau}^2_{(n+1)} = \frac{\sum \frac{\left(X_i-\hat{\mu}_{(n+1)}\right)^2-V_i}{\left(V_i+\hat{\tau}^2_{(n)}\right)^2}}{\sum \frac{1}{\left(V_i+\hat{\tau}^2_{(n)}\right)^2}}.$$

When the iterative procedure converges to the maximum likelihood estimates, the likelihood ratio test can be obtained as follows:

$$S_{New} = -2 \times \log(\lambda) = \sum \log\left(\frac{V_i}{V_i+\hat{\tau}^2}\right) + \sum \frac{X_i^2}{V_i} - \sum \frac{(X_i-\hat{\mu})^2}{V_i+\hat{\tau}^2}.$$

The statistical significance can be best assessed by an asymptotic distribution. Since μ is unrestricted and τ^2 is positive, μ corresponds to a normal distribution and τ^2 to a half of a normal distribution. Then, the statistic asymptotically follows an equal mixture of

1 degree of freedom and a 2 degrees of freedom χ^2 distribution. To overcome potential misinterpretations, with regards to the number of studies, tabulated values are provided. An interesting aspect of this statistic is that it can be decomposed into two parts. The first part relates to the fixed effects statistic while the second part is a statistic which can be used to test the same hypothesis as the Cochran's Q statistic of heterogeneity. This can assist the interpretation because one can assess the proportion of the statistic due to mean effect and to heterogeneity respectively.

In an attempt to interpret further the results of a meta-analysis, which might well include cohorts which do not exhibit an effect at each variant under study, Han and Eskin in [35] describe a new statistic representing the posterior probability that the effect exists in each study, which is estimated using cross-study information. In addition, a new association testing procedure can be applied which accounts for the existence of effect. The new framework makes two assumptions. The first assumption states that effect is either present or absent in the studies. The second states that if the effects exist, these are similar between studies. These form the *binary effects assumption*. The posterior probability that the effect exists in each study of a meta-analysis is called the *m-value*. Let $X_1, ..., X_N$ be the observed effect sizes for N studies and V_i the estimated variance of X_i then:

$$
\begin{aligned}
P(X_i|\text{no effect}) &= N(X_i; 0, V_i) \\
P(X_i|\text{effect}) &= N(X_i; \mu, V_i) \\
\mu &\sim N(0, \sigma^2) \\
T_i &\sim Bernoulli(\pi) \\
\pi &\sim B(\alpha, \beta).
\end{aligned}
$$

The vector $T = (T_1, ..., T_N)$ is the vector of values indicating which studies show an effect; T_i takes the value 1 if study i has an effect and zero otherwise. The vector T can take 2^N different values and U denotes this set. The m-value for cohort i, m_i, is defined using

Bayes' theorem:

$$\begin{aligned} m_i &= P(T_i = 1|X_1, ..., X_N) \\ &= \frac{P(X_1, ..., X_N|T_i = 1)P(T_i = 1)}{P(X_1, ..., X_N|T_i = 0)P(T_i = 0) + P(X_1, ..., X_N|T_i = 1)P(T_i = 1)} \\ &= \frac{\sum_{t \in U_i} P(X_1, ..., X_N|T = t)P(T = t)}{\sum_{t \in U} P(X_1, ..., X_N|T = t)P(T = t)}, \end{aligned}$$

where U_i is a subset of U whose elements' i^{th} value is 1. After deriving the terms of the above expression, the prior probability of T and the probability of $X_1, ..., X_N$ given T, a Metropolis-Hastings algorithm is employed in order to calculate the m-value as its exact calculation would require iteration over all t which is exponential to N. The m-value has a valid probabilistic interpretation and a threshold can be imposed in order to classify a study as having an effect or not. The authors suggest that if the m-value is greater than 0.9 then the study has an effect while if it is less than 0.1 the study is not associated. Further, to test association a new model is proposed called the *binary effects model*. It is a weighted sum of z-scores method where the m-values are incorporated into the weights which essentially assigns a greater weight to the studies having an effect. The binary effects model statistic can then be defined as:

$$Z_{BE} = \frac{\sum m_i \sqrt{W_i} \frac{X_i}{\sqrt{V_i}}}{\sqrt{\sum m_i^2 W_i}}.$$

Under a simulation study, the m-value has proven to be more powerful than a p-value or a Bayes factor in identifying correctly the studies both with and without an effect. Additionally, the power of the binary effects model to detect association increases for decreasing number of associated cohorts in a meta-analysis. In the presence of heterogeneity in the effect sizes, the binary effects model is generally less powerful than the random effects model but outperforms the traditional fixed effects model.

The notion of "missing heritability" can also be addressed in a meta-analysis level. One source of genetic variation is the differentiation between males and females, both in magnitude or direction. Such gender-differentiated effects are becoming increasingly appar-

ent in complex traits. Magi, Lindgren and Morris in [46] review the methodology for meta-analysis of gender-specific genome wide association studies and propose a gender-differentiated test of association allowing for heterogeneity of genetic effects between genders. Let the allelic effect of study i at SNP j be b_{ij} and the corresponding standard error s_{ij}. Additionally, k_i takes two values: 1 if the study is male-specific and 0 if the study is female-specific. Using a fixed effects model setting the respective test statistics and their associated variances are:

$$B_{M_j} = \frac{\sum_i k_i b_{ij}/s_{ij}^2}{\sum_i k_i/s_{ij}^2} \qquad V_{M_j} = \left[\sum_i k_i/s_{ij}^2\right]^{-1}$$
$$B_{F_j} = \frac{\sum_i (1-k_i) b_{ij}/s_{ij}^2}{\sum_i (1-k_i)/s_{ij}^2} \qquad V_{F_j} = \left[\sum_i (1-k_i)/s_{ij}^2\right]^{-1}.$$

Gender-specific tests of association at the j^{th} SNP are given by $X_{M_j}^2 = B_{M_j}^2/V_{M_j}$ and $X_{F_j}^2 = B_{F_j}^2/V_{F_j}$ where the test statistics both follow a χ^2 distribution with one degree of freedom. Further, the heterogeneity can be tested via gender-specific Cochran's Q-statistics:

$$Q_{M_j} = \sum_i \frac{k_i(b_{ij} - B_{M_j})}{s_{ij}^2} \quad \text{and} \quad Q_{F_j} = \sum_i \frac{(1-k_i)(b_{ij} - B_{F_j})}{s_{ij}^2},$$

the test statistics have χ^2 distributions with $\{(\sum_i k_i) - 1\}$ and $\{(\sum_i (1-k_i)) - 1\}$ degrees of freedom respectively. A gender-differentiated test of association can also be performed by combining the gender-specific test statistics defined above. The resulting test statistic, $X_{D_j}^2 = X_{M_j}^2 + X_{F_j}^2$, has a χ_2^2 distribution. Then, using $X_{D_j}^2$, one can test for heterogeneity between gender-specific allelic effects at the j^{th} SNP with the test statistic $X_{H_j}^2 = X_{D_j}^2 - X_{C_j}^2$ (following a χ_1^2 distribution), where $X_{C_j}^2$ is the traditional fixed effects statistic assuming the same allelic effect across both genders. This framework, offering a range of statistics, performs gender-specific meta-analyses which offer a substantial gain in power and should be used due to the evidence of the important role of gender-differentiated effects in the architecture of complex traits.

We now explore a novel methodology which allows data and results from different ethnic backgrounds to be combined in a unified meta-analysis framework. Transethnic meta-analysis offers an opportunity to gain power to detect novel loci to improve the resolution of fine-mapping of causal variants by combining differences in local LD structure between groups. Morris in [58] proposed a novel transethnic meta-analysis model which accounts for similarity in allelic effects between the most closely related populations and allows for heterogeneity between more diverse populations. The traditional fixed-effects model will clearly not account for such effects while the traditional random effects model assumes that each population has a different genetic effect. This might not capture the variability correctly as there is a possibility that populations from the same ethnic group exhibit more homogeneous effects than in more diverse groups. The novel methodology models this event by the means of a Bayesian partition model. Once the group specific allelic effects with their corresponding standard errors are obtained, groups are clustered according to their shared ancestry relatedness and allelic effects at the variant. Populations belonging in the same cluster are assumed to have the same effect while the between-cluster effects differ, hence allowing for heterogeneity. The ultimate goal is the calculation of the Bayes factor comparing the null model to an alternative model which is characterized by K, the number of clusters, C_K, the cluster centres and ψ, the allelic effects describing the clusters. The prior density function under a given model is defined to be a combination of different components relating to the parameters that make up a model configuration. The estimation of the marginal likelihood is achieved by the means of a Metropolis-Hastings algorithm since its direct calculation is not possible. Over R recorded MCMC outputs the marginal likelihood $f(b, s|M)$, where b is the collection of estimated allelic effects, s the associated standard errors and M is a model configuration, is approximated by

$$f(b, s|M) \approx \left[\frac{1}{R} \sum_{r=1}^{R} f(b, s|\theta^{(r)})^{-1} \right]^{-1},$$

the harmonic mean of sampled likelihood values, where $\theta^{(r)} = (K^{(r)}, C^{(r)}, \psi^{(r)})$. Output

from the MCMC chains can be used directly to approximate the posterior distribution of the allelic effect of each population as well as the posterior probability of heterogeneity in allelic effects between populations.

Meta-analysis is a vital tool in genetic epidemiology and has repeatedly proven to increase the evidence of causal variants. Numerous methodologies have been developed to facilitate such analysis under different characteristics in order to keep up with the availability of data, the different approaches to genome wide association analysis and fine mapping and the findings from these studies. These vary from frequentist to Bayesian, examining fixed and/or random effects, at single-SNP or multiple loci, for haplotype or genotype data. In addition to the different frameworks presented above, meta-analysis of haplotype association studies has been proposed in [4] as well as multipoint models for meta-analysis [5, 63]. Methodologies are publicly available implemented in different softwares, for example **PLINK** [68] and **META** [43] perform the traditional fixed and random effect meta-analysis while the latter is designed to work with **SNPTEST**'s output, **METAL** [107] calculates signed and overall z-scores as well as the inverse variance based statistic. **METASOFT** implements the fixed, random and binary effects models as decribed by Han amd Eskin [34, 35], **GWAMA** performs the gender-specific meta-analysis tests [46] and **MANTRA** encapsulates the trans-population meta-analysis methodology [58].

The published meta-analyses have increased rapidly over the past few years. They have enriched the evidence of causal loci and have further improved the power to do so. For example, a meta-analysis on six genome wide association studies increased the confirmed Crohn's disease loci to 71 where single-SNP z-scores showed a marked excess of significant associations [25]. A large scale analysis on the genetic architecture and pathophysiology of type 2 diabetes discovered ten previously, unreported susceptibility loci with two showing gender-differentiated effects [59]. Seven new risk alleles were identified, reaching genome-wid significance (p-value $< 5 \times 10^{-8}$), associated with rheumatoid arthritis loci, while two associations were refined and one was confirmed [83]. Another large scale meta-analysis

on 190,000 individuals aimed to identify additional susceptibility loci for coronary artery disease and escalated the total number to 46 [19]. Furthermore, multiple independent loci have been identified to affect smoking quantity. Two recent publications on meta-analysis have confirmed effects on smoking quantity at three genes encoding neuronal nicotinic acetylcholine receptor subunits [69, 44]. Additional diseases and conditions include kidney function [65], body mass index and glycemic traits [47], serum metabolite levels [39], thyroid function [66] and bipolar disorder [74].

1.5 Summary

Over the course of this chapter we explored the different methods available in the literature for the detection of associations on fine scale genetic studies which use genotype and/or haplotype data. Subsequently, we looked into the approaches used to combine results across studies and to carry out a meta-analysis. After the advance of genome wide association studies there has been a growing interest in refining and understanding better their results. Single-SNP analyses were soon succeeded by multipoint methods which search for combinations of variants which explain further the variation of the examined phenotype. The exploration of the variation along the genome via The HapMap Project and The 1000 Genomes Project has provided the researchers with plentiful data facilitating the development of novel methods which have shed light on the missing proportion of heritability. To further explore the findings of single-study association analyses, the procedure of meta-analysis is used to combine effects from many studies as an attempt to increase significance and boost power.

In the subsequent chapters we propose two novel methods which can be utilized in the analysis of genetic data. The first part describes a flexible and efficient model, suitable for fine mapping which analyses relatively small sets of haplotype data (up to a few thousand individuals and a few hundreds of variants). Using a Bayesian framework it seeks for the

most powerful combinations of SNPs and genetic models which explain association with the phenotype. The principal feature of our model is that it analyses haplotype data thus capturing additional biological information which is not observable unless we separate maternally and paternally derived sequence information. This property allows us to search for effects due to compound heterozygosity and at a later chapter we will show, using extensive simulations, that our new fine mapping approach has very large power to detect such effects and can outperform current popular approaches such as **SNPTEST**, **BAMSE** and **BIMBAM**.

The second part of this project describes a new framework for the accomplishment of meta-analysis using Bayesian hierarchical modelling. As a method for combining information from several parallel sources, meta-analysis is tightly linked to hierarchical modelling. The multiple parameters in a meta-analysis can be regarded as connected by the structure of the problem, implying that a joint probability model for these parameters reflects their dependence [29]. We set up a model space consisting of four models which capture homogeneity as well as heterogeneity between cohort-specific effect sizes accounting for possible non-association of some studies at each variant under consideration. The Gibbs sampler facilitates posterior inference while a choice of two straightforward approximations to the marginal likelihood allow us to calculate the Bayes factor to measure association. Detailed simulations show that this approach exhibits large power to detect the presence of fixed or random effects and outperforms analogous methods which also account for the existence of an effect in each cohort, for example **METASOFT**. Furthermore, we apply our method to a set of real samples to examine its ability to detect/confirm previously reported associations.

Part I

A Novel Model for Fine Mapping

Chapter 2

Methodology

We have seen in Chapter 1 that there exists an extensive collection of methods in the literature which facilitate the completion of genetic association studies of various characteristics, from quantitative traits to case-control studies, from typed to imputed variants, from frequentist approaches to more advanced bayesian methodologies. In this chapter we propose a new model suitable for fine mapping aiming at analyzing of a single cohort at each time described by relatively small datasets including a few thousand individuals typed at a few hundred SNPs. In a flexible and computationally tractable manner this model enables the detection of multiple loci that influence a quantitative phenotype of interest.

The novel approach we propose uses the Bayesian linear model framework where covariates are primarily haplotypes and not genotypes, as is the most frequent practise so far. This allows us to explore further underlying effects, which are not necessarily visible by just examining genotype data, for example compound heterozygote effects. Moreover we move away from the classic additive model and allow for a variety of genetic models to fit the haplotype data. These haplotype models consider interactions of the haplotype configurations at the different loci. Thus a variable selection procedure is employed, making use of a Markov Chain Monte Carlo algorithm in order to identify the set of variants that most likely affect the phenotype and how they do so.

A formal description of the model follows; it is, however, essential to introduce some notation which is used throughout the sections of this chapter and the ones that follow:

- S is the set of SNPs which are selected as causal. Its size ranges from 0 to 2;
- M_S is the genetic model which is selected to fit the SNPs in S. According to the cardinality of S different models are allowed to be fitted. The complete naming system of the models can be found in Appendix B.

2.1 The Bayesian Linear Model

The Bayesian approach to model fitting does not differ greatly from the Frequentist approach. The main differences involve assigning prior distributions to all the unknown parameters and determining the posterior distribution of the parameters given the data using Bayes Theorem [20].

The phenotype is related to the haplotypes (or genotypes) through the following relationship

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \epsilon, \tag{2.1}$$

where

- $\mathbf{Y} = (y_1, ..., y_n)$ is the phenotype vector for n individuals;
- $\mathbf{X}$ is the design matrix of the regression;
- $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_n)$ is a vector of normally distributed errors ($\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$);
- $\boldsymbol{\beta} = (\beta_1, ..., \beta_k)$ is the vector of parameters.

2.1.1 The Likelihood

The likelihood for a model is the joint probability of observing the data as a function of the model parameters. From Equation 2.1 we can denote the likelihood of the present model as

$$P(D|\boldsymbol{\beta}, \sigma^2) = N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}). \tag{2.2}$$

With reference to the probability density function of the multivariate normal distribution the full form of the likelihood is

$$P(D|\boldsymbol{\beta}, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-\frac{(Y - \mathbf{X}\boldsymbol{\beta})^T(Y - \mathbf{X}\boldsymbol{\beta})}{2\sigma^2}\right). \tag{2.3}$$

2.1.2 The Prior Distributions

In order to proceed in a Bayesian framework prior distributions have to be specified for the unknown quantities in the model. These are the parameter vector $\boldsymbol{\beta}$ and the regression variance σ^2. The choice of priors corresponds to the information available about the unknown parameters. For the sake of mathematical convenience we choose forms of distributions which lead to computationally tractable posterior distributions. This leads to the natural choice of conjugate priors. More precisely, in the situation of the Bayesian linear model the joint conjugate prior for $\boldsymbol{\beta}$ and σ^2 is the normal inverse-gamma(NIG) which we represent as follows:

$$\begin{aligned}
P(\boldsymbol{\beta}, \sigma^2) &= NIG(\boldsymbol{m}, \mathbf{V}, a, b) \\
&= P(\boldsymbol{\beta}|\sigma^2)P(\sigma^2) \\
&= N(\boldsymbol{m}, \sigma^2\mathbf{V})IG(a, b) \\
&= \frac{b^a}{(2\pi)^{k/2}|V|^{1/2}\Gamma(a)}(\sigma^2)^{-(a+(k/2)+1)} \\
&\quad \times \exp[-\{(\boldsymbol{\beta} - \boldsymbol{m})^T\mathbf{V}^{-1}(\boldsymbol{\beta} - \boldsymbol{m}) + 2b\}/(2\sigma^2)],
\end{aligned} \tag{2.4}$$

- $\boldsymbol{m}$ is the prior mean of the coefficients, $\boldsymbol{\beta}$;
- $\sigma^2\mathbf{V}$ is the prior variance of the coefficients, $\boldsymbol{\beta}$;
- a is the shape and b is the scale of the inverse-gamma distribution;
- k is the number of parameters in $\boldsymbol{\beta}$.

2.1.3 The Posterior Distribution

Using Bayes Theorem we can determine the form of the posterior distribution up to a normalising constant. Hence,

$$\begin{aligned} P(\boldsymbol{\beta}, \sigma^2|D) &\propto P(D|\boldsymbol{\beta}, \sigma^2) \times P(\boldsymbol{\beta}, \sigma^2) \\ &= (\sigma^2)^{-(a^*+(k/2)+1)} exp\left(-\frac{(\boldsymbol{\beta}-\boldsymbol{m}^*)^T(\mathbf{V}^*)^{-1}(\boldsymbol{\beta}-\boldsymbol{m}^*)+2b^*}{2\sigma^2}\right), \end{aligned} \tag{2.5}$$

where

$$\boldsymbol{m}^* = (\mathbf{V}^{-1} + \mathbf{X}^T\mathbf{X})^{-1}(\mathbf{V}^{-1}\boldsymbol{m} + \mathbf{X}^T\mathbf{Y}), \tag{2.6}$$

$$\mathbf{V}^* = (\mathbf{V}^{-1} + \mathbf{X}^T\mathbf{X})^{-1}, \tag{2.7}$$

$$a^* = a + n/2, \tag{2.8}$$

$$b^* = b + \{\boldsymbol{m}^T\mathbf{V}^{-1}\boldsymbol{m} + \mathbf{Y}^T\mathbf{Y} - (\boldsymbol{m}^*)^T(\mathbf{V}^*)^{-1}\boldsymbol{m}^*\}/2. \tag{2.9}$$

It is clear that the functional form of the posterior is the same as the functional for the prior in Equation 2.4. The difference lies in the four quantities, which are now updated, leading to the following posterior distribution:

$$P(\boldsymbol{\beta}, \sigma^2|D) = NIG(\boldsymbol{m}^*, \mathbf{V}^*, a^*, b^*). \tag{2.10}$$

2.1.4 The Design Matrix, X

In genetic association studies the design matrix of the regression model stores the information of each individual at each SNP. Depending on the genetic model M_S that is selected to fit the data the form and dimensions of **X** vary. The most common practice is to code the genotype of an individual at a specific SNP as 0, 1 and 2 copies of a particular reference allele. In this way we are assuming an additive genetic model. A more extensive model is the general genetic model where each genotype value contributes a different effect to the phenotype. These two are the only models which can be chosen to fit the data when the size of the SNP set S is 1. It would be straightforward to allow for dominant and recessive effects by changing slightly the coding of the alleles.

In the case where S contains 2 SNPs the additive and general models are available allowing for the both SNPs to be coded additively or generally. We extend the list of possible models to include a number of haplotype models, where information is extracted from the distinct haplotypes of each individual. Table 2.1 presents the haplotype combinations. The main idea behind these models is that we make the assumption of additivity of the haplotype risks. Initially, a particular haplotype combination is chosen as causal. Then, we simply consider how many copies of this causal haplotype configuration each individual has along with the number of copies of any of the remaining combinations. This idea generates four different models. The next model we consider can be characterised as the full haplotype model as each haplotype combination is assigned a respective genetic parameter. Table 2.2 describes the four single haplotype models with their respective names. Each model is assigned a different causal haplotype combination which is assigned a genetic effect β_1. The remaining three combinations are assigned a common effect of β_2. A comprehensive list of the models can be found in Appendix B.

H^1_{ji}	H^2_{ji}
0	0
0	1
1	0
1	1

Table 2.1: Haplotype combinations in the presence of 2 SNPs: H^1_{ji} and H^2_{ji} denote the alleles at the i^{th} haplotype ($i = 1, 2$) of the j^{th} individual ($j = 1, ..., n$) at the two SNPs in S.

	hap2_1	**hap2_2**	**hap2_3**	**hap2_4**
β_1	0 0	0 1	1 0	1 1
β_2	0 1	0 0	0 0	0 0
	1 0	1 0	0 1	0 1
	1 1	1 1	1 1	1 0

Table 2.2: Formulation of the four haplotype models: Each model is assigned a different causal haplotype combination which is assigned a genetic effect β_1. The remaining three combinations are assigned a common effect of β_2.

We clarify these formulations by the means of an example. Let the causal haplotype combination be $(0, 0)$ and hence fit the **hap2_1** model. Then we define the function I^{kl}_j for individual j at SNPs k and l as follows:

$$I^{kl}_j = \begin{cases} 0 & \text{if} \quad (H^k_{j1}, H^l_{j1}) \neq (0,0) \quad \text{and} \quad (H^k_{j2}, H^l_{j2}) \neq (0,0) \\ 1 & \text{if} \quad (H^k_{j1}, H^l_{j1}) = (0,0) \quad \text{or} \quad (H^k_{j2}, H^l_{j2}) = (0,0) \\ 2 & \text{if} \quad (H^k_{j1}, H^l_{j1}) = (0,0) \quad \text{and} \quad (H^k_{j2}, H^l_{j2}) = (0,0) \end{cases} \tag{2.11}$$

where (H^k_{ji}, H^l_{ji}) denotes the allele combinations for the i^{th} haplotype of the j^{th} individual at SNPs k and l. Then the phenotype for individual j is expressed as:

$$y_j = \beta_1 I^{kl}_j + \beta_2(2 - I^{kl}_j) + \epsilon_j. \tag{2.12}$$

To emphasize on the causality of specific pairs of haplotype combinations which only if they occur simultaneously lead to an elevated phenotype risk we include a further group of 6 haplotype models in the model space, when $|S| = 2$. One can argue that the full haplotype model **hap2_all** can be used to model combinations of haplotypes but since it consists

of 4 terms, one for each haplotype configuaration, it might not capture certain combinations as desired such as the compound heterozygosity model. We name these models as **hap2_comb*i*** where ***i*** ranges from 1 to 6. The compound heterozygosity model is described by **hap2_comb4**. As for the rest of the genetic models, the list of the 2-way haplotype models is also shown in Appendix B.

2.1.5 Model Comparison

It is often desirable to make comparisons between different competing models and find the one which fits the data best. The marginal likelihood of a model, say M, is a measure of the probability of observing the data given that M is true. In order to calculate this quantity we integrate out the model parameters:

$$P(D|M) = \int\int P(D|\boldsymbol{\beta}, \sigma^2) P(\boldsymbol{\beta}, \sigma^2) d\boldsymbol{\beta} d\sigma^2. \tag{2.13}$$

The notable thing here is that when we make use of the conjugate priors these integrals can be performed analytically (Appendix A). This proceeds by first integrating out $\boldsymbol{\beta}$ by writing the integrand as a Normal density and then by eliminating σ^2 by comparison with an Inverse-Gamma distribution [20]. Then

$$P(D|M) = \frac{|V^\star|^{1/2}(b)^a \Gamma(a^\star)}{|V|^{1/2}(b^\star)^{a^\star}\Gamma(a)(2\pi)^{n/2}}, \tag{2.14}$$

where the updated parameters are given in Equations 2.6 to 2.9. This quantity plays a key role in the MCMC setting that is described in the following section.

2.2 The Metropolis-Hastings Algorithm

In recent years Markov Chain Monte Carlo (MCMC) methods have been increasingly utilised to simulate complex, non-standard, multivariate distributions [13]. Sampling methods provide ways of generating samples with the property that the empirical distribution of the sample approximates the posterior distribution [100]. These computationally intensive schemes can be used for parameter learning, i.e. to obtain the posterior distribution of the parameters given the observed data [75, 32, 56] or for Bayesian learning, i.e. to obtain the posterior distribution of a particular model given the data in order to perform model comparisons.

We make use of an MCMC algorithm as it is computationally inefficient to determine the posterior distribution of the complete model space given the data. It is worth mentioning here, although it becomes clearer in the next chapter, that the cardinality of the model space increases exponentially as we increase the total number of SNPs under study. We make use of the Metropolis-Hastings algorithm which is applicable since the posterior distribution of a model given the data does not have a closed form. This is straight forward to implement and requires the definition of a proposal distribution. The proposal distribution is a density used for the generation of samples. This has to be a proper density and thus should integrate to 1. A targeted generation of a proposal distribution facilitates the convergence of the Metropolis-Hastings Algorithm.

2.2.1 Formulation of the Algorithm

Let q denote the proposal density. At each run of the Metropolis-Hastings algorithm a new state is drawn from the proposal density given the current configuration. Then, by examining the acceptance ratio α, we decide whether to accept or reject the proposed state as a sample from the actual posterior distribution.

We shall denote the current state as Θ and the proposed state as Θ^*. Then the acceptance

ratio is as follows:

$$\alpha(\Theta, \Theta^*) = min\left\{\frac{P(D|\Theta^*)P(\Theta^*)q(\Theta|\Theta^*)}{P(D|\Theta)P(\Theta)q(\Theta^*|\Theta)}, 1\right\}. \tag{2.15}$$

The proposed state is accepted if the acceptance probability exceeds the value of a randomly generated uniform random variable u. The algorithm is repeated until convergence is achieved. The initial state is chosen at random and does not affect the performance of the algorithm. In order to conclude with a representative approximation from the posterior distribution, a number of the initial iterations are discarded. This is called the *burn-in*.

Recall that each state consists of two components, the set of SNPs S and the genetic model to fit M_S, hence

$$\Theta = \{S, M_S\}.$$

In 2.15, the evidence, the probability of the data given the model $P(D|\Theta)$[1], can be estimated via 2.14. Now, the prior probability $P(\Theta)$ represents the prior belief for the two model components and is defined as follows

$$\begin{aligned} P(\Theta) &= P(S, M_S) \\ &= P(|S|)P(S||S|)P(M_S||S|) \\ &= \text{Bin}(2, p)\binom{L}{|S|}^{-1} P(M_S||S|), \end{aligned} \tag{2.16}$$

- $|S|$ denotes the cardinality of S, which is binomially distributed with $n = 2$ and probability of success p which is user defined,
- L is the total number of SNPs under examination.

The prior probability of the model is calculated as the probability of $|S|$ multiplied by the probability of choosing some specific SNPs according to $|S|$ and finally multiplied by the

[1]N.B.: Note that in the Metropolis-Hastings setting we replace M by Θ. It describes the model details which are fully characterised by S and M_S

$\|S\|$	0	1	2
$P(M_S \| \|S\|)$	1	$\frac{1}{2}$	$\frac{1}{13}$

Table 2.3: Probability distribution of the genetic models given $|S|$: Under a specific $|S|$ the number of possible genetic models differs and so do the probabilities since each genetic model under each $|S|$ has an equal probability of being obtained.

probability of which genetic model to fit, which also depends on $|S|$. The probability of choosing specific SNPs along with the probability of the genetic model are modelled as uniform allowing equal probabilities for each SNP combination and genetic model. This follows the idea of a prior imposed in [3]. Table 2.3 shows the probability distribution of the genetic models given $|S|$. Under a specific $|S|$ the number of possible genetic models differs and so do their probabilities since each genetic model under each $|S|$ has an equal probability of being obtained. We can see that the choice of model depends entirely on how many SNPs are to be included in S and each model has equal probability to be selected given the cardinality of S.

2.2.2 The Proposal Density, q

The Metropolis-Hastings algorithm is specified by its candidate generating density $q(\Theta, \Theta^*)$. One appealing property the proposal density should possess is the ease of drawing values from it. Generally, it is selected from a family of distributions that requires to specify parameters such as location and scale. The properties a good proposal distribution should possess include the ease to sample candidate states from as well as the ease to compute the acceptance ratio. Each jump should go a reasonable distance in the model space otherwise the algorithm would be slow and, finally, jumps should not rejected too frequently [29]. Several examples include the proposal density which defines a random walk chain, where the candidate value equals the current value with some noise added. Another example is the form of a proposal density which describes an independence chain (the candidate state is independent of the current state) [13]. Natural choices in the case of a continuous MCMC

setting are the multivariate Normal and the multivariate t distributions. In our case we need to build a proposal density which can serve the purpose of a discrete Markov chain and can draw candidates efficiently and speedy.

In order to move around the space of possible states we introduce the notion of an action. There are four possible actions which describe the possible alterations of a state described by S and M_S. They are as follows:

1. Keep SNP(s) in S, change model M_S,
2. Change a SNP in S , keep model M_S,
3. Add a SNP to S, change model M_S,
4. Delete a SNP in S, change model M_S.

There are two special cases that we need to consider at this point. If at the current state of the chain S is empty, there is only one action to be performed and this is clearly action 3. On the other hand if S has reached its full cardinality then action 3 is prohibited.

Under a multinomial distribution setting, the actions are assigned different success probabilities which sum up to 1. Then, a single random draw from this distribution will determine the action to be performed. Once this first step is performed the second one includes the update of the first component of the candidate state Θ^*. For action 1, S remains unchanged while for action 2 a SNP is chosen randomly to be replaced. Then a new SNP is chosen at random from the set of examined SNPs. Each SNP is assigned the same probability of being included in S. In the case of adding a SNP, the same setting occurs and each SNP is equally probable to be added to S. Finally, in deleting a SNP, a SNP in S is chosen at random to be dropped.

Once S is finalised, a genetic model needs to be chosen. Table 2.3 gives the probability distribution ofM_S given $|S|$. In most cases we choose a model uniformly from the collection of applicable models. There are situations, though, where this is not the case and we introduce the notion of the *closest model*.

The idea behind this notion is best explained by some examples. When we move from $|S| = 1$ to $|S| = 2$ the common genetic models are the additive and the general models in the respective dimension. Hence, we say that the closest model of the additive with a single SNP is the model where 2 SNPs act additively on the phenotype. The same occurs when we consider the general model. Therefore these closest models will have an elevated probability of being proposed.

The probability of a new state given the current one, $q(\Theta^*|\Theta)$, is simply the product of probability of choosing a specific action, the probability of changing/adding/deleting a SNP and the probability of changing a model (according to the action chosen). After the essential performance analysis, we are convinced that the proposal density is stable and fast. The fact that the multinomial probabilities assigned to the different actions can be adjusted provides the user the ability to give different weights to different actions.

2.3 Assessment Measures

Once the algorithm is executed for a sufficient number of iterations, we can extract different information to assess significance. The term "sufficient" is quite vague at this point. It surely depends on how big or small the entire model space is, something which is also examined this in detail in Chapter 3. We then focus on the following posterior quantities:

- $P(S = \{S_1, S_2\}|D)$, the posterior probability of detecting a particular pair of SNPs ;
- $P(S_i|D)$, $i = 1, ..., L$ the marginal posterior probability of each SNP;
- $P(M_S|D)$, the posterior probability of the genetic model.

These distributions are easily calculated and can contribute in assessing the causal associations. The posterior probability of each SNP given the data is obtained by counting how many times the algorithm “visits” it while the posterior distribution for S and M_S follow accordingly. We simply count the number of times each pair (if applicable) and genetic

model is chosen at every iteration. We can also extract the posterior probability distribution of $|S|$ which can give us a view of how many SNPs are most likely to explain variation.

2.4 Summary

Throughout this chapter we have presented the methodology of a new, formable and tractable model for fine mapping using haplotype data. With the use of an MCMC method and more specifically a Metropolis-Hastings algorithm and a local proposal density, it searches for the optimal SNP or pairs of SNPs which in combination with the appropriate genetic model are most strongly associated with phenotypic variation. The consideration of haplotypes rather than the direct analysis of genotypes provides additional underlying biological information and allows for effects which are not effortlessly visible to be uncovered, for example the compound heterozygosity effects. During the next chapter we assess the performance of our fine mapping model and compare it to existing methods.

Chapter 3

Simulation Study

Over the course of Chapter 2, we presented a flexible model for detecting associations between a disease and multiple genetic variants under the Bayesian linear model framework. Our goal in this chapter is to examine the performance and behaviour of this method using data simulated under the compound heterozygosity model. The data have different characteristics in order to assess the model under different circumstances. We examine the power of the approach to detect the true underlying relationships between the phenotypes and the haplotypes and compare its performance to existing methods. Additionally, we test the convergence of the Metropolis-Hastings via the Kullback-Leibler divergence measure. We begin by giving a description of the construction of the haplotype data using **HAPGEN**.

3.1 Generation of Simulated Data

3.1.1 HAPGEN

Haplotype datasets are simulated using the software **HAPGEN** [89]. The software constructs case-control datasets at SNP markers using known haplotypes from the HapMap and 1000 Genomes Projects. A control individual is constructed by separately consider-

ing its haplotypes in a given region. Each haplotype is made up as mosaics of the known HapMap haplotypes. The probability of breaks in the mosaic pattern is calculated by using fine-scale estimates of recombination. A case individual is simulated by first choosing a SNP to be causal. The genotype of the case at this SNP is simulated first and then by working outwards in each direction, the haplotypes carrying the alleles simulated at the causal SNP are constructed [82].

We simulate two batches of 1,000 datasets consisting of 1,000 individuals and 100 SNPs using as reference the European ancestry data of the HapMap Project, release 22, build 36 on chromosomes 7 and 15 respectively. The same SNPs were used for each batch of datasets, rs7794205 through to rs6950352 on chromosome 7 and rs11854593 through to rs12442995 on chromosome 15. We set the relative risk to 1 so that there is no genetic effect in the cases resulting in different haplotypes sets where there is no distinction between the cases and the controls. Subsequently, since the phenotypes we wish to examine are quantitative, we generate them in **R**.

3.1.2 Generation of Phenotypes

In order to assess the performance of the fine mapping model under compound heterozygosity we simulate different sets of phenotypes under this model which differ in the size of the genetic effect and the structure of the linkage disequilibrium (LD) both between the causal and the flanking SNPs. In detail, when an individual has the causal haplotype combination which leads to compound heterozygosity the effect size is set equal 0.30 or 0.20. When any other haplotype combination is present the effect size is zero. Let (H_{ji}^k, H_{ji}^l) be the allele combinations for the i^{th} haplotype of the j^{th} individual at SNPs k and l then the function I_j^{kl} is defined as follows:

$$I_j^{kl} = \begin{cases} 1 & \text{if} \quad (H_{j1}^k, H_{j1}^l) = (0,1) \quad \text{and} \quad (H_{j2}^k, H_{j2}^l) = (1,0) \\ 0 & \text{otherwise} \end{cases}$$

The phenotype for individual j is expressed as:

$$y_j = \beta_1 I_j^{kl} + \beta_2(1 - I_j^{kl}) + \epsilon_j \tag{3.1}$$

where ϵ_j $N(0, 1)$ while $\beta_1 = 0.30, 0.20$ and $\beta_2 = 0$.

Further, a set of null phenotypes, with no genetic effect size, is simulated in order to calculate false positive rates. We also vary the levels of linkage disequilibrium (LD) both between the causal variants and between the causal variants and the flanking sites. We divide the LD measure r^2 in three different bins which we label as low, medium and high (Table 3.1.2). We, then, create nine different cases where we consider all the different combinations of correlation. For each combination we choose a pair of SNPs to act as causal if it satisfies both conditions of the correlation structure. For example, when we refer to "Low causal/Low flanking" scenario, r^2 between the two causal SNPs is less than 0.4 and so is the correlation with the flanking SNPs If the scenario is "High causal/Medium flanking" then the correlation between the two causal sites is greater than 0.7, while the LD bewteen the causal and the flanking sites is between 0.4 and 0.7. For the three “Low flanking” scenarios we use the simulated data under chromosome 7 while for the remaining scenarios we use the simulated data under chromosome 15.

Linkage Disequilibrium	
Label	Levels
Low	$0 \leq r^2 < 0.40$
Medium	$0.40 \leq r^2 < 0.70$
High	$0.70 \leq r^2 \leq 1$

Table 3.1: LD Intervals used in the analysis.

As an attempt to avoid potentially problematic pairs of causal SNPs, for every 100^{th} dataset we set a different causal pair which also satisfies the LD constraints. We also set the prior probability on $|S|$ to be 0.95 and the probability of the closest model to 0.30. Each Metropolis-Hastings chain was performed for a total of $40,000$ iterations, with $20,000$ of those being the burn-in. The specification of the prior distributions is as follows:

- $\beta \sim N(\mathbf{0}, \boldsymbol{V_\beta}\sigma^2)$, where $\boldsymbol{V_\beta} = \left(\begin{smallmatrix} 0.02 & 0 \\ 0 & 0.02 \end{smallmatrix}\right)$;
- $\sigma^2 \sim IG(a, b)$ where $a = 3$ and $b = 2$.

The reason for this choice over the prior values for the parameters follows from the assumption that the genetic effects will have similar sizes with previously reported results of binary traits. We compare the non-centrality parameters under quantitative and binary traits. An F test statistic is used to compare the null hypothesis of no effect with the alternative hypothesis of a genetic effect. The distribution under the null hypothesis is a "central" F distribution while under the alternative it is "non-central". The non-centrality parameter is obtained to measure the shift of the F distributions under the two hypotheses.

- **Binary Traits**: $Np(1-p)\gamma^2$;
- **Quantitative Traits**: $Np(1-p)\dfrac{2\gamma^2}{\sigma^2}$,

where the total sample size is $2N$, γ denotes the genetic effect while σ^2 denotes the phenotypic variance and p the minor allele frequency. The values of a and b are chosen so that the total phenotypic variance lies well within the range of $IG(a, b)$. We choose a to be 3 and b to be 2 so that the mean and variance of $IG(3, 2)$ are both equal to 1 [50].

We proceed with the analysis of the data using three existing methods, **SNPTEST**, **BAMSE** and **BIMBAM**, in addition to our own bayesian fine mapping model. The assessment measure which can be extracted from each method is the marginal posterior probability of detecting each of the causal variants. Furthermore, all methods, excluding **SNPTEST**, provide the joint posterior probability of obtaining the causal pair of SNPs. Since our method also explores different genetic models, we can also observe the posterior probability distribution of the genetic models and more specifically the posterior probability of detecting the compound heterozygosity model.

3.2 Power Comparisons

This section presents the power of each method with the means of Receiving Operating Characteristic (ROC) curves. In brief, a ROC curve is created by plotting the false positive rates against the true positive rates at various thresholds. The true positive rate is also known as the sensitivity while the false positive rate is one minus the specificity. A ROC curve shows the trade off between sensitivity and specificity where an increase in sensitivity is accompanied by a decrease in specificity, ultimately providing the tool to select optimal models. The true and false positive rates are obtained by applying each method to the datasets simulated under the alternative and the null model of no association respectively. The marginal posterior probability of each SNP and the posterior probability of a pair of SNPs, if applicable, are extracted from each approach. These statistics are directly available for **BAMSE** and **FineMap MCMC** while for **SNPTEST** and **BIMBAM** we convert the resulting Bayes factors to posterior probabilities using the relationship presented in [85]. This links the posterior odds (PO) to the Bayes factor and to the posterior probability (PP) as follows:

$$PO = BF \times \frac{\pi}{1-\pi};$$
$$PP = \frac{PO}{1+PO},$$

where we set $\pi = 10^{-5}$, as it is also suggested in the WTCCC study [97], assuming 10^6 independent regions of the genome and an expectation of 10 detectable genes involved in the condition.

ROC curves are used to assess the different methods based on the three assessment measures: the marginal posterior probability to identify the susceptibility SNPs, the posterior probability to detect the causal pair of variants, along with the posterior probability to detect pairs in high LD with the causal pair ($r^2 \geq 0.80$ and $r^2 \geq 0.90$) and the posterior probability of detecting the compound heterozygosity model (only available from our fine

mapping model).

We start off by examining the results for the detection of the causal variants for the different levels of LD between the causal SNPs assuming low LD with the flanking sites (Figure 3.1). The left hand side column of ROC curves demonstrates the power of each method to detect the first component of S while the right hand side displays the outcome concerning the second component of S. It should be noted that the false positive rate on the x axis varies from 0 to 0.10 and the true positive rate on the y axis stretches from 0 to 1. The colours which represent each method are as follows: **SNPTEST-Additive** black, **SNPTEST-General** red, **BIMBAM** green, **BAMSE** blue, **FineMap MCMC** cyan. When the correlation between the causal SNPs is low, there is a clear distinction on the performance of each method. The additive model of **SNPTEST** performs the worst while **FineMap MCMC** the best. **BIMBAM** achieves higher power than the additive model but does not outperform **SNPTEST**'s general model, while **BAMSE** scores second best. When the between-variants LD increases ($0.40 \leq r^2 < 0.70$) the power of all methods increases although the additive model does not exhibit large evidence of association. In this instance, **BIMBAM** shows a boost in power and surpasses the general model of **SNPTEST** but does not do as well as **BAMSE**. Once again, the **FineMap MCMC** model exhibits the strongest results although for false positive rates closer to zero, **BAMSE** does better in identifying the second causal SNP. The final row of plots in Figure 3.1 concerns the high LD between the causal SNPs. There is a further increase in power where all methods, except the additive model, perform approximately the same as the false positive rate increases to 0.10. **SNPTEST**'s general model exhibits the largest power with **BAMSE** and **FineMap MCMC** scoring second for smaller false positive rates.

An additional posterior quantity of key importance is the posterior probability to detect the pair of SNPs associated with the phenotype. This information can be extracted from the Metropolis-Hastings algorithm of our fine mapping model as effectively as the marginal posterior probability for each SNP. It is also possible to obtain this information for **BAMSE**

and **BIMBAM**. The latter calculates the posterior probability for every possible pair of the examinable SNPs while **BAMSE** calculates the posterior probability for a given number of the risk sets which are above a certain threshold. We set the number of risk sets to 10 and the threshold at 0.0001. The results concerning the "Low causal/Low flanking" LD scenario are presented in Figure 3.2. **BAMSE** does not manage to exhibit power beyond roughly 0.10 while **BIMBAM** is outperformed by **FineMap MCMC**. Since the flanking sites' LD is characterized as low we only see some boost in power when we take them into account. It is unavoidable to have some highly correlated variants in the neighbouring sites but we do not see significant change when r^2 increases from 0.8 to 0.9. On the other hand, **BIMBAM** surpasses the other two approaches as the correlation between the causal sites increases (Figure 3.3). When we add the posterior probabilities of SNPs in high LD we see an increase in the power of the **FineMap MCMC** model while **BIMBAM**'s performance drops. The two methods perform approximately the same as the false positive rate increases. Analogous results are observed for the "High caus/Low flanking" scenario as we can see in Figure 3.4. The main difference though, is that although the false positives increase for **BIMBAM**, it still surpasses **FineMap MCMC** when we take into account flanking pairs.

We now move on to a slightly more complex LD structure. The causal pairs are chosen such that the correlation with the neighbouring sites lies between 0.40 and 0.70. It, now, becomes more difficult to detect the true associations since highly correlated SNPs are "visited" at a higher frequency by the MCMC chain. The power of detecting one of the two causal variants under the medium flanking LD scenario is presented in Figure 3.5. A collective comment concerning the three different levels of LD, is that **FineMap MCMC** outperforms the rest of the methods and as the LD between the causal variants increases we observe an analogous behaviour for the power. The greatest dissimilarity in the ROC curves occurs when r^2 is less than 0.40 while the power of **SNPTEST**'s additive model remains low across all values of r^2. While **BAMSE** is constantly the second best method, **BIMBAM** and **SNPTEST**'s general model alternate for the third place across the three

sets of plots. When between causal SNPs correlation escalates to beyond 0.70 **BAMSE**, **BIMBAM** and **SNPTEST**'s general model perform approximately equal in detecting the second component of S, while **BAMSE** surpasses the other two methods in detecting the first component.

Although **BAMSE** performs remarkably well in detecting marginal effects, it fails to exhibit comparable power in detecting the joint effect as we can see in Figure 3.6. This phenomenon is observed in all the results. **FineMap MCMC** does marginally better when we take in to account only the posterior information on actual causal pair but once we consider the highly correlated pairs adjacent the power increases substantially. **BIMBAM** fails to exhibit convincing true positive rates even if we consider flanking tag pairs of SNPs. On the other hand, when the LD between the causal SNPs increases (Figure 3.7) **BIMBAM** outperforms **FineMap MCMC** and **BAMSE**, although accounting for highly correlated flanking SNPs changes the conclusions. Although the two methods' power gets closer and closer to each other as the false positive rate increases, at very small values of the false positive rate the true positive rate of **FineMap MCMC** is twice as **BIMBAM**'s. The conclusions do not alter when we move to Figure 3.8 only in this case, the true positive rates reach higher values closer to 1.

The final scenario we study is that of high flanking LD, where r^2 is greater than 0.70. We expect the detection of the true associations to become more difficult since flanking SNPs can easily be perfect copies of the causal SNPs in which case the MCMC chain is not able to distinguish but still produce a combination of S and M_S that is associated with the phenotype. The detection of marginal effects suffers from this phenomenon when the LD between the causal SNPs is classified as medium or high. At low LD between causal SNPs, we once again see better performance by **FineMap MCMC** (Figure 3.9). **BAMSE** appears to perform better than **FineMap MCMC** when we explore the marginal effects of SNPs in medium LD between them, while later it is joined by **BIMBAM** and **SNPTEST**'s general model in the last pair of plots of Figure 3.9. **SNPTEST**' s additive model fails to show any

increased power than in any of the previously examined cases.

Figures 3.10, 3.11 and 3.12 have amongst others one thing in common, **BIMBAM** outperforms **FineMap MCMC** when accounting for the posterior probability of detecting the actual causal pair of SNPs. As it has previously been observed, the power at small false positive rates increases with the correlation between the causal variants. Once we incorporate the information of pairs with r^2 greater than 0.90, **BIMBAM** still performs better at medium and high causal scenarios. When, eventually, we account for the paris with r^2 greater than 0.80 we achieve highest power of **FineMap MCMC** across all three levels of between causal SNPs dependence.

Last but not least we examine the power to detect the compound heterozygosity model. This is directly accessible from the Metropolis-Hastings algorithm of the **FineMap MCMC** model as the proportion of "visits" the chain pays to the genetic model of interest. The resulting ROC curves are displayed in Figure 3.13. Each pair of axes includes three curves, one for each level of LD characterizing the correlation between the causal SNPs. We observe a trend across the three levels of flanking sites LD where the low causal LD scenario produces the highest power to detect the **hap2comb_4** model. Then follows, with a slight decrease, the power to detect the true model when the between causal SNPs r^2 is between 0.40 and 0.70. Furthermore, when the flanking sites LD is low or medium, the power drops substantially for between causal SNPs $r^2 > 0.70$. These observations are somewhat expected since as the LD between the components of S increases then the compound heterozygote effects become less frequent.

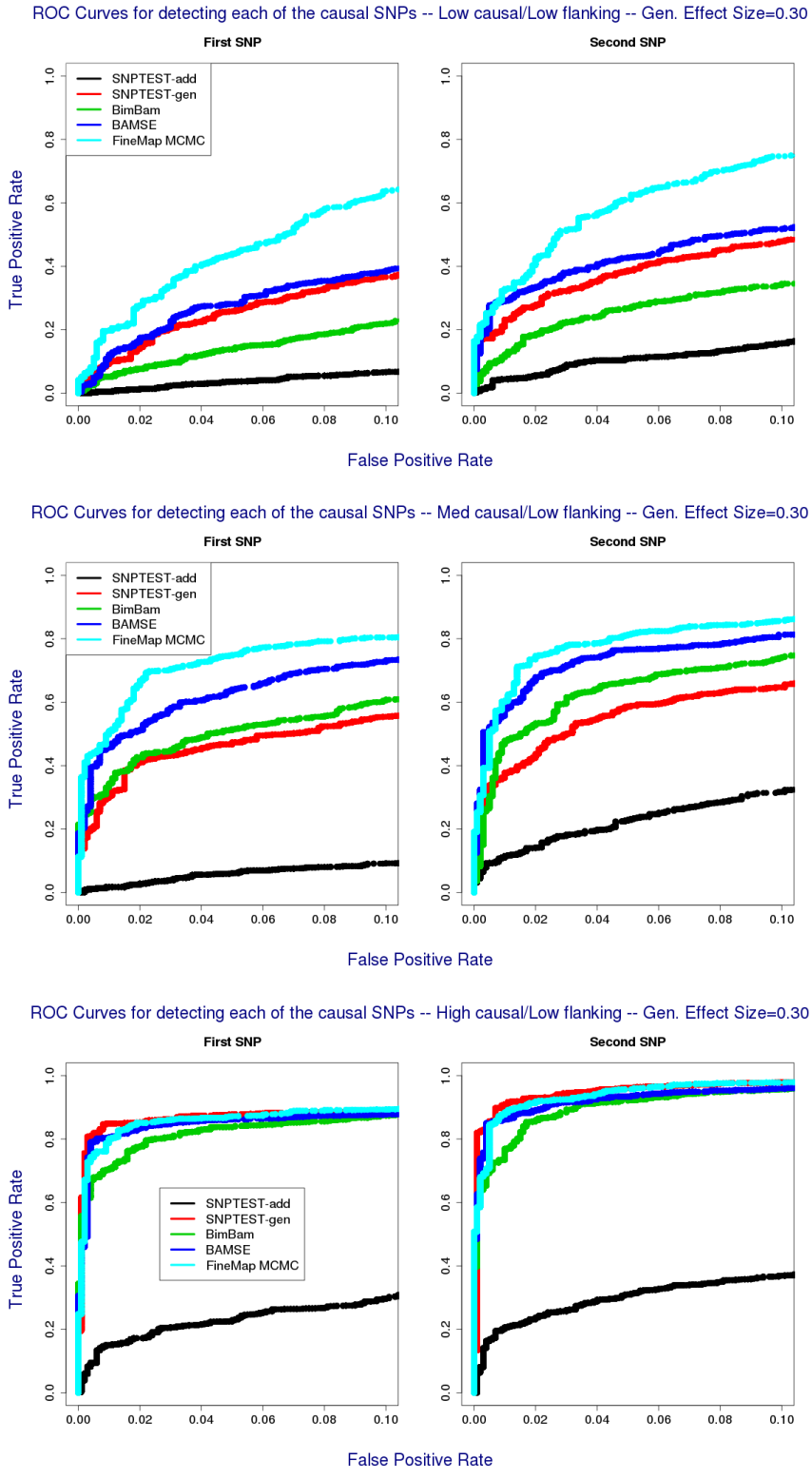

Figure 3.1: ROC curves to identify the marginal effects of the true causal variants for the different method examined. The horizontal axes represent the false positive rates while the vertical axes represent the true positive rates. The different methods are coded using the following colours: **SNPTEST-Additive** (black), **SNPTEST-General** (red), **BIMBAM** (green), **BAMSE** (blue), **FineMap MCMC** (cyan) (LD: low, med, high causal/low flanking).

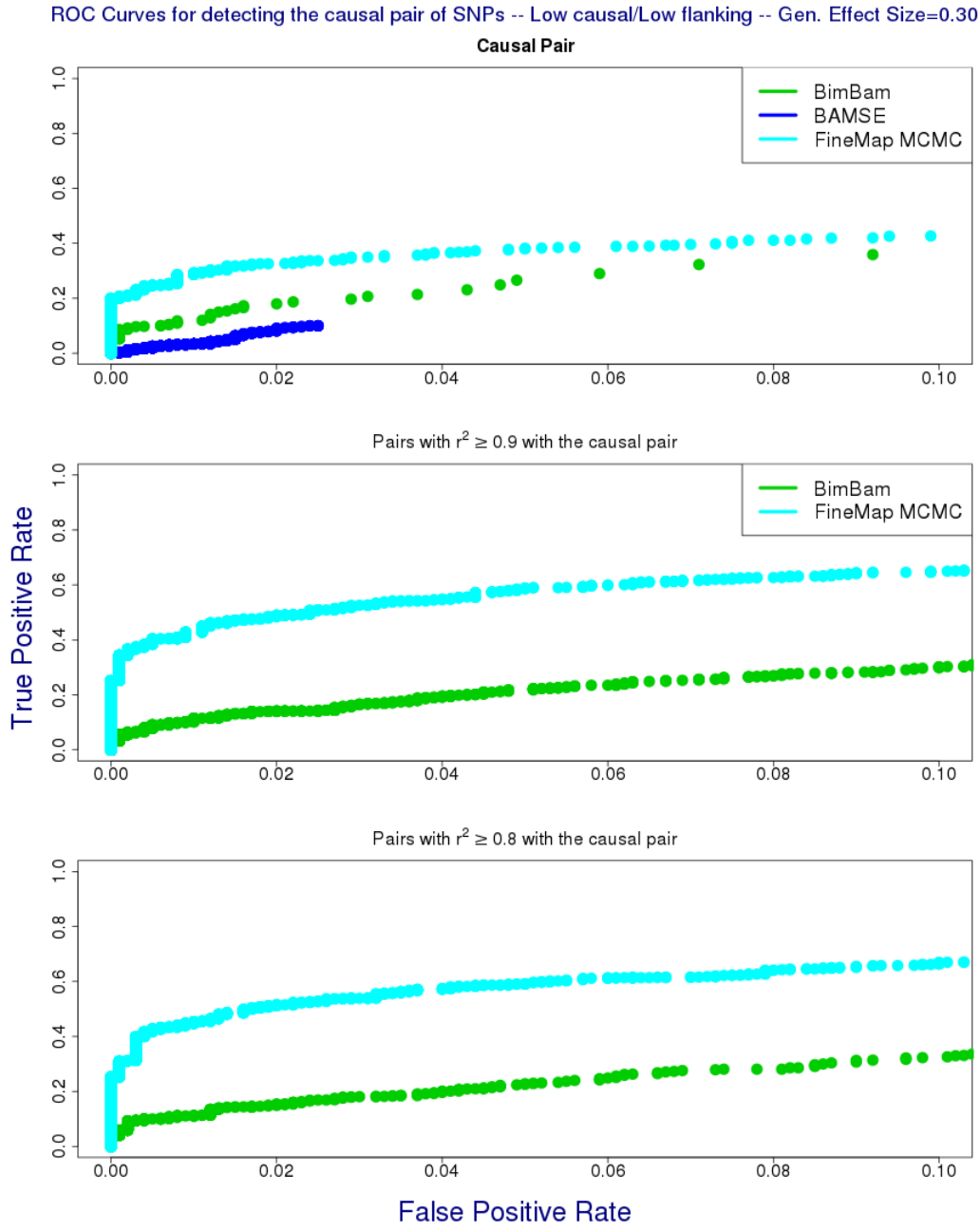

Figure 3.2: ROC curves to detect the causal pair of variants (**BAMSE**, **BIMBAM** and **FineMap MCMC**) and to detect the pairs in high LD ($r^2 = 0.80$ and 0.90) with the causal combination of SNPs (**BIMBAM** and **FineMap MCMC**). The horizontal axes represent the false positive rates while the vertical axes represent the true positive rates. The different methods are coded using the following colours: **BIMBAM** (green), **BAMSE** (blue), **FineMap MCMC** (cyan) (LD: low causal/low flanking).

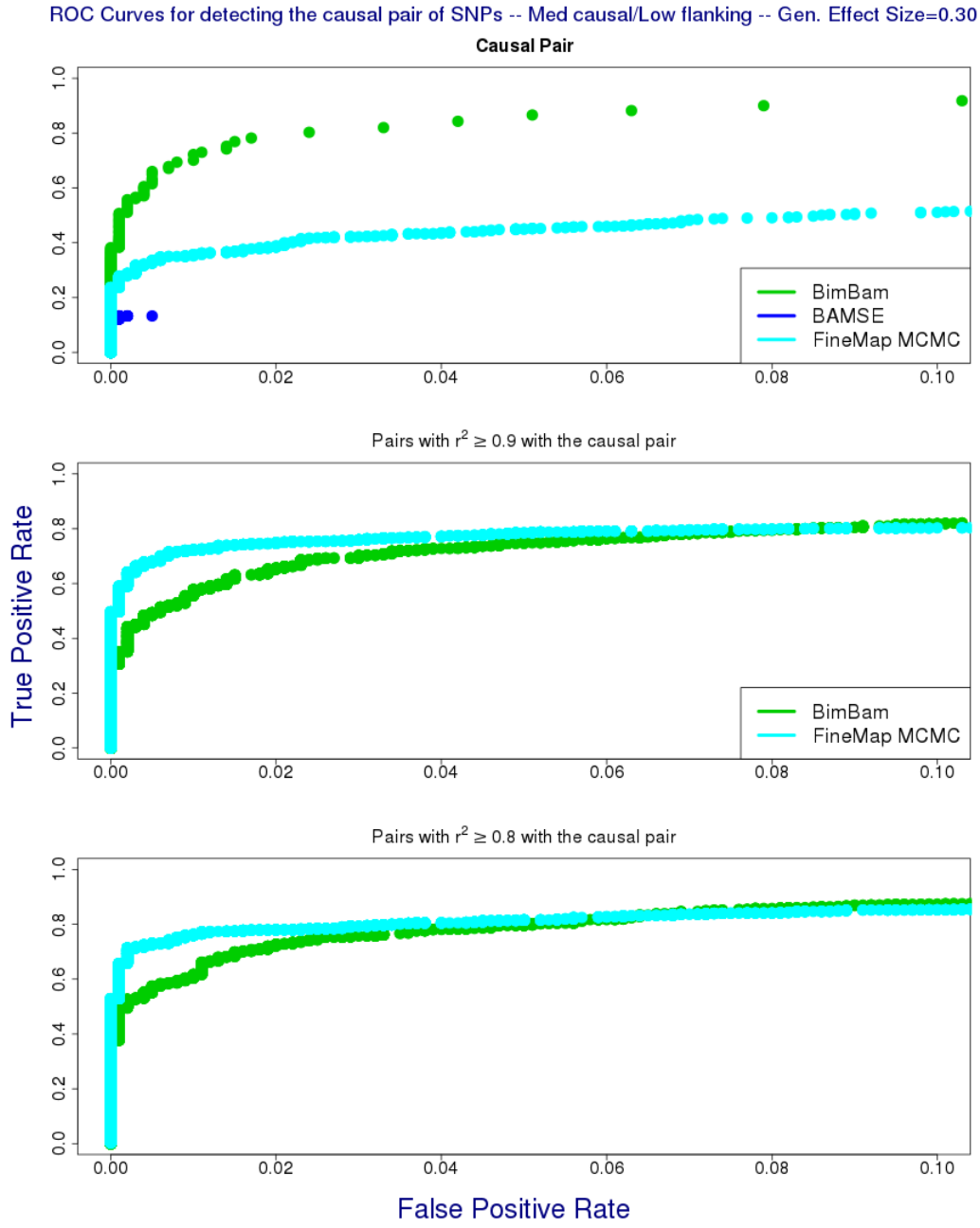

Figure 3.3: ROC curves to detect the causal pair of variants (**BAMSE**, **BIMBAM** and **FineMap MCMC**) and to detect the pairs in high LD ($r^2 = 0.80$ and 0.90) with the causal combination of SNPs (**BIMBAM** and **FineMap MCMC**). The horizontal axes represent the false positive rates while the vertical axes represent the true positive rates. The different methods are coded using the following colours: **BIMBAM** (green), **BAMSE** (blue), **FineMap MCMC** (cyan) (LD: med causal/low flanking).

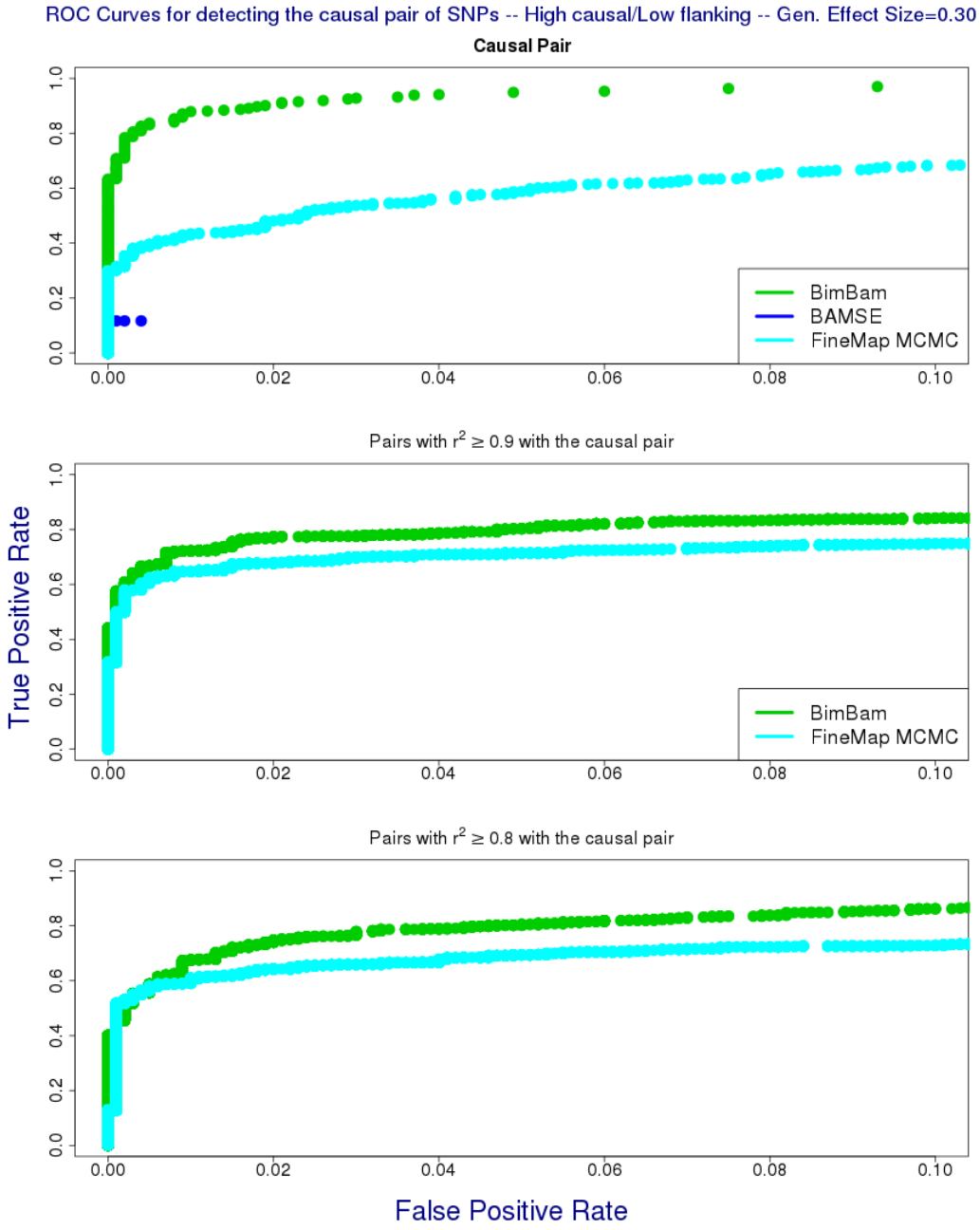

Figure 3.4: ROC curves to detect the causal pair of variants (**BAMSE**, **BIMBAM** and **FineMap MCMC**) and to detect the pairs in high LD ($r^2 = 0.80$ and 0.90) with the causal combination of SNPs (**BIMBAM** and **FineMap MCMC**). The horizontal axes represent the false positive rates while the vertical axes represent the true positive rates. The different methods are coded using the following colours: **BIMBAM** (green), **BAMSE** (blue), **FineMap MCMC** (cyan) (LD: high causal/low flanking).

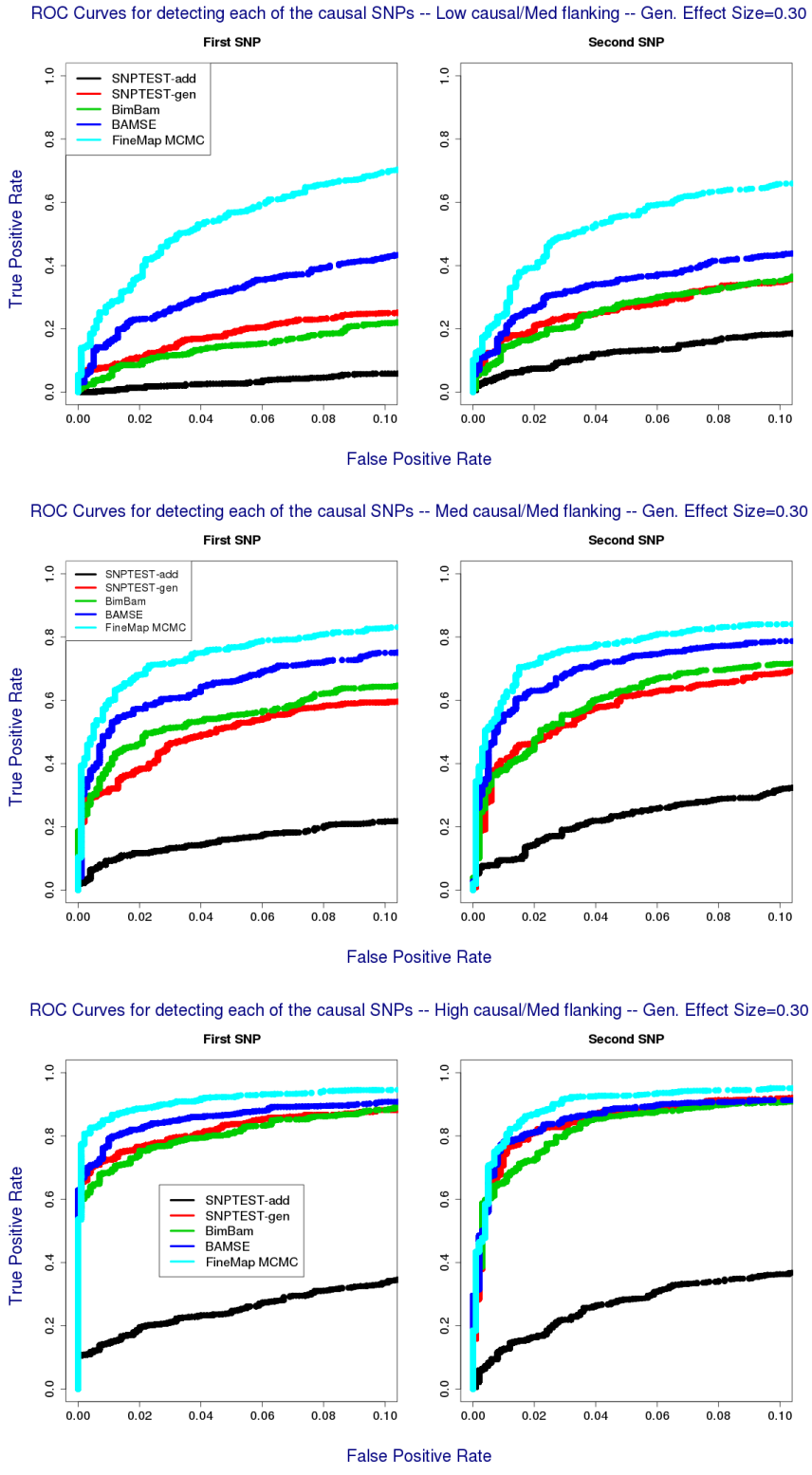

Figure 3.5: ROC curves to identify the marginal effects of the true causal variants for the different method examined. The horizontal axes represent the false positive rates while the vertical axes represent the true positive rates. The different methods are coded using the following colours: **SNPTEST-Additive** (black), **SNPTEST-General** (red), **BIMBAM** (green), **BAMSE** (blue), **FineMap MCMC** (cyan) (LD: low, med, high causal/med flanking).

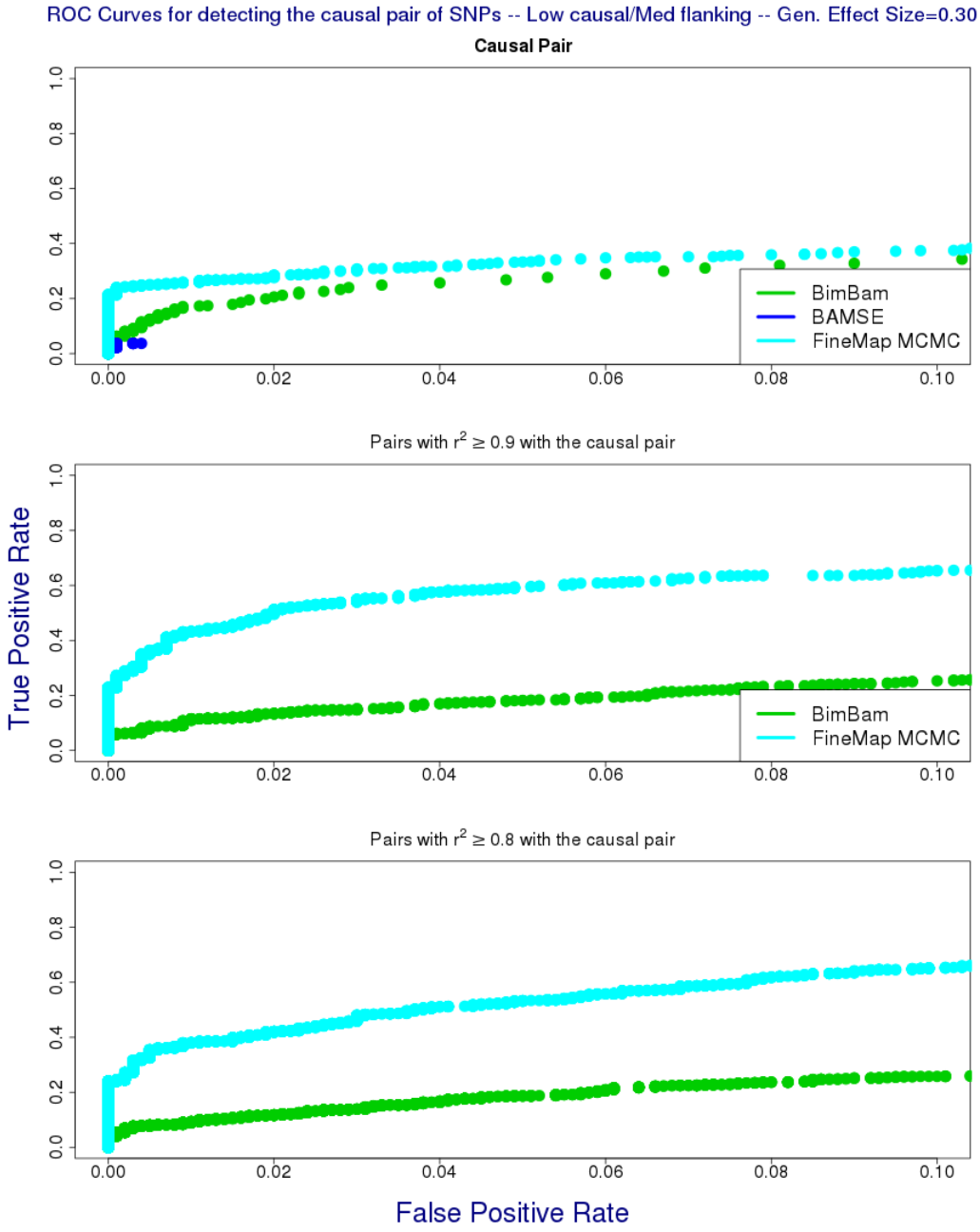

Figure 3.6: ROC curves to detect the causal pair of variants (**BAMSE**, **BIMBAM** and **FineMap MCMC**) and to detect the pairs in high LD ($r^2 = 0.80$ and 0.90) with the causal combination of SNPs (**BIMBAM** and **FineMap MCMC**). The horizontal axes represent the false positive rates while the vertical axes represent the true positive rates. The different methods are coded using the following colours: **BIMBAM** (green), **BAMSE** (blue), **FineMap MCMC** (cyan) (LD: low causal/med flanking).

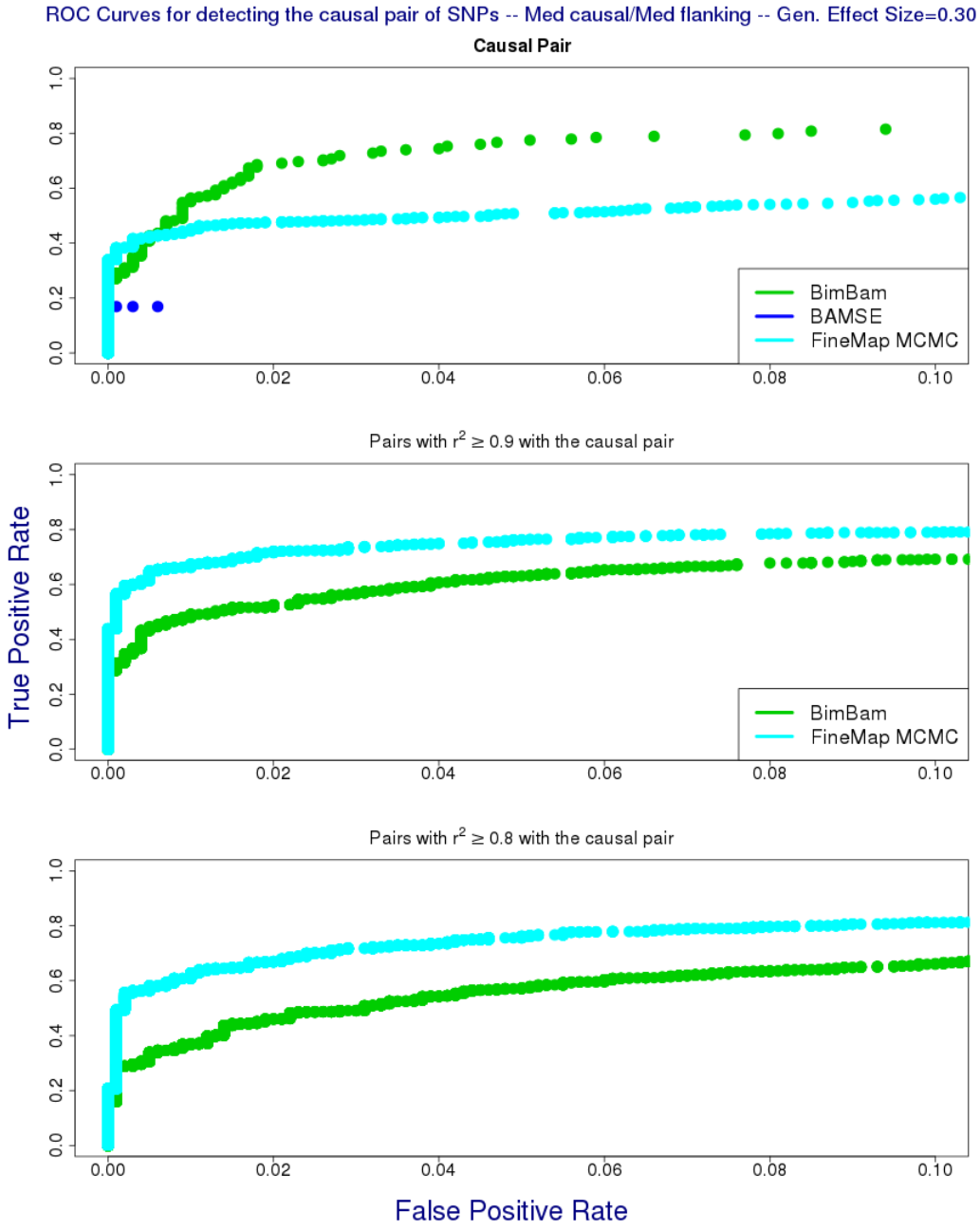

Figure 3.7: ROC curves to detect the causal pair of variants (**BAMSE**, **BIMBAM** and **FineMap MCMC**) and to detect the pairs in high LD ($r^2 = 0.80$ and 0.90) with the causal combination of SNPs (**BIMBAM** and **FineMap MCMC**). The horizontal axes represent the false positive rates while the vertical axes represent the true positive rates. The different methods are coded using the following colours: **BIMBAM** (green), **BAMSE** (blue), **FineMap MCMC** (cyan) (LD: med causal/med flanking).

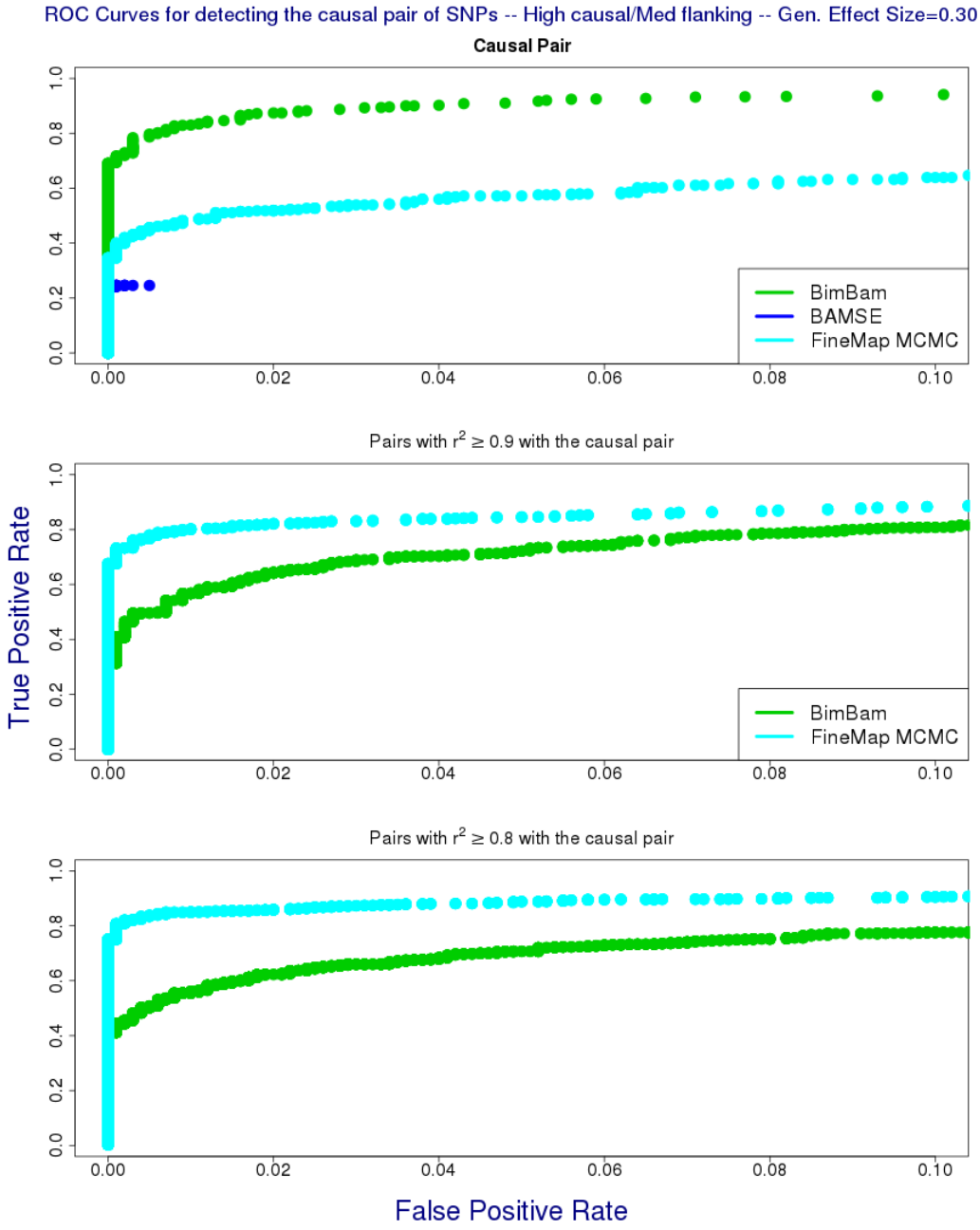

Figure 3.8: ROC curves to detect the causal pair of variants (**BAMSE**, **BIMBAM** and **FineMap MCMC**) and to detect the pairs in high LD ($r^2 = 0.80$ and 0.90) with the causal combination of SNPs (**BIMBAM** and **FineMap MCMC**). The horizontal axes represent the false positive rates while the vertical axes represent the true positive rates. The different methods are coded using the following colours: **BIMBAM** (green), **BAMSE** (blue), **FineMap MCMC** (cyan) (LD: high causal/med flanking).

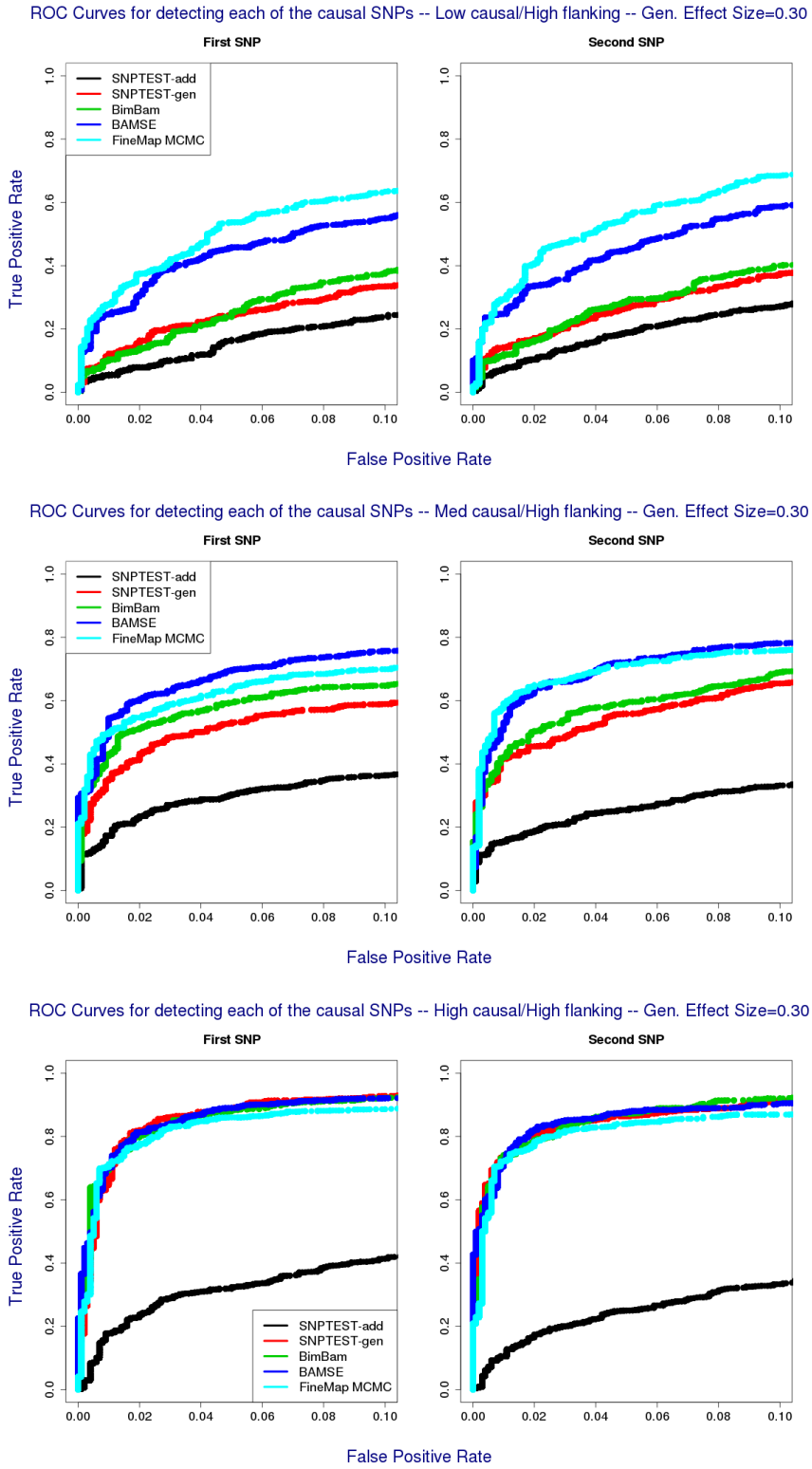

Figure 3.9: ROC curves to identify the marginal effects of the true causal variants for the different method examined. The horizontal axes represent the false positive rates while the vertical axes represent the true positive rates. The different methods are coded using the following colours: **SNPTEST-Additive** (black), **SNPTEST-General** (red), **BIMBAM** (green), **BAMSE** (blue), **FineMap MCMC** (cyan) (LD: low, med, high causal/high flanking).

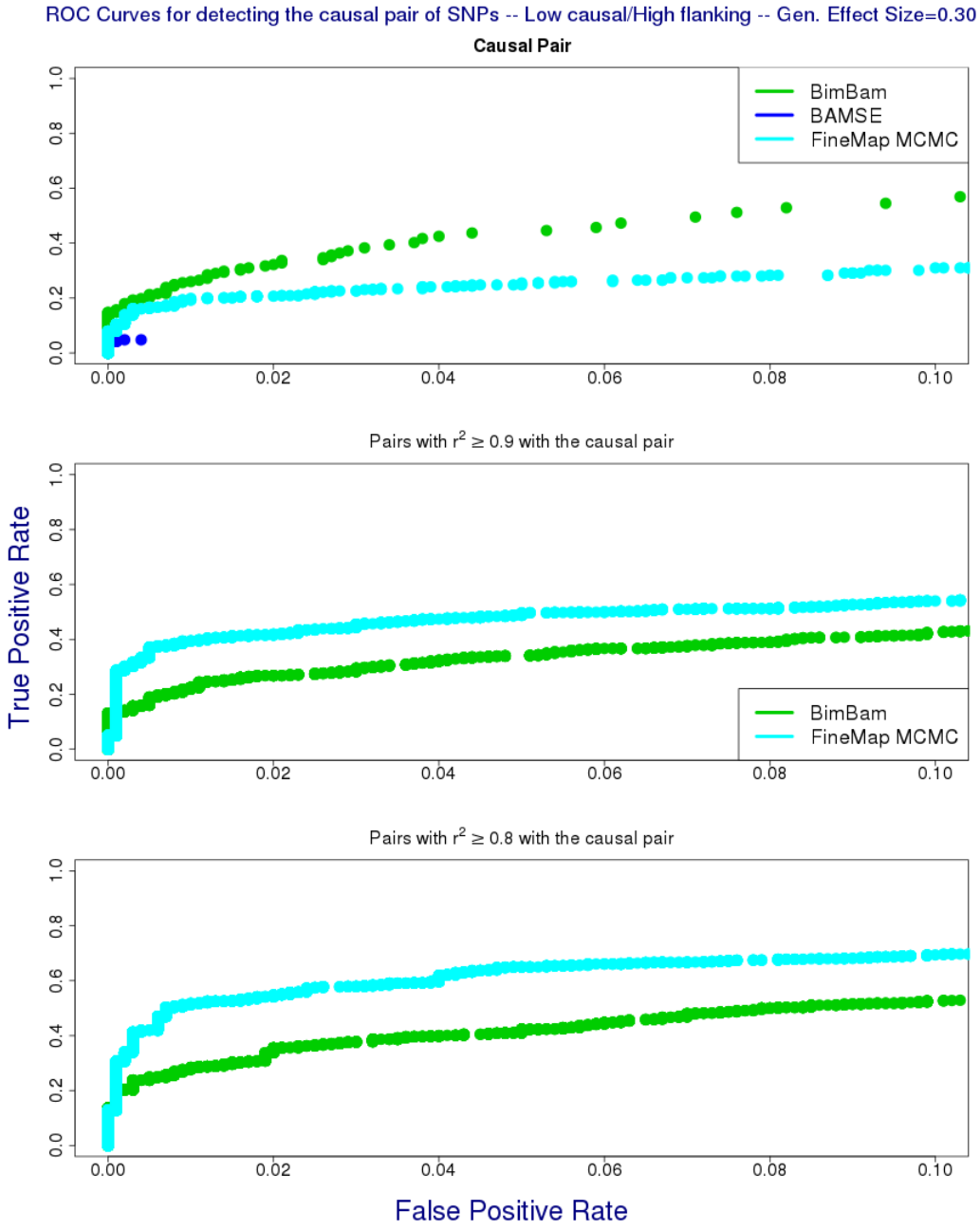

Figure 3.10: ROC curves to detect the causal pair of variants (**BAMSE**, **BIMBAM** and **FineMap MCMC**) and to detect the pairs in high LD ($r^2 = 0.80$ and 0.90) with the causal combination of SNPs (**BIMBAM** and **FineMap MCMC**). The horizontal axes represent the false positive rates while the vertical axes represent the true positive rates. The different methods are coded using the following colours: **BIMBAM** (green), **BAMSE** (blue), **FineMap MCMC** (cyan) (LD: low causal/high flanking).

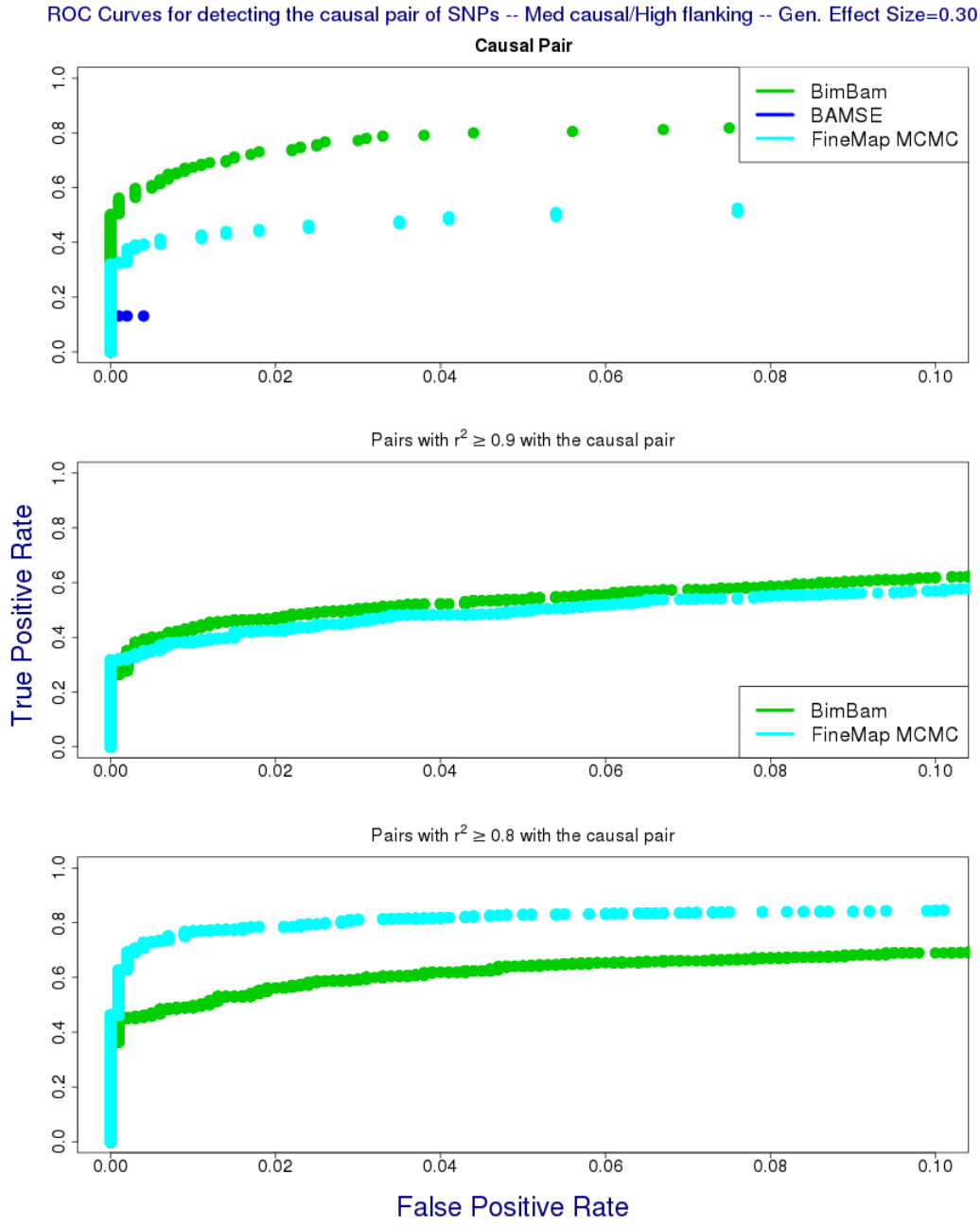

Figure 3.11: ROC curves to detect the causal pair of variants (**BAMSE**, **BIMBAM** and **FineMap MCMC**) and to detect the pairs in high LD ($r^2 = 0.80$ and 0.90) with the causal combination of SNPs (**BIMBAM** and **FineMap MCMC**). The horizontal axes represent the false positive rates while the vertical axes represent the true positive rates. The different methods are coded using the following colours: **BIMBAM** (green), **BAMSE** (blue), **FineMap MCMC** (cyan) (LD: med causal/high flanking).

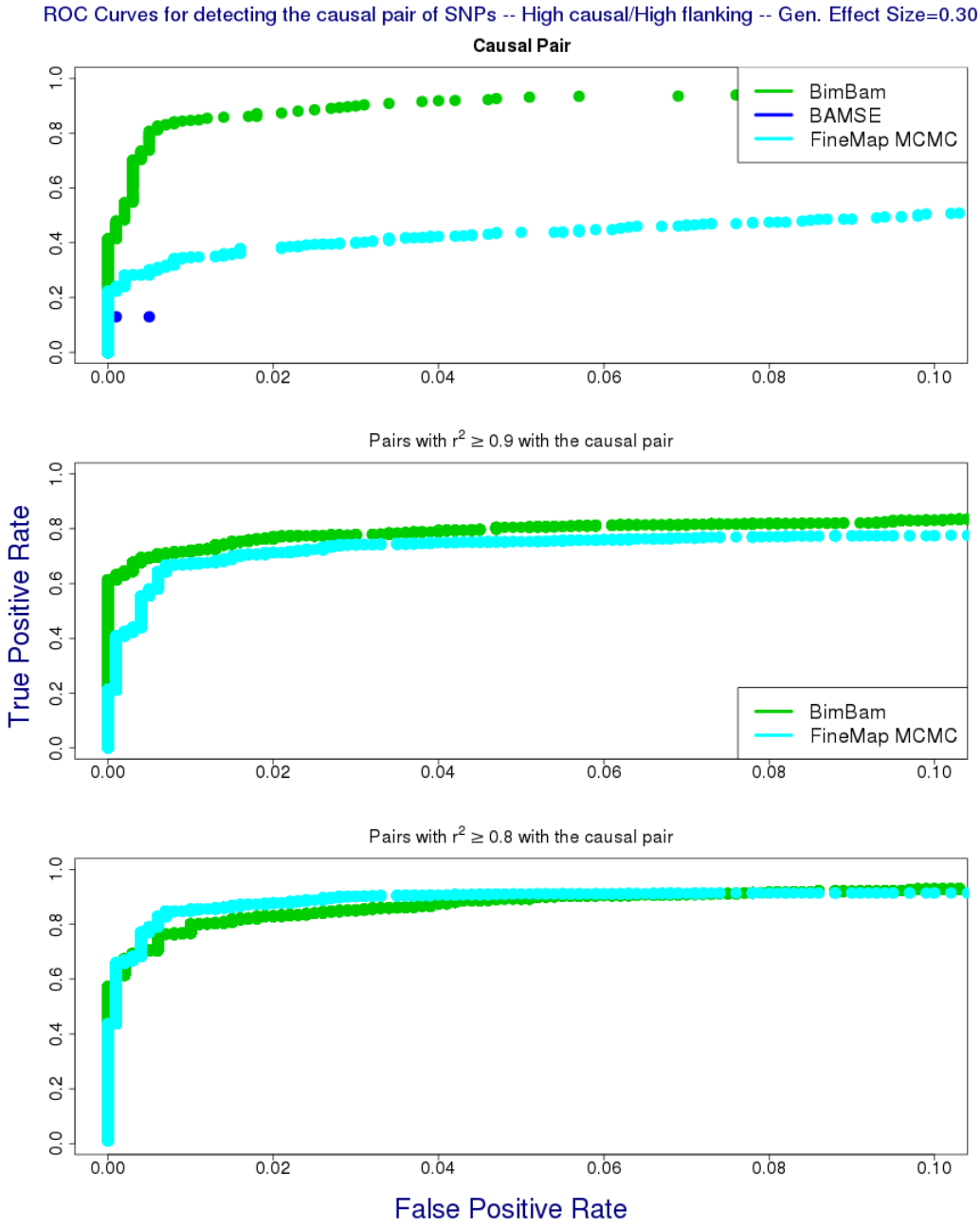

Figure 3.12: ROC curves to detect the causal pair of variants (**BAMSE**, **BIMBAM** and **FineMap MCMC**) and to detect the pairs in high LD ($r^2 = 0.80$ and 0.90) with the causal combination of SNPs (**BIMBAM** and **FineMap MCMC**). The horizontal axes represent the false positive rates while the vertical axes represent the true positive rates. The different methods are coded using the following colours: **BIMBAM** (green), **BAMSE** (blue), **FineMap MCMC** (cyan) (LD: high causal/high flanking).

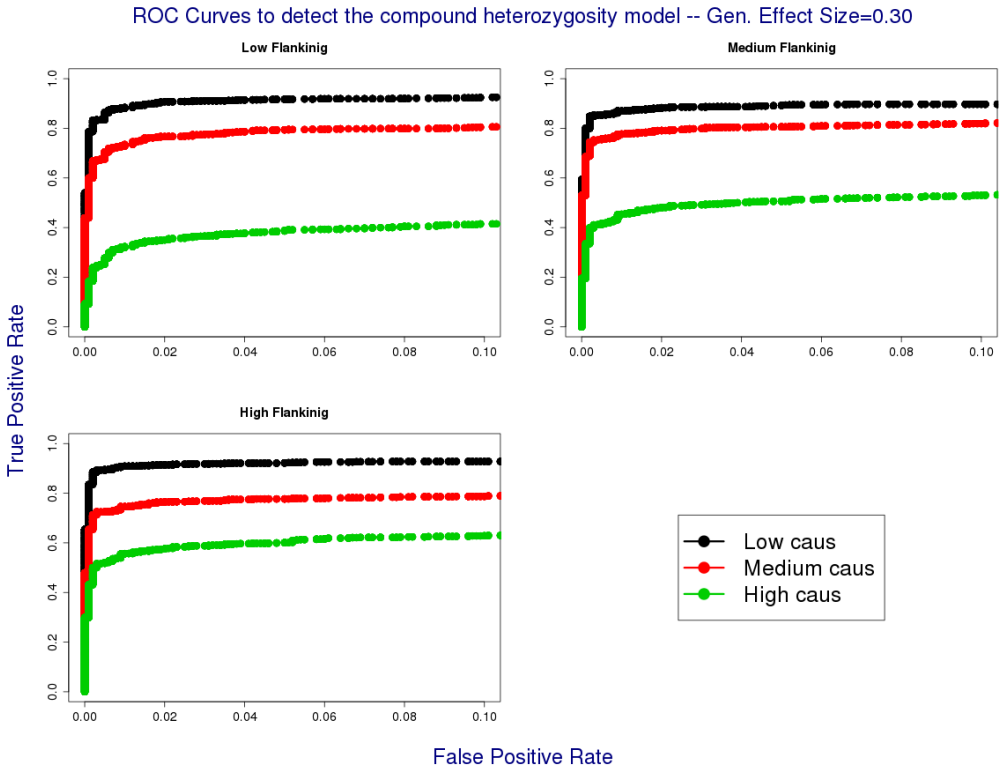

Figure 3.13: ROC curves to detect the compound heterozygosity model for the data under the examined structures of LD. The horizontal axes represent the false positive rates while the vertical axes represent the true positive rates. The three pairs of axes describe the LD structures between the causal SNPs and the flanking sites. The black ROC curves show the results for low LD between the causal SNPs, the red ROC curve is for medium LD between the causal SNPs and the green ROC curve is for high LD between the causal SNPs.

3.3 Examination of the Speed to Convergence

One of the important aspects of a Markov Chain Monte Carlo is its convergence to the true posterior distribution. Furthermore, this has to accompanied by a reasonable number of iterations which does not surpass the actual number of states in the model space. To perform more iterations than the total size of the model space would be counteractive since the reason for using an MCMC algorithm is to obtain a good approximation of the posterior distribution without having to calculate it analytically. To assess the speed to convergence of the Metropolis-Hastings algorithm, we calculate the Kullback-Leibler divergence between the estimated and true posterior marginal distributions of the SNPs given the data. We perform this for different scenarios of the proposal density where in each scenario, we consider different probabilities (weights) for the allowed actions. Table 3.2 presents these probabilities. We refer to the scenario where each action has equal weight as the uniform. Then, in turn for each action, we increase its weight hence constructing a medium and high scenario. For example, when we increase the weight for the first action, "Keep SNP, change model", the medium weight scenario assigns a probability of 0.5 to this action, while the second and third actions have probability equal to 0.2 and lastly, the action of deleting a SNP is given a probability of 0.1. When we increase 0.5 to 0.7, each of the remainig actions is given a probability of 0.1.

The true marginal posterior distribution of the SNPs is straight forward to calculate, but computationally expensive. The size of the model space makes it impractical to perform the calculations for each combination of SNP(s) and model, especially for a large number of variants. For the simulation study under consideration, the size of the model space sums to 64,551 possible states, calculated as follows:

$$\underbrace{1}_{\text{Null Model}} + \underbrace{100}_{\text{Single SNP}} \times \underbrace{2}_{\text{Model}} + \underbrace{\binom{100}{2}}_{\text{Pair of SNPs}} \times \underbrace{13}_{\text{Model}}.$$

The benefit of using the Metropolis-Hastings algorithm over the deterministic calculation of the entire model space is to achieve the true posterior distribution by using a smaller number of iterations than the total cardinality of the model space. For this reason, we examine the Kullback-Leibler divergence for four different number of iterations (burn in iterations are given in brackets): $10,000(5,000)$, $20,000(10,000)$, $30,000(15,000)$ and $40,000(20,000)$.

3.3.1 The Kullback Leibler Divergence

The Kullback-Leibler(KL) divergence is calculated in order to compare two probability distributions $P(x)$ and $Q(x)$ that are defined over the same values. This divergence is also known as the relative entropy and plays a key role in information theory and statistics. In discrete cases it is formulated as

$$KL(P||Q) = \sum_{\forall x \in X} P(x) \log \frac{P(x)}{Q(x)}. \tag{3.2}$$

Respectively, if P and Q are continuous distributions the summation becomes an integral over all possible values. The properties of the KL divergence are:

- $KL(P||Q) \geq 0$;
- $KL(P||Q) = 0 \quad iff \quad P(x) = Q(x) \quad \forall x \in X$;
- $KL(P||Q) \neq KL(Q||P)$.

Special conventions in the calculation of the Kullback-Leibler divergence are:

- $0 \log \frac{0}{0} = 0$;
- $0 \log \frac{0}{Q(x)} = 0 \; \forall \; x$;
- In the case where $Q(x) = 0$ for some x, we have $P(x) \log \frac{P(x)}{0}$. In this case we add a small deterministic error to each value of Q, obtained by taking the half of the minimum value of all the non-zero values of $Q(x)$.

	Weight – Action 1					
Actions	Uniform		Medium		High	
	$\|S\| = 1$	$\|S\| = 2$	$\|S\| = 1$	$\|S\| = 2$	$\|S\| = 1$	$\|S\| = 2$
Keep SNP, change model	0.25	$1/3$	0.5	0.6	0.7	0.8
Change SNP, keep model	0.25	$1/3$	0.2	0.3	0.1	0.1
Add SNP, change model	0.25	-	0.2	-	0.1	-
Delete SNP, change model	0.25	$1/3$	0.1	0.1	0.1	0.1
	Weight – Action 2					
Actions	Uniform		Medium		High	
	$\|S\| = 1$	$\|S\| = 2$	$\|S\| = 1$	$\|S\| = 2$	$\|S\| = 1$	$\|S\| = 2$
Keep SNP, change model	0.25	$1/3$	0.2	0.3	0.1	0.1
Change SNP, keep model	0.25	$1/3$	0.5	0.6	0.7	0.8
Add SNP, change model	0.25	-	0.2	-	0.1	-
Delete SNP, change model	0.25	$1/3$	0.1	0.1	0.1	0.1
	Weight – Action 3					
Actions	Uniform		Medium		High	
	$\|S\| = 1$	$\|S\| = 2$	$\|S\| = 1$	$\|S\| = 2$	$\|S\| = 1$	$\|S\| = 2$
Keep SNP, change model	0.25	$1/3$	0.2	$1/3$	0.1	$1/3$
Change SNP, keep model	0.25	$1/3$	0.2	$1/3$	0.1	$1/3$
Add SNP, change model	0.25	-	0.5	-	0.7	-
Delete SNP, change model	0.25	$1/3$	0.1	$1/3$	0.1	$1/3$
	Weight – Action 4					
Actions	Uniform		Medium		High	
	$\|S\| = 1$	$\|S\| = 2$	$\|S\| = 1$	$\|S\| = 2$	$\|S\| = 1$	$\|S\| = 2$
Keep SNP, change model	0.25	$1/3$	0.2	0.3	0.1	0.1
Change SNP, keep model	0.25	$1/3$	0.2	0.1	0.1	0.1
Add SNP, change model	0.25	-	0.1	-	0.1	-
Delete SNP, change model	0.25	$1/3$	0.5	0.6	0.7	0.8

Table 3.2: Probabilities assigned to the different actions in order to assess the speed of the model to converge to the truth. When $|S| = 2$ the action of adding a SNP is not allowed and therefore the probabilities are adjusted accordingly.

We treat the true marginal posterior distribution of the SNPs as P while Q will serve the role of the estimated marginal posterior distribution. The reasoning behind this is that the KL measures the additional amount of information when we use Q to code samples from P when theoretically P is unknown [10].

We now present the results of comparing the KL divergence between the true and estimated marginal posterior SNP distributions focusing on optimizing the circumstances under which we run the MCMC algorithm.

3.3.2 Results

In this section we present the KL divergence calculated for a single dataset from each of the following scenarios: Low causal/Low flanking, Low causal/Med flanking and Low causal/High flanking under both genetic effect sizes. We divide each figure into 4 pairs of axes, one for each action.

Figures 3.14 and E.14 show the KL divergence when r^2 between the two causal SNPs and between the causal and flanking sites is less than 0.4. We should bear in mind that the underlined causal model has $|S| = 2$. In Figure 3.14 when the probability on action "Keep SNP, change model" is 0.5, the KL divergence is slightly larger than for the default scenario at 10,000 iterations. Once we increase the iterations, we observe approximately equal divergence measures with the smallest value occurring at 40,000 iterations. As expected, the divergence between the true and the estimated posterior distributions tends to zero as the number of iterations increase. On the other hand, when we increase the probability of the same action to 0.7 there is an indication that the estimated distribution does not match the true distribution as closely. This is in accordance with the fact that such a high probability on the action which does not allow a change in S will not result in a convincing Metropolis-Hastings marginal posterior distribution. Similar behaviour is observed when we decrease the effect size to 0.20 for a larger number of iterations where for 10,000 and

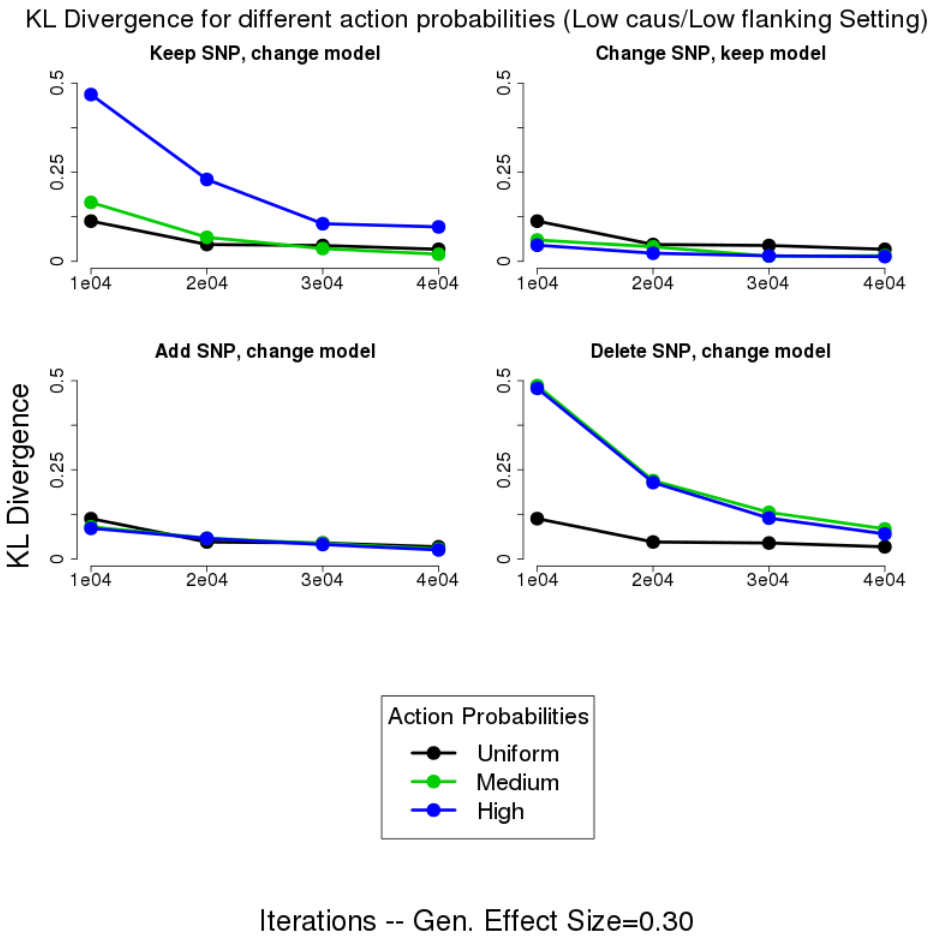

Figure 3.14: Kullback-Leibler divergence between the true and estimated marginal posterior distribution of each SNP for a single dataset for different number of iterations. The horizontal axis describes the number of iterations while the vertical axis shows the KL divergence. The black line represents the uniform action probabilities scenario, the green line represents the medium action probabilities scenario and the blue line denotes the high action probabilities scenario (LD: low causal/low flanking).

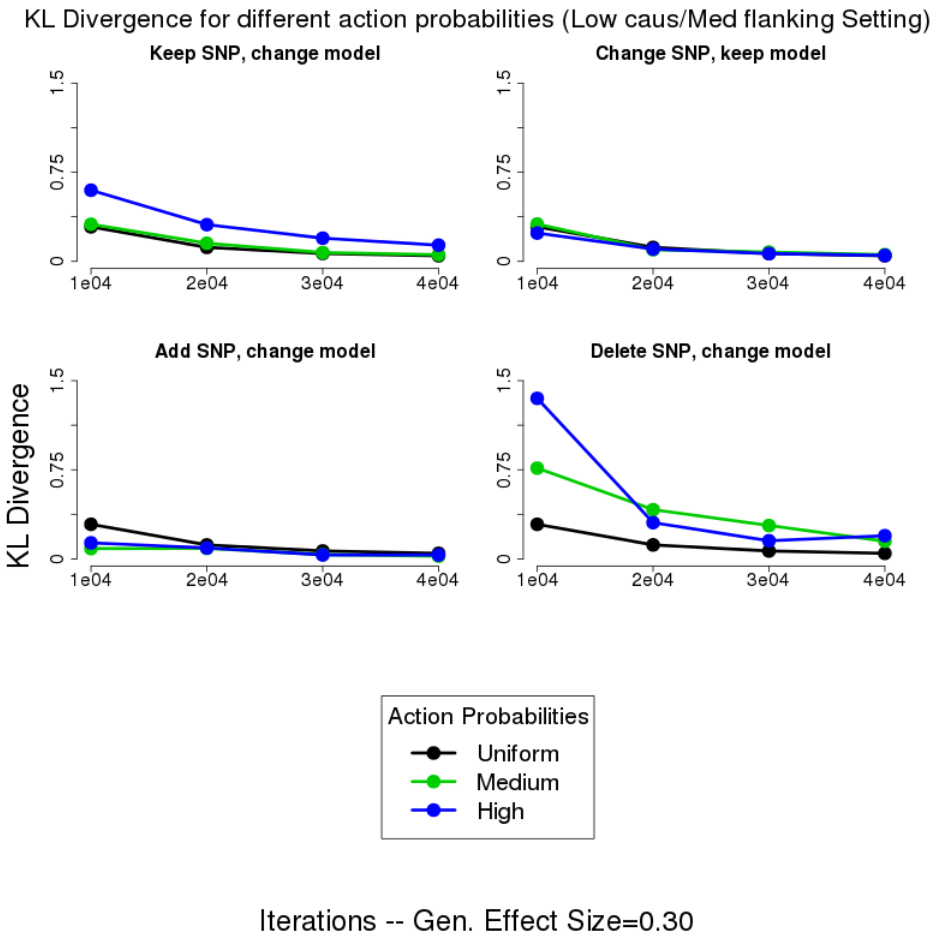

Figure 3.15: Kullback-Leibler divergence between the true and estimated marginal posterior distribution of each SNP for a single dataset for different number of iterations. The horizontal axis describes the number of iterations while the vertical axis shows the KL divergence. The black line represents the uniform action probabilities scenario, the green line represents the medium action probabilities scenario and the blue line denotes the high action probabilities scenario (LD: low causal/med flanking).

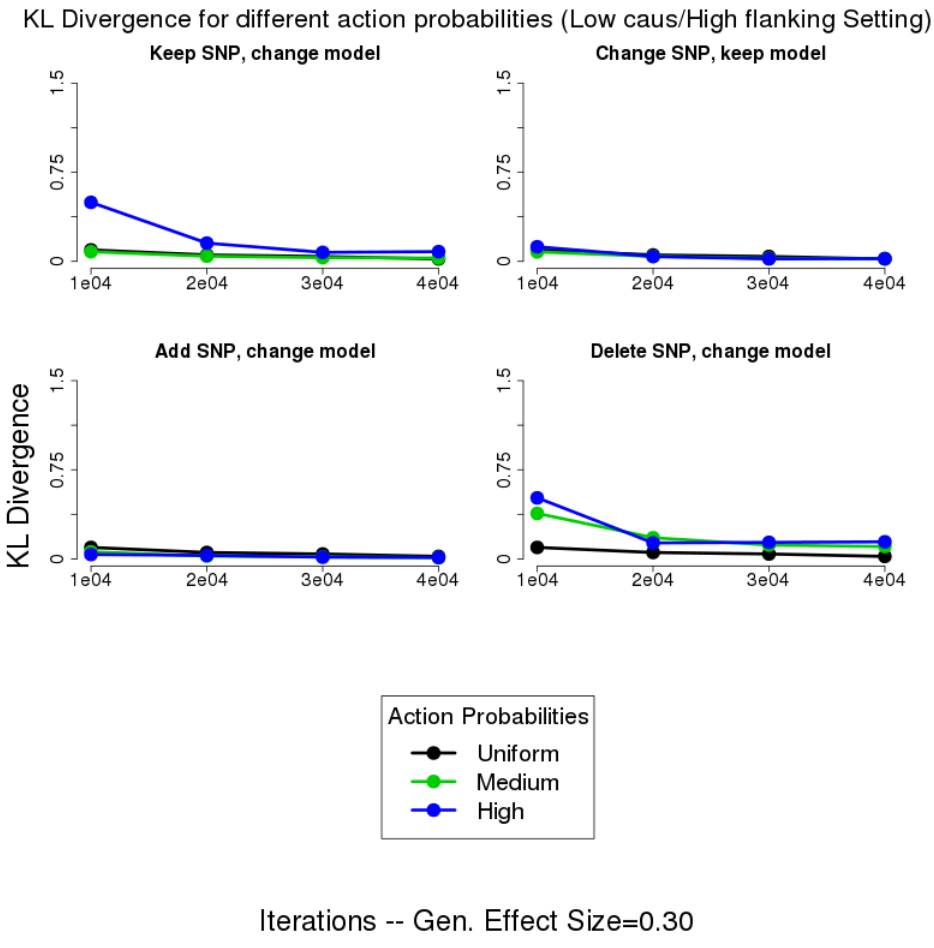

Figure 3.16: Kullback-Leibler divergence between the true and estimated marginal posterior distribution of each SNP for a single dataset for different number of iterations. The horizontal axis describes the number of iterations while the vertical axis shows the KL divergence. The black line represents the uniform action probabilities scenario, the green line represents the medium action probabilities scenario and the blue line denotes the high action probabilities scenario (LD: low causal/high flanking).

20,000 iterations the KL divergence is generally lower.

Unlike for the first action, when we add more probability to "Change SNP, keep the model" the MCMC output converges quicker to the truth and is more accurate, since the green and blue lines lie below the black which denotes the default scenario. Furthermore, a higher probability for adding a SNP to S does not alter the performance of the model and the three coloured lines are almost indistinguishable, especially for the smaller effect size. The fourth pair of axes in each of Figures 3.14 and E.14 shows the results when we add more weight to the action of deleting a SNP. We observe a significant difference between the default and the other two scenarios. The KL measure appears to be larger when this action has a higher probability, particularly for the larger genetic effect size.

Moving now to the setting where the causal SNPs are in low LD between them but in medium correlation with the flanking sites, we do not see significant fluctuations in the interpretation of the results (Figure 3.15 and E.15). A notable change in this setting is the expanded scale on the vertical axis as we observe higher KL divergence estimates for one of the actions ("Delete SNP, change model"). As in the above setting of low flanking LD, high probability on the action which keeps the current SNP(s) selection and changes the model, leads to an elevated KL divergence while for two of the actions which perform a change to the SNP set S ("Change SNP, keep model" and "Add SNP, change model"), different weights give approximately equal results for larger numbers of iterations. At 10,000 iterations we see some differentiation and the default weights allocation gives the least accurate results ("Add SNP, change model"). On the other hand, the default weights scenario produces the best divergence estimates for the action "Delete SNP, change model".

Lastly, when the r^2 between the causal variants and flanking loci increases to more than 0.70 we can confidently say that, when we impose equal probabilities for each action the MCMC chains produces the target posterior distribution using a smaller number of iterations (Figure 3.16 and E.16). The general aftermath resembles that of the preceeding two LD scenarios.

3.4 Sensitivity to Prior Assumptions

In general, the posterior distribution of the model parameters can either overestimate or underestimate different aspects of the true posterior uncertainty. It is sensible, when using an MCMC algorithm, to assess the sensitivity of its output for different specifications of the prior distribution and to examine how these speceifcations account for this posterior uncertainty [29].

In this section we consider a range of prior probabilities on the cardinality of the SNP set S. Different prior probability on $|S|$ leads to different prior belief on the number of SNPs which account for the variation in the phenotype. We calculate the KL divergence between the estimated and true marginal posterior distribution of the SNPs when the Metropolis-Hastings algorithm is executed for $p \in \{0.20, 0.40, 0.60, 0.80, 0.95\}$. As p increases, more support is added to $|S|$ being equal to 2. Figure 3.17 demonstrates the behaviour of the KL divergence under the scenarios of low causal and low, medium and high flanking LD respectively. If apriori we support smaller values for p, the estimated posterior distribution does not converge to the truth. This discrepancy increases when r^2 between the causal and flanking sites is more than 70%. The estimated posterior distribution approaches the true underlying distribution of the SNPs when p equals 0.6, whereas the two become approximately equal for p equal to 0.80 and 0.95, for all three examinded LD structures. The analogous results under a genetic effect of size 0.20 can be found in Appendix E and Figure E.17.

3.5 Acceptance Rate of the Metropolis-Hastings Algorithm

A Metropolis-Hastings algorithm can be characterized by the proportion of jumps accepted. This is a key aspect of this MCMC implementation and should be addressed, as the endless choices in the jumping rules contribute to different results. An efficient Metropolis-Hastings

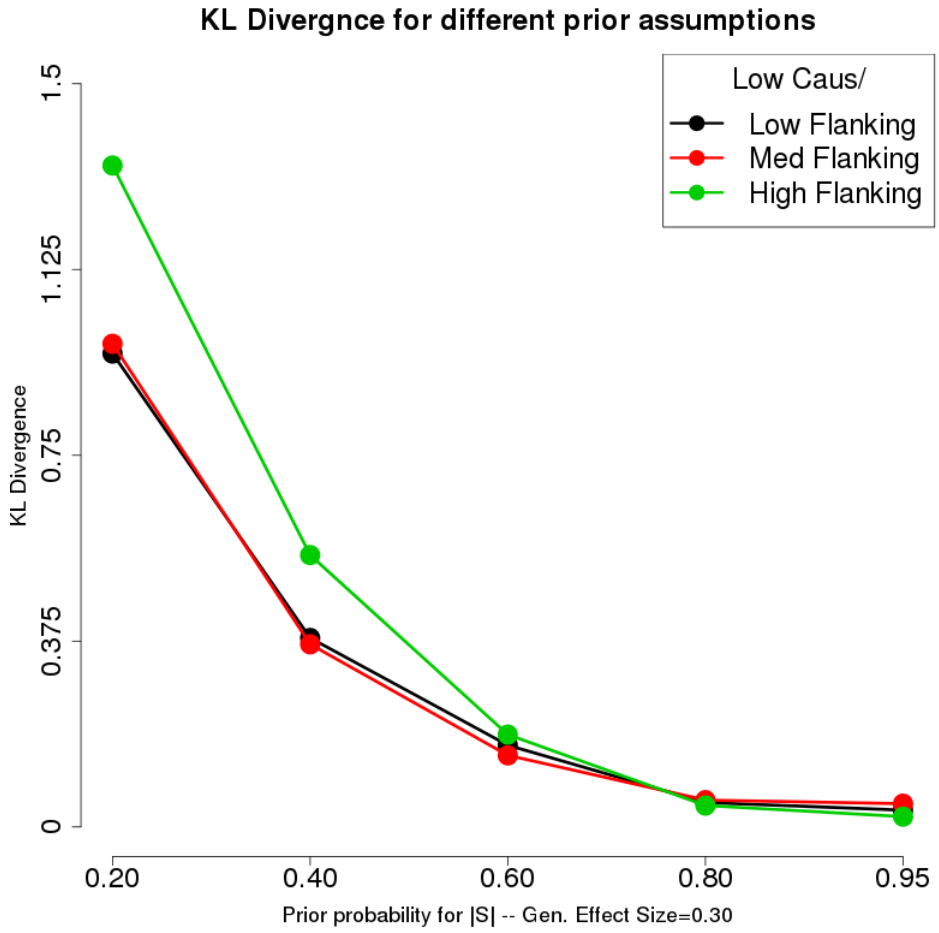

Figure 3.17: Kullback-Leibler divergence between the true and estimated marginal posterior distribution of each SNP for a single dataset for different prior probabilities for $|S|$. The horizontal axis presents the different values of p and the vertical axis shows the KL divergence. The LD between the causal SNPs is low while the flanking LD is low (black line), medium (red line) and high (green line).

algorithm should explore the model space freely in order converge to the target distribution. In this case we say the chain should be mixing well. Therefore, what should be avoided is a sticky, slow-mixing chain which produces a highly correlated set of samples. It has been suggested that acceptance rates should be near 20% when altering a vector of parameters or 40% when altering a single parameter at a time [29].

Effect Size		LD Setting		
		Low Flanking	Medium Flanking	High Flanking
0.30	Low Causal	0.377	0.251	0.244
	Medium Causal	0.135	0.132	0.157
	High Causal	0.140	0.155	0.134
		Low Flanking	Medium Flanking	High Flanking
0.20	Low Causal	0.396	0.366	0.355
	Medium Causal	0.316	0.298	0.320
	High Causal	0.271	0.283	0.269

Table 3.3: Mean acceptance rate across the 1000 datasets for each LD setting and the different genetic effect sizes.

Table 3.3 shows the resulting acceptance rates of the Metropolis-Hastings algorithm averaged over the 1,000 examined datasets for both genetic effect sizes and all LD scenarios. When the LD between the causal SNPs is low and the flanking LD increases, we observe a decrease in the average of the acceptance rates. Since we have examined the convergence of such chains in a previous section via the KL divergence, we can say that proposed states within the algorithm are more frequently rejected and convergence is achieved before completing 40,000 iterations. When we assume a smaller effect size, the average acceptance rate is near 40% for the aforementioned LD scenarios although the decreasing behaviour persists as the flanking LD increases. Additionally, as the LD between the causal SNPs increases the rates decrease but without falling below 26.9%. On the other hand, rates fall between 13% and 15% for these LD cases for the larger genetic effect size. This leads us to the conclusion that such complex LD structures require more than 40,000 iterations to achieve convergence. This is also justified by the KL divergence for the high caus/high flanking scenario, which at 40,000 iterations is 0.12, while for the med caus/high flanking scenario it equals 0.05.

3.6 Summary

Over the course of this chapter, we have performed extensive simulations in order to assess the performance and behaviour of our proposed fine mapping model and how they relate to existing approaches. We have constructed haplotype sets using **HAPGEN** and simulated phenotypes under the compound heterozygosity model under two genetic effect sizes and different LD structures amongst causal and flanking variants.

We have compared the power of our fine mapping model and of existing models to identify the causal variants, and have concluded that our approach indeed leads to a higher power to detect haplotype effects which are not directly visible when dealing with genotype data. In order to be more explicit, under low or medium levels of LD, where r^2 does not go over 0.70, the haplotype model outperforms the rest of the models in the detection of the causal SNPs marginally while for larger correlation estimates, the power of the different approaches reaches the same levels. Along with **BAMSE** and **BIMBAM**, we have also tested the power of our framework to detect the causal pair. As the flanking LD increases, **BIMBAM** outperforms the other two in detecting the actual causal pair while the haplotype fine mapping model achieves higher power when we account for the detection of pairs highly correlated with the truth. When the LD structure is simpler and there is not any significant correlation neither between the pair of causal SNPs nor between the flanking positions, the fine mapping model performs better. Further, the power to detect the true genetic model is affected once again by the correlation grading, where higher LD leads to lower power.

Further, we use the Kullback-Leibler (KL) divergence measure to assess the speed of the Metropolis-Hastings algorithm to converge. We examine how different weights to the actions of the proposal density affect the discrepancy between the true and estimated marginal posterior distributions of the SNPs for different lengths of the MCMC chains. As a way of assessing the sensitivity of the algorithm to prior assumptions, we consider a set of prior

probabilities on $|S|$ and use the KL divergence. Finally, we study the acceptance rates of the Metropolis-Hastings algorithm and how these vary according to the size of the genetic effect and distinctive LD structures.

Part II

A Novel Approach to Meta Analysis

Chapter 4

Methodology

As we have seen in Chapter 1 it is often desirable, when a collection of studies is accessible, to perform a joint analysis on the collection of possibly tens of thousands of individuals. The use of meta analysis is becoming more and more popular as it leads to elevated power to identify associated variants with potentially small effect sizes. This approach has become possible due to the advent of genotype imputation [52] which has allowed studies to be combined at common sets of SNPs. The literature is rich in such studies for several diseases and conditions including smoking quantity [44], type 2 diabetes [59], Crohn's disease [25], ischemic stroke [9] and rheumatoid arthritis [84]. An important factor that drives the analysis of such studies is the similarity or dissimilarity of the effect sizes of each study. The most popular approach to conduct a joint analysis is the fixed effects model where the main assumption is that the cohort specific effect sizes are very much alike. There are, though, circumstances for which this assumption will not hold and the effect sizes between studies may differ leading to the presence of heterogeneity. Heterogeneity can be attributed to numerous factors such as different populations between studies, environmental factors due to different regions of origin of the individuals, difference in linkage disequilibrium structures as well as various genotyping platforms and imputation algorithms. Statistical heterogeneity can occur due to design factors [35].

We propose an approach which adequately accommodates both fixed and random effects within a Bayesian framework. In order to make the connection with the fine mapping model presented in the previous chapters, we assume that analysis has been performed on each cohort individually and has concluded to a causal SNP set S and a causal genetic model M_S have been chosen for further analysis. Our meta-analysis approach can be employed in the usual way at all SNPs in a given region but it can also allow for combining models from more than one SNP. The input data can either be genotype or haplotype data and different genetic models can be chosen to fit the data, as they have been previously described. The novelty in our approach occurs when we account for the probability that the effect exists in each study hence not "forcing" spurious associations. We then conclude to a model space of cardinality four:

1. Random Effects Model, all cohorts associated ;
2. Fixed Effects Model, all cohorts associated ;
3. Random Effects Model, subset of cohorts associated ;
4. Fixed Effects Model, subset of associated.

We assume the additive model at each SNP as it is the most commonly used. Nevertheless, our approach is easily extended to other genetic models. Before describing the formulation for each model, we introduce some essential notation that will be used throughout this section:

- K is the total number of cohorts ;
- $\boldsymbol{N} = (N_1, ..., N_K)$ is the vector of sample size for each study;
- $\boldsymbol{Y}_i$ is the phenotype vector for cohort i, $i = 1, ..., K$;
- $\boldsymbol{X}_i$ is he design matrix for cohort i, $i = 1, ..., K$;
- μ and μ_F is the phenotype mean of individuals carrying the reference genotype under the random and fixed effects models respectively and ν its precision ;

- τ_i and $\tau_{F,i}$ is the phenotypic precision (inverse to the phenotypic variance) for the random and fixed effects models respectively for study i, $i = 1, ..., K$.

4.1 The Random Effects Model, All Cohorts Associated

The formulation of the random effects model assuming an additive genetic model and all cohorts to be associated is as follows:

$$\mathbf{Y}_i|\mu_i, \boldsymbol{\beta}_i, \tau_i \sim N(\mu_i + \mathbf{G}_i\boldsymbol{\beta}_i, \tau_i^{-1}\mathbf{I}_{N_i}), \qquad i = 1, ..., K \tag{4.1}$$

$$\mu_i \sim N(0, \nu^{-1}) \tag{4.2}$$

$$\boldsymbol{\beta}_i|\boldsymbol{\beta}, \lambda \sim N(\boldsymbol{\beta}, \lambda^{-1}\mathbf{V}_p) \tag{4.3}$$

$$\boldsymbol{\beta}|\psi \sim N(\mathbf{0}, \psi^{-1}\mathbf{W}_p) \tag{4.4}$$

$$\tau_i \sim \Gamma(j, l) \tag{4.5}$$

$$\lambda \sim \Gamma(c, d) \tag{4.6}$$

$$\psi \sim \Gamma(a, b) \tag{4.7}$$

where

- $\mathbf{G}_i$ is the genotypes matrix of the individuals and $\mathbf{X}_i = (\mathbf{1}_{N_i}, \mathbf{G}_i)$;
- $\boldsymbol{\beta}_i$ is the vector of parameters for cohort i;
- $\boldsymbol{\beta}$ is the mean of the Normal distribution that $\boldsymbol{\beta}_i$ is drawn from, for all i;
- λ is the between cohort precision;
- ψ is the precision of $\boldsymbol{\beta}$;
- p is the dimension of $\boldsymbol{\beta}_i$ and $\boldsymbol{\beta}$;
- $\mathbf{V}_p$ and $\mathbf{W}_p$ are diagonal matrices of dimension $p \times p$.

Each cohort's effect size is characterised by $\boldsymbol{\beta}_i$ which is drawn from a Normal distribution with mean $\boldsymbol{\beta}$ and precision λ. The value of λ quantifies the amount of heterogeneity present in the data. We proceed by expressing the joint likelihood of the model

$$P(\mathbf{Y}_1, ..., \mathbf{Y}_K | \mu_1, ..., \mu_K, \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_K, \boldsymbol{\beta}, \tau_1, ..., \tau_K, \lambda, \psi) = \prod_{i=1}^{K} P(\mathbf{Y}_i | \mu_i, \boldsymbol{\beta}_i, \boldsymbol{\beta}, \tau_i, \lambda, \psi), \tag{4.8}$$

while the joint prior distribution of the parameters is

$$P(\mu_1, ..., \mu_K, \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_K, \boldsymbol{\beta}, \tau_1, ..., \tau_K, \lambda, \psi) = \prod_{i=1}^{K} P(\beta_i | \boldsymbol{\beta}, \lambda) P(\mu_i) P(\tau_i) \times P(\boldsymbol{\beta} | \psi) P(\lambda) P(\psi). \tag{4.9}$$

We obtain the joint posterior distribution of the parameters given the data up to the normalising constant by multiplying together the joint likelihood and the joint prior distribution

$$\begin{aligned} P(\mu_1, ..., \mu_K, \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_K, \boldsymbol{\beta}, \tau_1, ..., \tau_K, \lambda, \psi | \mathbf{Y}_1, ..., \mathbf{Y}_K) &\propto \prod_{i=1}^{K} \{P(\mathbf{Y}_i | \mu_i, \boldsymbol{\beta}_i, \boldsymbol{\beta}, \tau_i, \lambda, \psi) P(\beta_i | \boldsymbol{\beta}, \lambda) \\ &\quad P(\mu_i) P(\tau_i)\} \times P(\boldsymbol{\beta} | \psi) P(\lambda) P(\psi) \\ &\propto \prod_{i=1}^{K} \Big(\tau_i^{(N_i/2)+j-1} \times \\ &\quad \exp\{-\frac{\tau_i}{2}\{(\mathbf{Y}_i - \mu_i \mathbf{1}_{N_i} - \mathbf{G}_i \boldsymbol{\beta}_i)^T (\mathbf{Y}_i - \mu_i \mathbf{1}_{N_i} - \mathbf{G}_i \boldsymbol{\beta}_i) + 2l\}\} \\ &\quad \times \lambda^{p/2} \exp\{-\frac{\lambda}{2} (\boldsymbol{\beta}_i - \boldsymbol{\beta})^T \mathbf{V}_p^{-1} (\boldsymbol{\beta}_i - \boldsymbol{\beta})\} \\ &\quad \times \nu^{\frac{1}{2}} \exp\{-\frac{\nu}{2} \mu_i^2\}\Big) \\ &\quad \times \psi^{p/2+a-1} \exp\{-\frac{\psi}{2} (\boldsymbol{\beta}^T \mathbf{W}_p^{-1} \boldsymbol{\beta} + 2b)\} \\ &\quad \times \lambda^{c-1} \exp\{-d\lambda\}. \end{aligned} \tag{4.10}$$

4.1.1 The Gibbs Sampler

In order to perform inference on this model we need to employ an MCMC algorithm and in particular the Gibbs sampler which allows inferences to be drawn from complex, multidimensional posterior distributions. The Gibbs sampler which can be characterised by computational simplicity is a special case of the Metropolis-Hastings algorithm. A proposal is generated from the appropriate conditional posterior distribution and is always accepted [79]. It generates a sequence of samples form the joint distribution by sampling successively from all fully conditional posterior distributions. A draw of a single parameter is regarded as a draw from its marginal posterior distribution. The derivation of each of the fully conditional posterior distributions requires extracting terms which are function of each parameter from the joint posterior density. The general framework of the Gibbs sampling algorithm is presented below.

4.1.1.1 Full Conditional Posterior Distributions

Let $(\theta_1, \theta_2, ..., \theta_p)$ be the vector of parameters whose posterior distribution is proportional to $p(\theta_1, ...\theta_{i-1}, \theta_i, \theta_{i+1}, ...\theta_p|\boldsymbol{y})$. We define the vector $\boldsymbol{\theta}_{-i}$ to be the vector equal to $\boldsymbol{\theta}$ with its i^{th} component, $\boldsymbol{\theta}_i$, deleted. Then the fully conditional posterior distribution of $\boldsymbol{\theta}_i$ is

$$\begin{aligned} p(\boldsymbol{\theta}_i|\boldsymbol{\theta}_{-i}, \boldsymbol{y}) &= \frac{p(\theta_1, ...\theta_{i-1}, \theta_i, \theta_{i+1}, ...\theta_p|\boldsymbol{y})}{\int p(\theta_1, ...\theta_{i-1}, \theta_i, \theta_{i+1}, ...\theta_p|\boldsymbol{y}) d\boldsymbol{\theta}_i} \\ &\propto p(\theta_1, ...\theta_{i-1}, \theta_i, \theta_{i+1}, ...\theta_p|\boldsymbol{y}). \end{aligned} \tag{4.11}$$

4.1.1.2 The Gibbs Sampling Algorithm

Consider the above vector of parameters with posterior posterior density $p(\theta_1, \theta_2, ..., \theta_p|\boldsymbol{y})$, known up to proportionality.

Assigning reasonable starting values $(\theta_1^{(0)}, \theta_2^{(0)}, ..., \theta_p^{(0)})$ the Gibbs sampler, at each iteration t, cycles through the components of $\boldsymbol{\theta}$, drawing each component conditional on all the others [29]:

$$\theta_i^t \sim p(\theta_i | \theta_{-i}^{t-1}, \boldsymbol{y}),$$

where $\boldsymbol{\theta}_{-i}^{t-1} = (\theta_1^t, ..., \theta_{i-1}^t, \theta_{i+1}^{t-1}, ..., \theta_p^{t-1})$, the vector of the current values of $\boldsymbol{\theta}$ except for θ_i. The chain's initial iterations are discarded (burn-in) while the rest of the draws $(\theta_1^{(i)}, \theta_2^{(i)}, ..., \theta_p^{(i)})$ for a large number of iterations i are regarded as samples from the normalized posterior distribution with density

$$p(\theta_1, \theta_2, ..., \theta_p | \boldsymbol{y}) / \int p(\theta_1, \theta_2, ..., \theta_p | \boldsymbol{y}) d\theta_1, ..., \theta_p$$

and the component $\theta_j^{(i)}$ is a draw from its marginal posterior distribution with density [79]

$$p(\theta_j) / \int p(\theta_j | \boldsymbol{y}) d\theta_j.$$

The simplicity of the Gibbs sampler is enriched in the present setting and the derivation of the fully conditional posterior distributions becomes even more straight forward as we use conjugate prior distributions for each of our parameters resulting to kernels of the densities of well known distributions such as the Normal and the Gamma. The details of each parameter's fully conditional posterior distribution can be found in Appendix D.

4.2 The Fixed Effects Model, All Cohorts Associated

The simplest model in our meta-analysis approach is the fixed effects model where all cohorts are believed to contribute to the phenotype. It has the smallest number of parameters since each cohort's effect size $\boldsymbol{\beta}_i$, $i = 1, ..., K$ is collapsed to a single parameter which we denote by $\boldsymbol{\beta}_F$ following the assumption of the fixed effects model of homogeneity across studies. The formulation of this model is as follows:

$$\mathbf{Y}_i|\mu_{F,i}, \boldsymbol{\beta}_F, \tau_{F,i} \sim N(\mu_{F,i} + \mathbf{G}_i\boldsymbol{\beta}_F, \tau_{F,i}^{-1}\mathbf{I}_{N_i}), \qquad i = 1, ..., K \tag{4.12}$$

$$\mu_{F,i} \sim N(0, \nu^{-1}) \tag{4.13}$$

$$\boldsymbol{\beta}_F|\psi_F \sim N(\mathbf{0}, \psi_F^{-1}\mathbf{W}_p) \tag{4.14}$$

$$\tau_{F,i} \sim \Gamma(j_F, l_F) \tag{4.15}$$

$$\psi_F \sim \Gamma(a_F, b_F) \tag{4.16}$$

where

- $\boldsymbol{\beta}_F$ is the effect size parameter, identical for all cohorts;
- ψ_F is the precision of $\boldsymbol{\beta}_F$.

We also add the same subscript to the hyperparameters, the scale and shape of the two Gamma distributions in order to distinguish them from those of the random effects model but this does not necessarily imply that they differ.

The joint likelihood can then be written as

$$P(\mathbf{Y}_1, ..., \mathbf{Y}_K|\mu_{F,1}, ..., \mu_{F,K}, \boldsymbol{\beta}_F, \tau_{F,1}, ..., \tau_{F,K}, \psi_F) = \prod_{i=1}^{K} P(\mathbf{Y}_i|\mu_{F,i}, \boldsymbol{\beta}_F, \tau_{F,i}, \psi_F) \tag{4.17}$$

while the joint distribution of the parameters is

$$P(\mu_{F,1}, ..., \mu_{F,K}, \boldsymbol{\beta}_F, \tau_{F,1}, ..., \tau_{F,K}, \psi_F) = \prod_{i=1}^{K} P(\mu_{F,i})P(\tau_{F,i}) \times P(\boldsymbol{\beta}_F|\psi_F)P(\psi_F). \tag{4.18}$$

The posterior distribution of the parameters given the data can be then obtained as

$$P(\mu_{F,1},...,\mu_{F,K},\boldsymbol{\beta}_F,\tau_{F,1},...,\tau_{F,K},\psi_F|\mathbf{Y}_1,...,\mathbf{Y}_K) \propto \prod_{i=1}^{K} P(\mathbf{Y}_i\mu_{F,i},\boldsymbol{\beta}_F,\tau_{F,i},\psi_F)P(\mu_{F,i})P(\tau_{F,i}) \tag{4.19}$$

$$\times P(\boldsymbol{\beta}_F|\psi_F)P(\psi_F)$$

$$\propto \prod_{i=1}^{K} \Big(\tau_{F,i}^{(N_i/2)+j_F-1}$$

$$\times \exp\{-\frac{\tau_{F,i}}{2}\{(\mathbf{Y}_i-\mu_{F,i}\mathbf{1}_{N_i}-\mathbf{G}_i\boldsymbol{\beta}_F)^T(\mathbf{Y}_i-\mu_{F,i}\mathbf{1}_{N_i}-\mathbf{G}_i\boldsymbol{\beta}_F)+2l_F\}\}$$

$$\times\nu^{1/2}\exp\{-\frac{\nu}{2}\mu_{F,i}^2\}\Big) \times \psi_F^{p/2+a_F-1}\exp\{-\frac{\psi_F}{2}({\boldsymbol{\beta}_F}^T\mathbf{W}_p^{-1}\boldsymbol{\beta}_F+2b_F)\}.$$

We proceed, as with the random effects model in the previous section, with the derivation of the full conditional posterior distributions for each parameter and the formulation of the Gibbs sampler in order to obtain samples of the joint posterior distribution in 4.19. The details of each parameter's fully conditional posterior distribution can be found in Appendix D.

4.3 The Random Effects Model, Subset of Cohorts Associated

The third model that composes our meta-analysis methodology is the random effects model which allows for cohorts *not* to exhibit any genetic effect at the examined variant or variants. We hope that this will increase the power to detect association when only a subset of cohorts is truly associated without predicting spurious associations by forcing all cohorts to have a non-zero effect size. To achieve this we introduce an extra vector of parameters named $\boldsymbol{h} = (h_1,...h_K)$, $i = 1,...,K$. Each h_i takes the value 1 or 0 denoting association or

non-association.

The formulation of this model is very similar to that of the random effects model with all cohorts believed to be associated. The additional information incorporated now can be seen in the first line of the following

$$\mathbf{Y}_i|\mu_i, \boldsymbol{\beta}_i, \tau_i, h_i \sim N(\mu_i + h_i\mathbf{G}_i\boldsymbol{\beta}_i, \tau_i^{-1}\mathbf{I}_{N_i}), \qquad i = 1, ..., K \tag{4.20}$$

$$\mu_i \sim N(0, \nu^{-1}) \tag{4.21}$$

$$\boldsymbol{\beta}_i|\boldsymbol{\beta}, \lambda \sim N(\boldsymbol{\beta}, \lambda^{-1}\mathbf{V}_p) \tag{4.22}$$

$$\boldsymbol{\beta}|\psi \sim N(\mathbf{0}, \psi^{-1}\mathbf{W}_p) \tag{4.23}$$

$$h_i \sim Bernoulli(\pi) \tag{4.24}$$

$$\tau_i \sim \Gamma(j, l) \tag{4.25}$$

$$\lambda \sim \Gamma(c, d) \tag{4.26}$$

$$\psi \sim \Gamma(a, b) \tag{4.27}$$

$$\pi \sim B(e, f) \tag{4.28}$$

The distribution of cohort's i phenotype differs according to the value of h_i and more specifically this difference is observed at the mean of the distribution. Furthermore, we impose a Bernoulli distribution on the h_i's with probability of success π which, subsequently, has a Beta distribution with parameters e and f. This can be regarded as the full model since the other three that make up this meta-analysis approach are special cases of it. Before we write down the joint likelihood of this model we define the sets S_1 and S_0 to contain the indices for the associated and non-associated cohorts respectively. Hence, $S_1 = \{i : h_i = 1\}$ and $S_0 = \{i : h_i = 0\}$. Then

$$P(\mathbf{Y}_1, ..., \mathbf{Y}_K|\mu_1, ..., \mu_K, \boldsymbol{\beta}_{i:i\in S_1}, \boldsymbol{\beta}, \tau_1, ..., \tau_K, \lambda, \psi, h_1, ..., h_K, \pi) = \tag{4.29}$$

$$\prod_{i\in S_1} P(\mathbf{Y}_i|\mu_i, \boldsymbol{\beta}_i, \boldsymbol{\beta}, \tau_i, \lambda, \psi, h_i, \pi) \times \prod_{i\in S_0} P(\mathbf{Y}_i|\mu_i, \tau_i, h_i, \pi),$$

with joint prior distribution of the parameters

$$P(\mu_1, ..., \mu_K, \boldsymbol{\beta}_{i:i\in S_1}, \boldsymbol{\beta}, \tau_1, ..., \tau_K, \lambda, \psi, h_1, ..., h_K, \pi) = \tag{4.30}$$
$$\prod_{i\in S_1} P(\mu_i)P(\boldsymbol{\beta}_i|\boldsymbol{\beta}, \lambda)P(\tau_i)P(h_i|\pi) \times \prod_{i\in S_0} P(\mu_i)P(\tau_i)P(h_i|\pi)$$
$$\times P(\boldsymbol{\beta}|\psi)P(\lambda)P(\psi)P(\pi)$$

and finally the joint posterior distribution

$$P(\mu_1, ..., \mu_K, \boldsymbol{\beta}_{i:i\in S_1}, \boldsymbol{\beta}, \tau_1, ..., \tau_K, \lambda, \psi, h_1, ..., h_K, \pi|\mathbf{Y}_1, ..., \mathbf{Y}_K) \propto \tag{4.31}$$
$$\prod_{i\in S_1} P(\mathbf{Y}_i|\mu_i, \boldsymbol{\beta}_i, \boldsymbol{\beta}, \tau_i, \lambda, \psi, h_i, \pi)P(\mu_i)P(\beta_i|\boldsymbol{\beta}, \lambda)P(\tau_i)P(h_i|\pi)$$
$$\times \prod_{i\in S_0} P(\mathbf{Y}_i|\mu_i, \tau_i, h_i, \pi)P(\mu_i)P(\tau_i)P(h_i|\pi) \times P(\boldsymbol{\beta}|\psi)P(\lambda)P(\psi)P(\pi) \propto$$
$$\prod_{i\in S_1} \Big(\tau_i^{(N_i/2)} \exp\{-\frac{\tau_i}{2}(\mathbf{Y}_i - \mu_i\mathbf{1}_{N_i} - \mathbf{G}_i\boldsymbol{\beta}_i)^T(\mathbf{Y}_i - \mu_i\mathbf{1}_{N_i} - \mathbf{G}_i\boldsymbol{\beta}_i)\}$$
$$\times\nu^{1/2} \exp\{-\frac{\nu}{2}\mu_i^2\}\lambda^{p/2} \exp\{-\frac{\lambda}{2}(\boldsymbol{\beta}_i - \boldsymbol{\beta})^T\mathbf{V}_p^{-1}(\boldsymbol{\beta}_i - \boldsymbol{\beta})\} \times \tau_i^{j-1} \exp\{-l\tau_i\} \times \pi\Big)$$
$$\times \prod_{i\in S_0} \Big(\tau_i^{(N_i/2)} \exp\{-\frac{\tau_i}{2}(\mathbf{Y}_i - \mu_i\mathbf{1}_{N_i})^T(\mathbf{Y}_i - \mu_i\mathbf{1}_{N_i})\}$$
$$\times\nu^{1/2} \exp\{-\frac{\nu}{2}\mu_i^2\}\tau_i^{j-1} \exp\{-l\tau_i\} \times (1 - \pi)\Big)$$
$$\times\psi^{p/2} \exp\{-\frac{\psi}{2}\boldsymbol{\beta}^T\mathbf{W}_p^{-1}\boldsymbol{\beta}\} \times \lambda^{c-1} \exp\{-d\lambda\} \times \psi^{a-1} \exp\{-b\psi\} \times \pi^{e-1}(1 - \pi)^{f-1}.$$

Although a complex model from the prospective of a large number of parameters, the associated Gibbs sampler step does not involve overwhelming difficulties. The fully conditional posterior distributions can be found in Appendix D.

4.4 The Fixed Effects Model, Subset of Cohorts Associated

The fourth and final model which makes up the approach of our multiple cohorts analysis is the fixed effects model allowing for cohorts not to be associated. Hence, we take the model presented in the previous section and employ the fixed effects assumption replacing $\boldsymbol{\beta}_i$ with $\boldsymbol{\beta}_F$ and omitting the random effects assumption:

$$\mathbf{Y}_i|\mu_{F,i}, \boldsymbol{\beta}_F, \tau_{F,i}, h_i \sim N(\mu_i + h_i\mathbf{G}_i\boldsymbol{\beta}_F, \tau_{F,i}^{-1}\mathbf{I}_{N_i}), \qquad i = 1, ..., K \tag{4.32}$$

$$\mu_{F,i} \sim N(0, \nu^{-1}) \tag{4.33}$$

$$\boldsymbol{\beta}_F|\psi_F \sim N(\mathbf{0}, \psi_F^{-1}\mathbf{W}_p) \tag{4.34}$$

$$h_i \sim Bernoulli(\pi) \tag{4.35}$$

$$\tau_{F,i} \sim \Gamma(j_F, l_F) \tag{4.36}$$

$$\psi_F \sim \Gamma(a_F, b_F) \tag{4.37}$$

$$\pi \sim B(e, f). \tag{4.38}$$

The joint likelihood, then, takes the following form

$$P(\mathbf{Y}_1, ..., \mathbf{Y}_K|\mu_{F,1}, ..., \mu_{F,K}, \boldsymbol{\beta}_F, \tau_{F,1}, ..., \tau_{F,K}, \psi_F, h_1, ..., h_K, \pi) = \tag{4.39}$$
$$\prod_{i\in S_1} P(\mathbf{Y}_i|\mu_{F,i}, \boldsymbol{\beta}_F, \tau_{F,i}, \psi_F, h_i, \pi) \times \prod_{i\in S_0} P(\mathbf{Y}_i|\mu_i, \tau_i, h_i, \pi)$$

while the joint prior distribution of the parameters is

$$P(\mu_{F,1}, ..., \mu_{F,K}, \boldsymbol{\beta}_F, \tau_{F,1}, ..., \tau_{F,K}, \psi_F, h_1, ..., h_K, \pi) = \tag{4.40}$$
$$\prod_{i\in S_1} P(\mu_{F,i})P(\tau_{F,i})P(h_i|\pi) \times \prod_{i\in S_0} P(\mu_{F,i})P(\tau_{F,i})P(h_i|\pi)$$
$$\times P(\boldsymbol{\beta}_F|\psi_F)P(\psi_F)P(\pi),$$

and finally the joint posterior distribution

$$P(\mu_{F,1},...,\mu_{F,K},\boldsymbol{\beta}_F,\tau_{F,1},...,\tau_{F,K},\psi_F,h_1,...,h_K,\pi|\mathbf{Y}_1,...,\mathbf{Y}_K) \propto \prod_{i\in S_1} P(\mathbf{Y}_i|\mu_{F,i},\boldsymbol{\beta}_F, \tag{4.41}$$

$$\tau_{F,i},\psi_F,h_i=1,\pi)P(\mu_i)P(\tau_{F,i})P(h_i|\pi)$$

$$\times \prod_{i\in S_0} P(\mathbf{Y}_i|\mu_{F,i},\tau_{F,i},h_i,\pi)P(\mu_{F,i})P(\tau_{F,i})P(h_i|\pi) \times P(\boldsymbol{\beta}_F|\psi_F)P(\psi_F)P(\pi) \propto$$

$$\prod_{i\in S_1} \Big(\tau_{F,i}^{(N_i/2)} \exp\{-\frac{\tau_{F,i}}{2}(\mathbf{Y}_i-\mu_{F,i}\mathbf{1}_{N_i}-\mathbf{G}_i\boldsymbol{\beta}_F)^T(\mathbf{Y}_i-\mu_{F,i}\mathbf{1}_{N_i}-\mathbf{G}_i\boldsymbol{\beta}_F)\}$$

$$\times\nu^{1/2}\exp\{-\frac{\nu}{2}\mu_{F,i}^2\} \times \tau_{F,i}^{j_F-1}\exp\{-l_F\tau_{F,i}\} \times \pi\Big)$$

$$\times \prod_{i\in S_0} \Big(\tau_{F,i}^{(N_i/2)} \exp\{-\frac{\tau_{F,i}}{2}(\mathbf{Y}_i-\mu_i\mathbf{1}_{N_i})^T(\mathbf{Y}_i-\mu_i\mathbf{1}_{N_i})\}$$

$$\times\nu^{1/2}\exp\{-\frac{\nu}{2}\mu_{F,i}^2\}\tau_{F,i}^{j_F-1}\exp\{-l_F\tau_{F,i}\} \times (1-\pi)\Big)$$

$$\times\psi^{p/2}\exp\{-\frac{\psi_F}{2}\boldsymbol{\beta}_F^T\mathbf{W}_p^{-1}\boldsymbol{\beta}_F\} \times \psi_F^{a_F-1}\exp\{-b_F\psi_F\} \times \pi^{e-1}(1-\pi)^{f-1}.$$

This model resembles the *Binary Effects Model* of Han and Eskin in [35] which is implemented in a Frequentist framework. In the following chapter where we assess the performance of our models wee compare the two under different circumstances. Appendix D describes the fully conditional posterior distributions which compose the Gibbs sampling algorithm.

4.5 The Marginal Likelihood from the Gibbs Output

The essential quantity to perform model comparison, as we have previously seen, is the marginal likelihood. The crucial difference we face in the current setting is that, unlike in the analysis of a single cohort, we do not obtain a closed form of the posterior distributions of the four models, only samples from those through the use of the Gibbs sampling algorithm. This complicates the calculation of the marginal likelihood and urges us to employ approaches which approximate the marginal likelihood from the Gibbs output.

We have investigated two such approaches, which we refer to as the "Chib's method" described in [12] and the "Pritchard's method" described in [67]. The approach developed by Chib serves the setting where the Gibbs sampling algorithm has been used to provide a sample of draws from the posterior distribution and makes use of the fact that the marginal likelihood is the normalizing constant of the posterior density. Pritchard, Stephens and Donnelly, on the other hand, use an approach which is based on the sampled distribution of the deviance. We proceed by describing the calculation of the marginal likelihood under the two approaches using general notation explained in the respective subsections.

4.5.1 Chib Approach

The approach developed by Chib in [12] is suitable in the setting of a Gibbs sampling algorithm where all integrating constants of the conditional distributions are known. This requirement is usually met when conjugate priors are used in the model. For a given model, let the likelihood function and prior density be $f(\boldsymbol{y}|\theta)$ and $\pi(\boldsymbol{\theta})$ respectively. Let $\{\boldsymbol{\theta}^{(g)}\}_{g=1}^{G}$ be the output of the Gibbs procedure, where G is the number of iterations (an initial number of these can be discarded as the burn-in). The approach is based on two ideas. First, $m(\boldsymbol{y})$, by virtue of being the normalizing constant of the posterior density, can be expressed as:

$$m(\boldsymbol{y}) = \frac{f(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta}|\boldsymbol{y})}, \tag{4.42}$$

where the numerator expressions are defined above and the denominator is the posterior density of $\boldsymbol{\theta}$. Second, for a given $\boldsymbol{\theta}$, which we denote as $\boldsymbol{\theta}^*$, the posterior quantity $\pi(\boldsymbol{\theta}^*|\boldsymbol{y})$ can be estimated by exploiting the information in the collection of complete conditional densities. Let the estimate of $\pi(\boldsymbol{\theta}^*|\boldsymbol{y})$ be denoted by $\hat{\pi}(\boldsymbol{\theta}^*|\boldsymbol{y})$ we get the following:

$$\log_{10} \hat{m}(\boldsymbol{y}) = \log_{10} f(\boldsymbol{y}|\boldsymbol{\theta}^*) + \log_{10} \pi(\boldsymbol{\theta}^*) - \log_{10} \hat{\pi}(\boldsymbol{\theta}^*|\boldsymbol{y}) \tag{4.43}$$

It is of convenience to obtain the marginal likelihood on the logarithmic scale with base 10. The evaluation of the log likelihood function and the prior can be done rather easily while the estimation of the posterior ordinate follows in a straight forward manner. The estimation error arises from the estimation of the posterior ordinate. The choice of the point $\boldsymbol{\theta}^*$ is not critical since expression (4.42) holds for any $\boldsymbol{\theta}$ but more efficiently, for a given number of posterior draws, the density is likely to be more accurately estimated at a high density point where more samples are available. A modal value such as the the posterior mode or the maximum likelihood estimate can be computed from the Gibbs output, at least approximately. Alternatively, one can make use of the posterior mean.

The proposed estimation method only requires the complete conditional densities and produces a simulation consistent estimate. Let the parameter vector $\boldsymbol{\theta}$ consist of B components so that $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, ..., \boldsymbol{\theta}_B)$. At point $\boldsymbol{\theta}^*$ we write the posterior density as

$$\pi(\boldsymbol{\theta}^*|\boldsymbol{y}) = \pi(\boldsymbol{\theta}_1^*|\boldsymbol{y}) \times \pi(\boldsymbol{\theta}_2^*|\boldsymbol{y}, \boldsymbol{\theta}_1^*) \times ... \times \pi(\boldsymbol{\theta}_B^*|\boldsymbol{y}, \boldsymbol{\theta}_1^*, ..., \boldsymbol{\theta}_{B-1}^*), \tag{4.44}$$

where

- $\pi(\boldsymbol{\theta}_1^*|\boldsymbol{y})$ is the marginal ordinate. This quantity can be estimated in an obvious manner via the ergodic average of the full conditional density with the posterior draws of

all the parameters but $\boldsymbol{\theta}_1$, for which we use its star value:

$$\hat{\pi}(\boldsymbol{\theta}_1^*|\boldsymbol{y}) = \frac{1}{G}\sum_{g=1}^{G}\pi(\boldsymbol{\theta}_1^*|\boldsymbol{y}, \boldsymbol{\theta}_2^{(g)}, \boldsymbol{\theta}_3^{(g)}, ..., \boldsymbol{\theta}_B^{(g)}). \tag{4.45}$$

- the typical term $\pi(\boldsymbol{\theta}_r^*|\boldsymbol{y}, \boldsymbol{\theta}_1^*, ..., \boldsymbol{\theta}_{r-1}^*)$ is the reduced conditional ordinate.

The reduced conditional ordinate is expressed as follows:

$$\pi(\boldsymbol{\theta}_r^*|\boldsymbol{y}, \boldsymbol{\theta}_1^*, ..., \boldsymbol{\theta}_{r-1}^*) = \int \pi(\boldsymbol{\theta}_r^*|\boldsymbol{y}, \boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, ..., \boldsymbol{\theta}_{r-1}^*, \boldsymbol{\theta}_l(l > r)) \tag{4.46}$$
$$d\pi(\boldsymbol{\theta}_{r+1}, ..., \boldsymbol{\theta}_B|y, \boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, ..., \boldsymbol{\theta}_{r-1}^*).$$

To estimate this term we continue the Gibbs procedure, sampling from the complete conditional densities of $\{\boldsymbol{\theta}_r, \boldsymbol{\theta}_{r+1}, ..., \boldsymbol{\theta}_B\}$ where in each of these full conditionals, $\boldsymbol{\theta}_s$ is set equal to $\boldsymbol{\theta}_s^*$ for $s \leq r-1$. If we denote these draws as $\{\boldsymbol{\theta}_r^{(j)}, \boldsymbol{\theta}_{r+1}^{(j)}, ..., \boldsymbol{\theta}_B^{(j)}\}$ then an estimate of (4.46) is

$$\hat{\pi}(\boldsymbol{\theta}_r^*|\boldsymbol{y}, \boldsymbol{\theta}_1^*, ..., \boldsymbol{\theta}_{r-1}^*) = \frac{1}{G}\sum_{j=1}^{G}\pi(\boldsymbol{\theta}_r^*|\boldsymbol{y}, \boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, ..., \boldsymbol{\theta}_{r-1}^*, \boldsymbol{\theta}_l^{(j)}(l > r)). \tag{4.47}$$

An estimate of the joint density is $\prod_{r=1}^{B}\hat{\pi}(\boldsymbol{\theta}_r^*|\boldsymbol{y}, \boldsymbol{\theta}_s^{(j)}(s < r))$ whereas the estimate of the $\log_{10}$ marginal likelihood is

$$\log_{10}\hat{m}(\boldsymbol{y}) = \log_{10} f(\boldsymbol{y}|\boldsymbol{\theta}^*) + \log_{10}\pi(\boldsymbol{\theta}^*) - \sum_{r=1}^{B}\log_{10}\hat{\pi}(\boldsymbol{\theta}_r^*|\boldsymbol{y}, \boldsymbol{\theta}_s^{(j)}(s < r)). \tag{4.48}$$

Thus, this estimate of the $\log_{10}$ marginal likelihood is used to obtain for example the $\log_{10}$ Bayes factor of association. The estimate of the marginal likelihood can also be used to asses the convergence of the Gibbs sampler by computing and monitoring its stability for different iterations.

4.5.2 Pritchard Approach

Pritchard, Stephens and Donnelly present in [67] an approach which is based on the sampled distribution of the deviance. The simplest way of estimating the marginal likelihood uses the harmonic mean identity

$$f_m^{-1}(x) = E_\pi f_m(x|\theta)^{-1}, \tag{4.49}$$

where

$$\pi(\theta) = \frac{f_m(x|\theta)\pi_0(\theta)}{f_m(x)}.$$

If we then have samples θ_j from the posterior distribution $\pi(\theta)$, the sample means of the inverse likelihood function $f_m(x|\theta_j)^{-1}$ will converge almost surely to $f_m(x)^{-1}$. Hence we can re-write the harmonic mean estimate for the marginal likelihood as:

$$\hat{f}_m(x) \stackrel{def}{=} \left\{ \frac{1}{n} \sum_{j=1}^{n} f_m(x|\theta_j)^{-1} \right\}^{-1}. \tag{4.50}$$

We denote the Bayesian deviance as $D(\theta)$. Then,

$$D(\theta) = -2 \log f_m(x|\theta). \tag{4.51}$$

The conditional mean and variance of D given x are estimated by

$$E(D(\theta)|x) \approx \frac{1}{n} \sum_{j=1}^{n} -2 \log f_m(x|\theta^{(j)}) = \hat{\mu}; \tag{4.52}$$

$$Var(D(\theta)|x) \approx \frac{1}{n} \sum_{j=1}^{n} \left(-2 \log f_m(x|\theta^{(j)}) - \hat{\mu}\right)^2 = \hat{\sigma}^2. \tag{4.53}$$

If we make the hesitating assumption that the conditional distribution of D given x is normal, it follows from Equations 4.49 and 4.50 that

$$-2 \log \hat{f}_m(x) \approx \hat{\mu} + \frac{\hat{\sigma}^2}{4}. \tag{4.54}$$

An alternative interpretation of this method is that the model selection is based on penalizing the mean of the Bayesian deviance by a quarter of its variance. The samples θ_j from the posterior distribution $\pi(\theta)$ result from the implementation of the Gibbs sampler.

4.6 Assessment Measures

In order to perform inference for our meta-analysis model we consider several quantities which are used to answer different questions. In the interest of investigating which model fits the data better we calculate the posterior distribution of each model given the data. A priori we do not have any indication as to which model might be favoured therefore we use a uniform prior on the four models. Then, we write the posterior probability for model k, M_k, as follows:

$$P(M_k|Data) = \frac{P(Data|M_k)P(M_k)}{\sum_{k=1}^{4} P(Data|M_k)P(M_k)}, \tag{4.55}$$

where $P(M_k) = \frac{1}{4}$ for $k = 1, ..., 4$ and $P(Data|M_k)$ is the marginal likelihood for model k which can be obtained either using Chib's or Pritchard's methods.

Furthermore, we can obtain the posterior probability that the effect exists in each study which is readily available from the Gibbs output. This information is conentrated in the parameter h_i, $i = 1, ..., K$ hence we can use the estimate of the marginal posterior distribution $P(h_i|Data)$ for all i to provide evidence for the association or non-association of the cohort with the phenotype. This can be thought of as the analogous statitic of the *m-value* presented by Han and Eskin in [35].

We can also measure the strength of evidence for association via the Bayes factor for each SNP where we compare the marginal likelihood under each of the four models with the marginal likelihood of the null model of no association:

$$BF_{ass,k} = \frac{P(Data|M_k)}{P(Data|M_{NULL})}, \tag{4.56}$$

where we denote $P(Data|M_{NULL})$ as the marginal likelihood under the null model.

The Null Model

We define the null model of no association for the joint cohort analysis as follows

$$\mathbf{Y}_i|\tau_i \sim N(\mu_i, \tau_i^{-1}\mathbf{I}_{N_i}), \qquad i = 1, ..., K \tag{4.57}$$

$$\mu_i \sim N(0, \nu^{-1}) \tag{4.58}$$

$$\tau_i \sim \Gamma(j, l). \tag{4.59}$$

The joint likelihood and prior distribution are written as:

$$P(\mathbf{Y}_1, ..., \mathbf{Y}_K|\mu_1, ..., \mu_K, \tau_1, ..., \tau_K) = \prod_{i=1}^{K} P(\mathbf{Y}_i|\mu_i, \tau_i), \tag{4.60}$$

$$P(\mu_1, ..., \mu_K, \tau_1, ..., \tau_K) = \prod_{i=1}^{K} P(\mu_i)P(\tau_i). \tag{4.61}$$

the joint posterior is then formulated as:

$$P(\mu_1, ..., \mu_K, \tau_1, ..., \tau_K|\mathbf{Y}_1, ..., \mathbf{Y}_K) \propto \prod_{i=1}^{K} P(\mathbf{Y}_i|\mu_i, \tau_i)P(\mu_i)P(\tau_i) \tag{4.62}$$

$$= \frac{l^j \nu^{\frac{1}{2}}}{(2\pi)^{\frac{N_i}{2}+\frac{1}{2}}\Gamma(j)}(\tau_i)^{(N_i/2)+j-1} \exp\{-\frac{\tau_i}{2}(\mathbf{Y}_i - \mu_i\mathbf{1}_{N_i})^T(\mathbf{Y}_i - \mu_i\mathbf{1}_{N_i}) - \frac{\nu}{2}\mu_i^2 - l\tau_i\}.$$

In order to obtain the marginal likelihood of this model we need to evaluate the following integral if we were to proceed analytically.

$$P(Data|M_{NULL}) = \prod_{i=1}^{K} \int \int P(\mu_i, \tau_i|\mathbf{Y}_i)d\mu_i d\tau_i. \tag{4.63}$$

Alternativelly we set up a Gibbs sampler to obtain samples from the joint posterior distribution of the parameters given the data and then use Chib's approach to evaluate the desired

marginal likelihood. The full posterior conditionals are as follows:

$$P(\mu_i|\tau_i, \mathbf{Y}_i) = N(\frac{\tau_i \sum_{N_i} \mathbf{Y}_i}{\nu + N_i\tau_i}, \frac{1}{\nu + N_i\tau_i}); \tag{4.64}$$

$$P(\tau_i|\mu_i, \mathbf{Y}_i) = \Gamma(j_i^*, l_i^*), \tag{4.65}$$

where

$$j_i^* = \frac{N_i}{2} + j \quad \text{and} \quad l_i^* = \frac{(\mathbf{Y}_i - \mu_i \mathbf{1}_{N_i})^T(\mathbf{Y}_i - \mu_i \mathbf{1}_{N_i})}{2} + l.$$

4.7 Summary

Over the course of this chapter we have presented a meta analysis method which has been designed to fit quantitative data available on a number of cohorts. Apart from the classical approach to meta analysis which assumes similar effects between studies and fits a fixed effects model we additionally implement the random effects model which accounts for heterogeneity across studies, both at low and high level. The supplementary information we incorporate in the standard fixed and random effects model is to allow for a single or more cohorts not to be associated with the phenotype creating a more general framework which does not "force" every cohort to exhibit a non-zero genetic effect at the variant under examination. Due to the complexity and high-dimensionality of the models we use the Gibbs sampling algorithm to obtain samples from the marginal posterior distributions of the parameters which later assist us in the calculation of the marginal likelihood. To assess the power of the method to detect the true underlying model as well as its strength to detect phenotypic association and how this relates to existing methods, we perform an extensive simulation study which examines a variety of scenarios presented in the following chapter.

Chapter 5

Simulation Study and Analysis of Real Samples

In Chapter 4 we presented an intelligible approach to jointly analyze multiple cohorts, handling both homogeneity and heterogeneity between the effect sizes of the examined variants across cohorts. In this chapter we assess the performance of our model for different values of the mean of the random effects distribution, for different levels of heterogeneity and different number of associated cohorts. We compare its power to detect association and to detect an effect in each cohort with the power of existing models. Furthermore, we examine the convergence of the Gibbs sampling algorithm via trace plots and the corrected potential scale reduction factor. Finally, the method is applied to a set of real samples consisting of 11 cohorts and a total of 21,952 individuals of European descent where phenotypes concern smoking behaviour such as cigarettes-per-day.

5.1 Generation of Simulated Data

In order to assess the performance of our method, we construct sets of simulated data each with different characteristics. We make use of the haplotype data generated on chromosome

7 for the purpose of examining the performance of our fine mapping model in Chapter 3, hence, we have 1,000 datasets consisting of 1,000 individuals each in hand. Phenotypes are simulated under each of the four meta-analysis models:

- $\mathbf{Y}_i = \mu_i + \mathbf{G}_i\boldsymbol{\beta}_i + \epsilon_i$, $\epsilon_i \sim N(0, \tau_i^{-1}\mathbf{I}_{N_i})$, $i = 1, ..., K$;
- $\mathbf{Y}_i = \mu_{F,i} + \mathbf{G}_i\boldsymbol{\beta}_F + \epsilon_i$, $\epsilon_i \sim N(0, \tau_{F,i}^{-1}\mathbf{I}_{N_i})$;
- $\mathbf{Y}_i = \mu_i + h_i\mathbf{G}_i\boldsymbol{\beta}_i + \epsilon_i$, $\epsilon_i \sim N(0, \tau_i^{-1}\mathbf{I}_{N_i})$;
- $\mathbf{Y}_i = \mu_{F,i} + h_i\mathbf{G}_i\boldsymbol{\beta}_F + \epsilon_i$, $\epsilon_i \sim N(0, \tau_{F,i}^{-1}\mathbf{I}_{N_i})$.

We set the number of cohorts, K, equal to 10 and choose a SNP to be causal under the additive model. The cohort specific effect sizes $\boldsymbol{\beta}_i$, under the random effects models, follow a $N(\boldsymbol{\beta}, \lambda^{-1}\mathbf{V}_p)$. The number of cohorts for which $h_i = 1$ together with the values used for the other parameters are given in Table 5.1. The datasets are simulated under all combinations of the parameters, hence for example, for each value of $\boldsymbol{\beta}$ we explore five heterogeneity values with five different numbers of associated cohorts in each case. Additional to the parameters described in Table 5.1, we set τ_i and $\tau_{F,i}$ equal to 1 for $i = 1, ..., 10$ and ψ^{-1} equal to 0 while the phenotypes are standardised before the analysis. We let, $\mu = \mu_F = 0$, $\mathbf{V}_p$ and $\mathbf{W}_p$ equal to the identity while p equals 1 as the parameter vector consists of the additive effect at the chosen SNP, the prior precision of μ and μ_F, ν, is set equal to 0. 01 implying a highly variable prior distribution of μ. The different values of the heterogeneity parameter λ reflect different values of the standard deviation that the cohort specific effect sizes are drawn from. This varies from no variability amongst the effect sizes ($\lambda^{-1} = 0$) to very large variability ($\lambda = 5$). In the simulation study in [62] the between study standard deviation, τ, takes the values $\{0, 0.10, 0.20, 0.30\}$. Under the notation used in our simulations, these values correspond to $1/\lambda = 0$ and $\lambda = 100, 25, 11.1$. Addtionally, in [35] the heterogeneity is k times the effect size (logarithm of the relative risk). The effect size equals $\log(1.3)$ and $k = 0, 0.20, 0.40, 0.60, 0.80, 1$, which in terms of our heterogeneity parameter translates to $1/\lambda = 0$ and $\lambda = 363.2, 90.8, 40.4, 22.7, 14.5$. We choose similar values to adopt the same procedures. We also vary the number of associated cohorts aiming to focus on the models

Parameters	Values
Effect Size $\boldsymbol{\beta}$, $\boldsymbol{\beta}_F$ (Heritability h^2)	0.143(1%)
	0.101(0.5%)
	0.071(0.25%)
	0.045(0.1%)
Heterogeneity, λ^{-1}	0
	1/1000
	1/100
	1/15
	1/5
Associated Cohorts	10
	8
	6
	4
	2

Table 5.1: Parameter details for the simulation of phenotypes for the meta-analysis model.

which provide estimates on whether cohorts are associated or not. The mean of the random effects distribution $\boldsymbol{\beta}$ and the fixed effects parameter $\boldsymbol{\beta}_F$ are set to take different values which translate to different estimates of the heritability h^2. The following section gives a brief description on how these values are obtained.

5.1.1 Heritability Estimates

Heritability quantifies how much variation in the phenotype values is due to genetic variation and how much is due to external factors such as environmental causes. First, we define the additive genetic model as follows:

$$Y_i = \mu + \alpha G_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2). \tag{5.1}$$

We consider the mean of the phenotypic values $\boldsymbol{Y}$ under the above model, when the alleles present are a (major) and A (minor), as follows:

$$E(\boldsymbol{Y}|aa) = \mu;$$
$$E(\boldsymbol{Y}|Aa) = \mu + \alpha;$$
$$E(\boldsymbol{Y}|AA) = \mu + 2\alpha.$$

We also denote the frequency of the minor allele A as p and the frequency of the major allele A as q. Population parameters of interest are the population mean μ_{pop} and the genetic component of the variance, V_G. It is important to understand the partioning of the variation into components attributable to different causes. The total variation or phenotypic varation is the variance of the phenotype values. This is a sum of separate components such as the genotypic, the additive, the dominance, the interaction and the environmental components [23].

Assuming Hardy-Weinberg equilibrium we can calculate the population parameters. We have $G \in \{0, 1, 2\}$ with associated probabilities $\{q^2, 2pq, p^2\}$ and thus:

$$E(\boldsymbol{G}) = 0 \times q^2 + 1 \times 2pq + 2 \times p^2 = 2pq + 2p^2 = 2p;$$
$$E(\boldsymbol{G}^2) = 0^2 \times q^2 + 1^2 \times 2pq + 2^2 \times p^2 = 2p(q + 2p) = 2p(1 + p);$$
$$Var(\boldsymbol{G}) = E(\boldsymbol{G}^2) - E(\boldsymbol{G})^2 = 2p(1 + p) - (2p)^2 = 2p(1 - p).$$

We can write μ_{pop} as:

$$\begin{aligned} \mu_{pop} &= E(\boldsymbol{Y}) \\ &= E(\mu + \alpha\boldsymbol{G} + \boldsymbol{\epsilon}) \\ &= E(\mu) + \alpha E(\boldsymbol{G}) + E(\boldsymbol{\epsilon}) \\ &= \mu + 2p\alpha. \end{aligned} \tag{5.2}$$

Then the derivation of V_G is as follows:

$$\begin{aligned} V_G &= Var(\boldsymbol{Y}_G) \\ &= Var(\alpha \boldsymbol{G}) \\ &= \alpha^2 Var(\boldsymbol{G}) \\ &= 2p(1-p)\alpha^2. \end{aligned}$$

The overall phenotypic variation, V_P can be written as:

$$\begin{aligned} V_P &= Var(\alpha \boldsymbol{G} + \boldsymbol{\epsilon}) \\ &= \alpha^2 Var(\boldsymbol{G}) + Var(\boldsymbol{\epsilon}) + 2Cov(\alpha \boldsymbol{G}, \boldsymbol{\epsilon}). \end{aligned}$$

Since by definition $Cov(\boldsymbol{G}, \boldsymbol{\epsilon}) = 0$,

$$V_P = 2p(1-p)\alpha^2 + \sigma^2. \tag{5.3}$$

In this case the two types of heritability presented in Falconer & Mackay, $\frac{V_G}{V_P}$ and $\frac{V_A}{V_P}$ are equal since there is no dominance effect in our meta analysis models. Therefore, the heritability term H_A is:

$$\begin{aligned} H_A &= \frac{V_G}{V_P} \\ &= \frac{2p(1-p)\alpha^2}{2p(1-p)\alpha^2 + \sigma^2}. \end{aligned} \tag{5.4}$$

Now, for a given value of the heritability we can solve for α:

$$\alpha = \pm \sqrt{\frac{H_A \sigma^2}{2p(1-p)(1-H_A)}}. \tag{5.5}$$

We set σ^2 equal to 1 while p at the chosen SNP is approximately 0.5 across the 1,000 datasets examined. We want to choose values for α such that the value of H_A is 0.1%, 0.25%,0.5%,1% respectively. The resulting values are reported in Table 5.1.

5.1.2 The Prior Distributions

It is natural to model a meta analysis problem hierarchically, where observable outcomes are modeled conditionally on certain parameters, which themselves are given a probabilistic specification in terms of further parameters, known as *hyperparameters* [29]. The choice of values of the hyperparameters is important, as a key concern in a Bayesian framework is the dependence of the posterior distribution on one's prior. We wish to choose prior distributions to reflect what we expect to see in real studies. Table 5.2 lists the hyperparameters which we need to consider along with their assigned values. These are chosen so that the prior distributions cover the spectrum of values of the respective parameters chosen in order to conduct a realistic simulation study. For example, the values of a, a_F, b, b_F are chosen such that the prior variance of $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_F$ is as large to include the chosen values to examine. The same applies to c and d. The chosen values that mark the heterogeneity levels cover both extremes of little or no heterogeneity and large heterogeneity therefore, the imposed prior on λ should satisfy this condition. Thus the prior mean and prior standard deviation of λ are 500 and 250 respectively, given the values of c and d. The prior distributions of τ and τ_F have mean 1 and standard deviation 3.16 which complies with the fact that the phenotypes are standardised prior to the analysis. Finally, we use a uniform prior on π so that no specific values are favoured on the probability of association of a cohort. Figure 5.1 shows the prior distributions for the parameters based on the hyperparameters given in Table 5.2.

Hyperparameter	Property	Value
a	Shape of the Gamma prior on ψ	1000
a_F	Shape of the Gamma prior on ψ_F	1000
b	Rate of the Gamma prior on ψ	30
b_F	Rate of the Gamma prior on ψ_F	30
c	Shape of the Gamma prior on λ	4
d	Rate of the Gamma prior on λ	0.008
j	Shape of the Gamma prior on τ_i, $i = 1, ..., 10$	0.1
j_F	Shape of the Gamma prior on $\tau_{F,i}$, $i = 1, ..., 10$	0.1
l	Rate of the Gamma prior on τ_i, $i = 1, ..., 10$	0.1
l_F	Rate of the Gamma prior on $\tau_{F,i}$, $i = 1, ..., 10$	0.1
e	Shape 1 of the Beta prior on π	1
f	Shape 2 of the Beta prior on π	1

Table 5.2: List of the hyperparameters used in our meta analysis approach with the values assigned to them in the simulation study and analysis of the real samples.

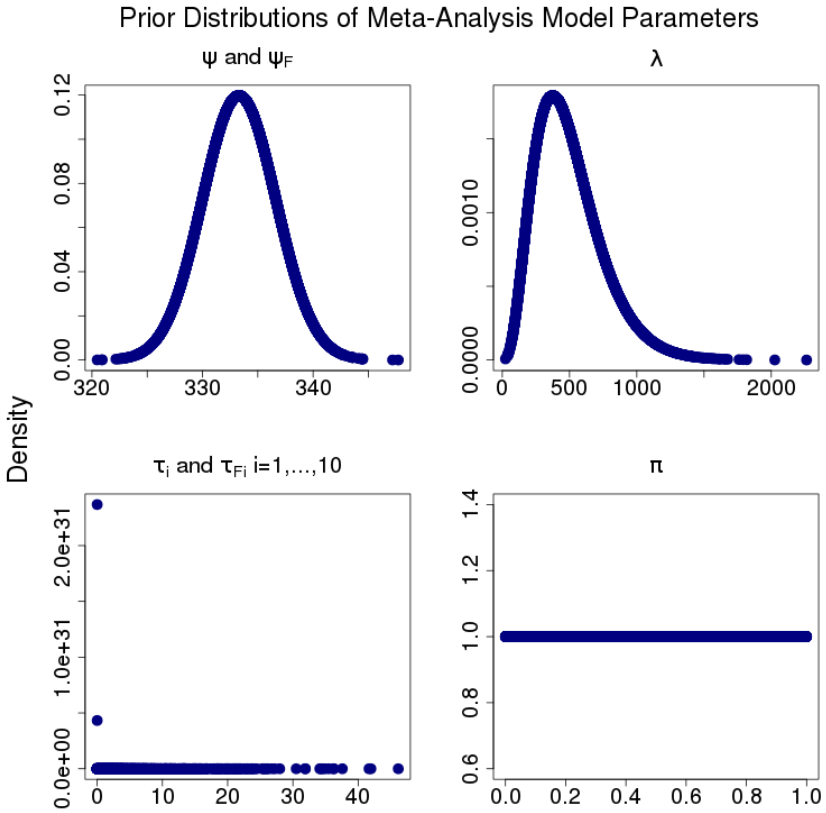

Figure 5.1: Prior distributions of the model parameters ψ, ψ_F, λ, τ_i, $\tau_{F,i}$ $i = 1, ..., 10$ and π used in the simulation study. The parameters ψ and ψ_F have the same prior distribution so do τ_i and $\tau_{F,i}$ for all i.

5.2 Probability to Detect the True Model

We set the number of total iterations to 5,000 and the burn-in equal to 2,000 as it is advisable to discard the first half of the chain [29]. The extra iterations required for any additional Gibbs chains within the calculation of the marginal likelihood via Chib's method, is set to 1,000. For each dataset we obtain the posterior probability for each model using both methods for the calculation of the marginal likelihood. In order to assess the ability of our meta-analysis approach to detect the correct model, we calculate the power in two ways. On the one hand we calculate the probability for each model via obtaining its percentage of exhibitng the highest posterior probability while on the other hand, we average the posterior probabilities for each model over the 1,000 datasets. These two techniques lead to the same conclusions for all $\boldsymbol{\beta}$ and λ therefore, we only present the results obtained using the highest posterior probability approach. The reason for which we obtain analogous results can be explained by an example. If we assume, for a given model, that its power under the highest posterior probability approach is large then this would imply that the majority of datasets exhibit, the largest posterior probability for that model. Then, given this power estimate and the exhibited certainty that the given model is the true model, it would lead to a large average of the posterior probabilities of the given model across the datasets.

Table 5.3 summarizes the probability of each model under all combinations of the different values of $\boldsymbol{\beta}$, λ and the number of associated cohorts. The values of $\boldsymbol{\beta}$ vary along the columns of the table, while each row describes a heterogeneity value and the number of associated cohorts that the data is simulated under. We highlight the probability of detecting the correct model in red while the highest probability is highlighted in blue. On a general note, the true model is identified when all cohorts are associated, simulated under the random and fixed effects models, for all levels of heterogeneity. We, also, observe that, when heterogeneity or homogeneity is present in the data, it is generally picked up since large probability is jointly attributed to the random or fixed effects models respectively.

When all cohorts have an effect and heterogeneity is small ($\lambda = 1000$) the probability to detect the random effects model, all cohorts associated, decreases as the mean of the random effects distribution becomes smaller. Such a level of heterogeneity can be also be considered as if the data have arisen from the fixed effects model since each $\boldsymbol{\beta}_i$ is drawn from a $N(\boldsymbol{\beta}, 0.032^2)$ distribution for any $\boldsymbol{\beta}$. This explains the relatively large probability for the fixed effects model with all cohorts associated. Undoubtedly, the probability estimate reduces for effect size values close to zero as it becomes more difficult to detect them. On the other hand, the power of the random effects model, subset of cohorts associated increases but is only the highest probability model for a small number of associated cohorts and $\boldsymbol{\beta} = 0.071$ and 0.045. If $\boldsymbol{\beta} = 0.143$ or 0.101 the fixed effects model, subset of cohorts associated is the highest probability model when four or two studies are associated with the phenotype.

As λ decreases we can clearly see a dominance of the random effects model, all cohorts associated when it is indeed the true model. In some cases when not all studies are associated the same model displays large probability although this is not entirely incorrect, the posterior distributions of effect sizes of the non-associated studies are simply centred around zero. We can also argue that when a study shows a small effect, near zero, the random effects model with subset of cohorts being associated will mistakenly classify it as non-associated hence in this case the random effects model all associated, will be more powerful. Furthermore, when only two out of ten cohorts are associated, we observe an elevation in probability as the heterogeneity increases of the general random effects model.

The last set of rows of Table 5.3 refers to the results when the simulated data are under the fixed effects models. When all cohorts are associated, the correct model displays the highest probability, although it decreases as $\boldsymbol{\beta}$ decreases. At 80% of the cohorts being associated, there is a preference for the random effects models: the all associated version has the largest probability for $\boldsymbol{\beta} = 0.143, 0.101, 0.071$ while for the $\boldsymbol{\beta} = 0.045$ the subset associated version is favoured. Similar results are obtained for six associated studies, although for

the largest value of $\boldsymbol{\beta}$ examined, the correct underlying model is given more evidence. This can be explained by the fact that when there is a mixture of association/non-assocation in a dataset under the fixed effects model, this creates heterogeneity in the sense that, under the random effects model(all cohorts associated), the effect sizes distribution has two modes, one at zero and one at the true effect size. At 40% of the cohorts being associated, the probability of the general random effects model has the tendency to increase as the simulated effect size decreases, while at 20% of associated studies we observe the same behaviour for $\boldsymbol{\beta} = 0.143, 0.101, 0.071$, while at $\boldsymbol{\beta} = 0.045$ the probability reduces, however the model still remains the most likely one.

We conclude that our meta-analysis approach confidently identifies the correct model when all cohorts are associated both under a random or a fixed effects scenario. Moreover, heterogeneity is also correctly detected when large probability is jointly attributed to both random effects models while homogeneity in the effect sizes when not all cohorts are associated is identified with conviction for larger effect sizes. The analogous results when the marginal likelihood is calculated under the Pritchard's method can be found in Appendix F.

		$\beta = 0.143$				$\beta = 0.101$				$\beta = 0.071$				$\beta = 0.045$			
		All		Subset		All		Subset		All		Subset		All		Subset	
	Ass.	Random	Fixed	Random	Fixed	Random	Fixed	Random	Fixed	Random	Fixed	Random	Fixed	Random	Fixed	Random	Fixed
$\lambda = 1000$	10	0.519	0.458	0.013	0.010	0.502	0.451	0.026	0.021	0.455	0.396	0.059	0.090	0.314	0.202	0.236	0.248
	8	0.642	0.017	0.052	0.289	0.669	0.135	0.062	0.134	0.524	0.180	0.129	0.167	0.602	0.004	0.196	0.198
	6	0.444	0.005	0.075	0.476	0.555	0.057	0.104	0.284	0.357	0.083	0.261	0.299	0.150	0.051	0.487	0.312
	4	0.293	0.001	0.152	0.554	0.317	0.031	0.228	0.424	0.170	0.026	0.427	0.377	0.113	0.059	0.449	0.379
	2	0.147	0.005	0.332	0.516	0.094	0.007	0.443	0.456	0.053	0.005	0.504	0.438	0.021	0.004	0.540	0.435
$\lambda = 100$	10	0.839	0.003	0.054	0.104	0.747	0.002	0.047	0.204	0.715	0.006	0.086	0.193	0.726	0.002	0.108	0.164
	8	0.637	0.002	0.086	0.275	0.614	0.005	0.096	0.285	0.567	0.004	0.137	0.292	0.602	0.004	0.196	0.198
	6	0.426	0.002	0.148	0.424	0.427	0.004	0.156	0.413	0.455	0.001	0.192	0.352	0.390	0.003	0.295	0.312
	4	0.216	0.003	0.295	0.486	0.251	0.003	0.313	0.433	0.256	0.004	0.314	0.426	0.263	0.005	0.353	0.379
	2	0.076	0.003	0.430	0.491	0.085	0.006	0.466	0.443	0.081	0.002	0.501	0.416	0.087	0.003	0.494	0.416
$\lambda = 15$	10	0.921	0.000	0.071	0.008	0.951	0	0.044	0.005	0.959	0	0.032	0.009	0.973	0	0.022	0.005
	8	0.833	0.000	0.148	0.019	0.869	0	0.103	0.028	0.891	0	0.085	0.024	0.913	0	0.068	0.019
	6	0.648	0.000	0.288	0.064	0.694	0	0.251	0.055	0.720	0	0.208	0.072	0.758	0	0.173	0.069
	4	0.419	0.000	0.441	0.140	0.447	0	0.429	0.124	0.470	0.001	0.408	0.121	0.503	0	0.355	0.142
	2	0.133	0.001	0.619	0.247	0.127	0.003	0.612	0.258	0.137	0	0.636	0.227	0.155	0	0.571	0.274
$\lambda = 5$	10	0.987	0	0.013	0	0.992	0	0.008	0	0.994	0	0.006	0	0.995	0	0.005	0
	8	0.948	0	0.051	0.001	0.952	0	0.046	0.002	0.962	0	0.037	0.001	0.973	0	0.026	0.001
	6	0.800	0	0.196	0.004	0.813	0	0.180	0.007	0.826	0	0.169	0.005	0.823	0	0.175	0.002
	4	0.478	0	0.491	0.031	0.49	0	0.48	0.03	0.480	0	0.471	0.049	0.505	0	0.467	0.028
	2	0.118	0.001	0.751	0.130	0.122	0.001	0.756	0.121	0.112	0.001	0.754	0.133	0.104	0	0.749	0.147
$\lambda^{-1} = 0$	10	0.221	0.766	0.013	0.000	0.223	0.730	0.044	0.003	0.229	0.681	0.034	0.056	0.122	0.347	0.244	0.287
	8	0.519	0.054	0.071	0.356	0.630	0.223	0.082	0.06	0.391	0.334	0.140	0.135	0.107	0.145	0.465	0.283
	6	0.444	0.003	0.080	0.473	0.593	0.072	0.127	0.208	0.340	0.133	0.298	0.229	0.079	0.046	0.556	0.319
	4	0.332	0.004	0.127	0.537	0.374	0.027	0.246	0.353	0.165	0.056	0.474	0.305	0.044	0.013	0.593	0.350
	2	0.152	0.004	0.291	0.553	0.102	0.008	0.498	0.392	0.031	0.004	0.599	0.366	0.014	0.006	0.526	0.454

Table 5.3: Probability to detect the correct model under which the data are simulated from for different values of β, λ and different number of associated cohorts (The correct model is highlighted in red while the model which exhibits the highest probability is highlighted in blue).

5.3 Power to Detect Association with the Phenotype

Although genome wide association studies have detected numerous loci associated with diseases, these loci only account for a small proportion of the genetic contribution to the disease. The remaining can be accounted for by loci with very small effect sizes requiring tens of thousands of samples to be identified. For this reason we calculate a meta-analysis Bayes factor for each of the four models $BF_{ass,k}$ for $k = 1, 2, 3, 4$ and subsequently obtain the power of each model to detect association. Furthermore, we compare the performance of our models to three of **METASOFT** statistics.

In order to obtain the power, at certain levels of significance, it is necessary to derive the empirical threshold values for each model using the empirical null distributions. We use 20,000 null datasets to construct the null distributions and extract the thresholds for different p-values and present the results for a p-value equal to 10^{-4}. On the same note, it is straightforward to compare the resulting **METASOFT** p-values simply by counting how many of these fall below 10^{-4}.

Figures 5.2 through to 5.6 present the power to detect association for our four meta analysis models and the three **METASOFT** statistics on the same axes for all heterogeneity values and $\boldsymbol{\beta}$. The horizontal axis displays the number of associated cohorts while the vertical axis shows the power. The colours which represent each method are as follows: random effects model, all cohorts associated in black, random effects model, subset of cohorts associated in red, fixed effects model, all cohorts associated in green, fixed effects model, subset of cohorts associated in blue, **METASOFT**'s fixed effects model in cyan, **METASOFT**'s random effects model in pink and **METASOFT**'s binary effects model in brown. It it worth noting that the results of **METASOFT**'s fixed effects model (**METASOFT FE**) are directly comparable to the results of the Bayesian fixed effects model with all cohorts associated, **METASOFT**'s random effects model (**METASOFT RE**) is analogous to the Bayesian random effects model with all cohorts and associated and lastly **METASOFT**'s binary effects

model (**METASOFT BE**) can be regarded as parallel to Bayesian fixed effects model with subset of cohorts associated.

Figure 5.2 presents the power for $\lambda = 1000$ (i.e. the standard deviation of the random effects distribution equals ~ 0.032) and all values of $\boldsymbol{\beta}$. As a general statement, the power to detect association with the phenotype decreases within each value of $\boldsymbol{\beta}$, for decreasing number of associated cohorts, as well as between values of $\boldsymbol{\beta}$. The random effects model (all cohorts associated) along with the fixed effects model (subset associated) outperform the others in all cases where we observe similar performance of the two with fixed effects model (subset associated) doing slightly better when $\boldsymbol{\beta} = 0.045$. When only two cohorts are associated, at $\boldsymbol{\beta} = 0.143$ and 0.101 the fixed effects model (subset associated) exhibits the highest power. The lines for the fixed effects model with all cohorts associated and the **METASOFT FE** are almost on top of each other with our Bayesian model performing slightly better. We also observe a drop in power for the more complicated random effects model as $\boldsymbol{\beta}$ decreases. The power loss is greatest when all studies are associated. This can be attributed to the small effect sizes and to the fact that the model does not have large enough power to correctly classify the studies as associated. On the other hand the random effects model (all associated) and **METASOFT RE** have larger power, although the latter does not perform as well as its analogous approach. The same conclusion applies for the two fixed effects model which account for possible non-association of studies. The power of the binary effects model does not surpass its Bayesian equivalent and the greatest difference is observed at two associated cohorts and $\boldsymbol{\beta} = 0.143$.

When the heterogeneity increases and $\lambda = 100$ (Figure 5.3), there is an obvious separation of the two fixed effects models, all cohorts associated form the rest of the models. As $\boldsymbol{\beta}$ decreases three groups of models are formed where the two fixed effects models, all cohorts associated exhibit the lowest power, then follow the binary effects model and the rest of the models form the group with similar power estimates. The random effects model with the assumption of all cohorts being associated outperforms the rest as $\boldsymbol{\beta}$ decreases while at two

associated cohorts the fixed effect model (subset of cohorts associated) comes first while the random effects (subset of cohorts associated) comes second best.

For large heterogeneity, $\lambda = 15, 5$, (Figure 5.4 and 5.5) we observe the same grouping of models as for $\lambda = 100$. **METASOFT FE** and its equivalent Bayesian model exhibit the lowest power as it is expected, with **METASOFT FE** doing slightly better. The second group consists of only one model, **METASOFT BE**. The rest of the models have very large power and approximately identical estimates as we can not distinguish between the different coloured lines. Only at the lowest number of associated cohorts the random and fixed effects models, subset associated perform marginally better.

Finally, the results for the data simulated under the fixed effects models are presented in Figure 5.6. It is worth noting here that the behaviour of the results matches significantly that of Figure 5.2 since the large value of λ essentially implies almost no heterogeneity. The power equals 1 or is very close to 1 when all or 80% of the cohorts are associated for all models and $\boldsymbol{\beta} = 0.143$ and 0.101. For smaller values of $\boldsymbol{\beta}$, the power decreases significantly, with the Bayesian random effects model, subset of cohorts associated, performing the worst. As the number of associated studies decreases, the power eventually reaches 0, especially at $\boldsymbol{\beta} = 0.045$ where even at four associated cohorts the power is also near zero.

The power is re-calculated when we obtain the marginal likelihood under Pritchard's method. Generally, when all or 80% of the studies are associated, the power estimates are approximately equal amongst the two approaches but at four and two associated cohorts the power given Pritchard's method is slightly higher while the power given Chib's method is higher for the general random effects model (subset of cohorts associated). Furthermore, the power under Pritchard's method is larger for the fixed effects model, all cohorts associated.

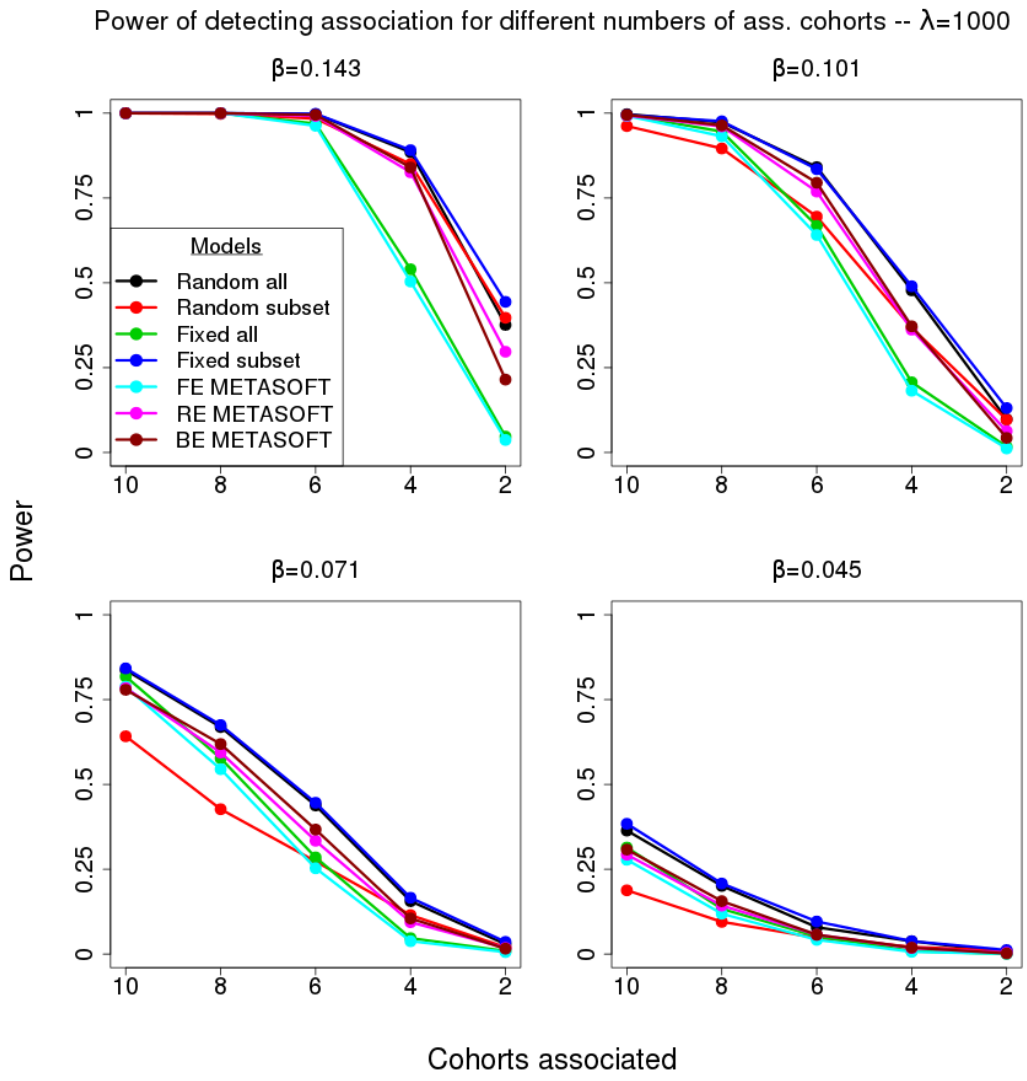

Figure 5.2: Power to detect assocation for the different meta analytic models examined at a significance level equal to 0.0001, all the values of β and $\lambda = 1000$. The horizontal axis displays the number of associated cohorts while the vertical axis shows the power. The colours which represent each method are as follows: random effects model, all cohorts associated in black, random effects model, subset of cohorts associated in red, fixed effects model, all cohorts associated in green, fixed effects model, subset of cohorts associated in blue,**METASOFT**'s fixed effects model in cyan, **METASOFT**'s random effects model in pink and **METASOFT**'s binary effects model in brown.

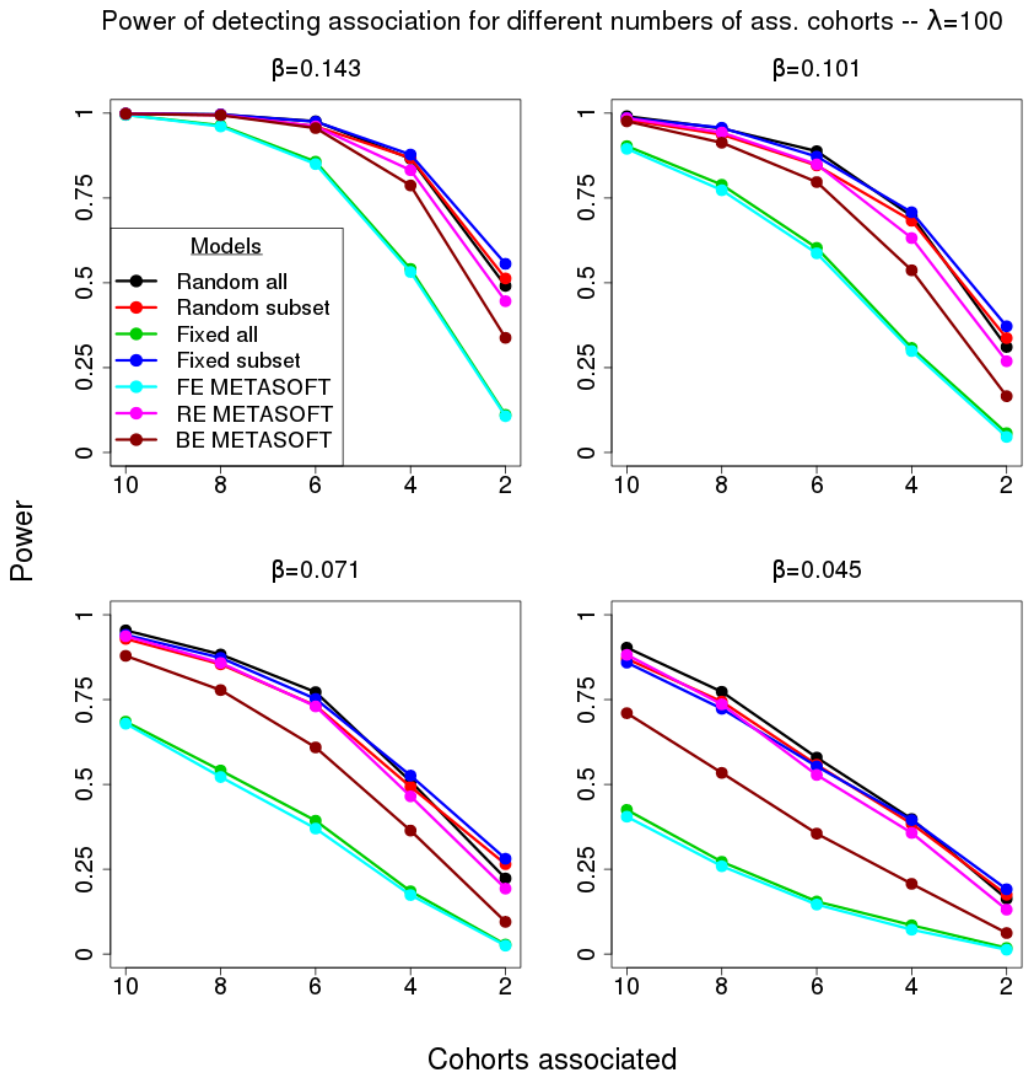

Figure 5.3: Power to detect assocation for the different meta analytic models examined at a significance level equal to 0.0001, all the values of β and $\lambda = 100$. The horizontal axis displays the number of associated cohorts while the vertical axis shows the power. The colours which represent each method are as follows: random effects model, all cohorts associated in black, random effects model, subset of cohorts associated in red, fixed effects model, all cohorts associated in green, fixed effects model, subset of cohorts associated in blue,**METASOFT**'s fixed effects model in cyan, **METASOFT**'s random effects model in pink and **METASOFT**'s binary effects model in brown.

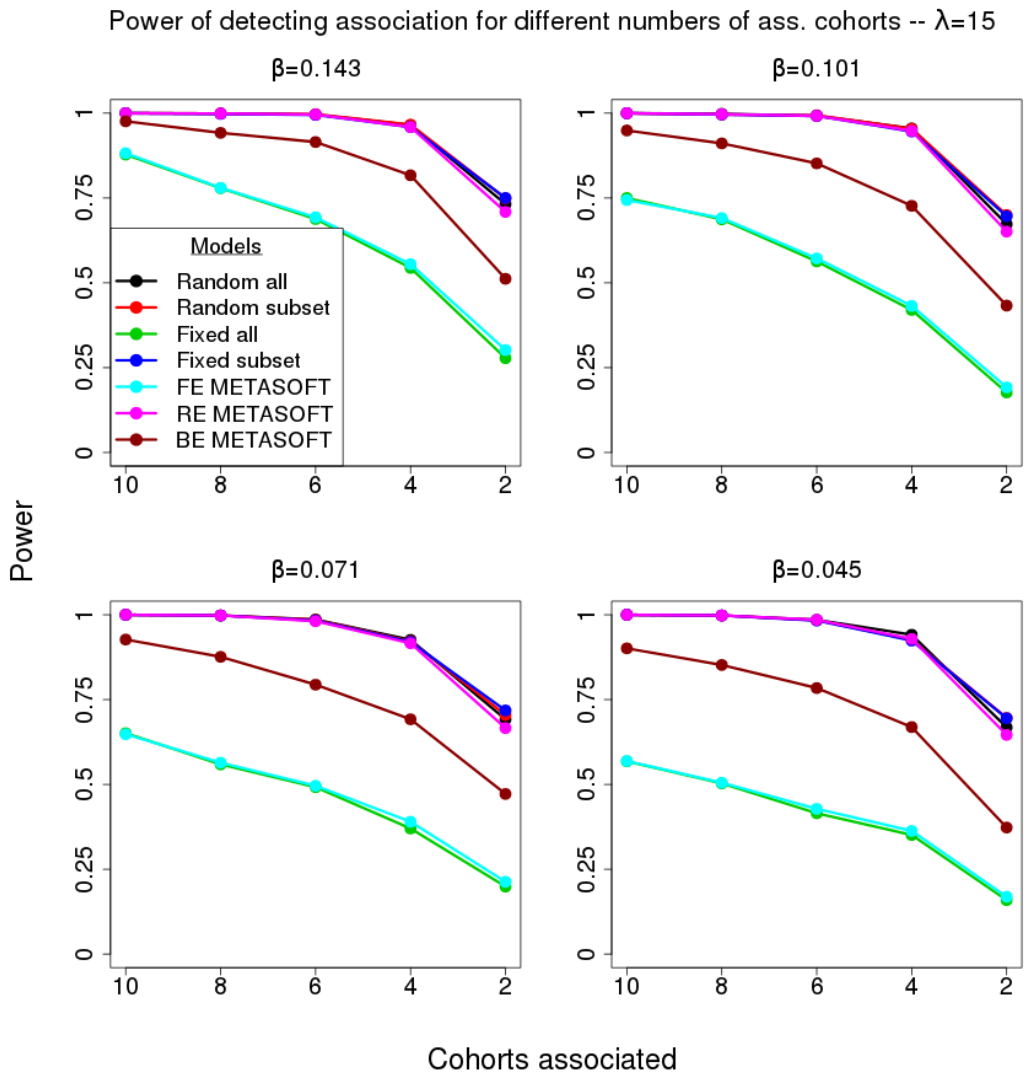

Figure 5.4: Power to detect assocation for the different meta analytic models examined at a significance level equal to 0.0001, all the values of β and $\lambda = 15$. The horizontal axis displays the number of associated cohorts while the vertical axis shows the power. The colours which represent each method are as follows: random effects model, all cohorts associated in black, random effects model, subset of cohorts associated in red, fixed effects model, all cohorts associated in green, fixed effects model, subset of cohorts associated in blue,**METASOFT**'s fixed effects model in cyan, **METASOFT**'s random effects model in pink and **METASOFT**'s binary effects model in brown.

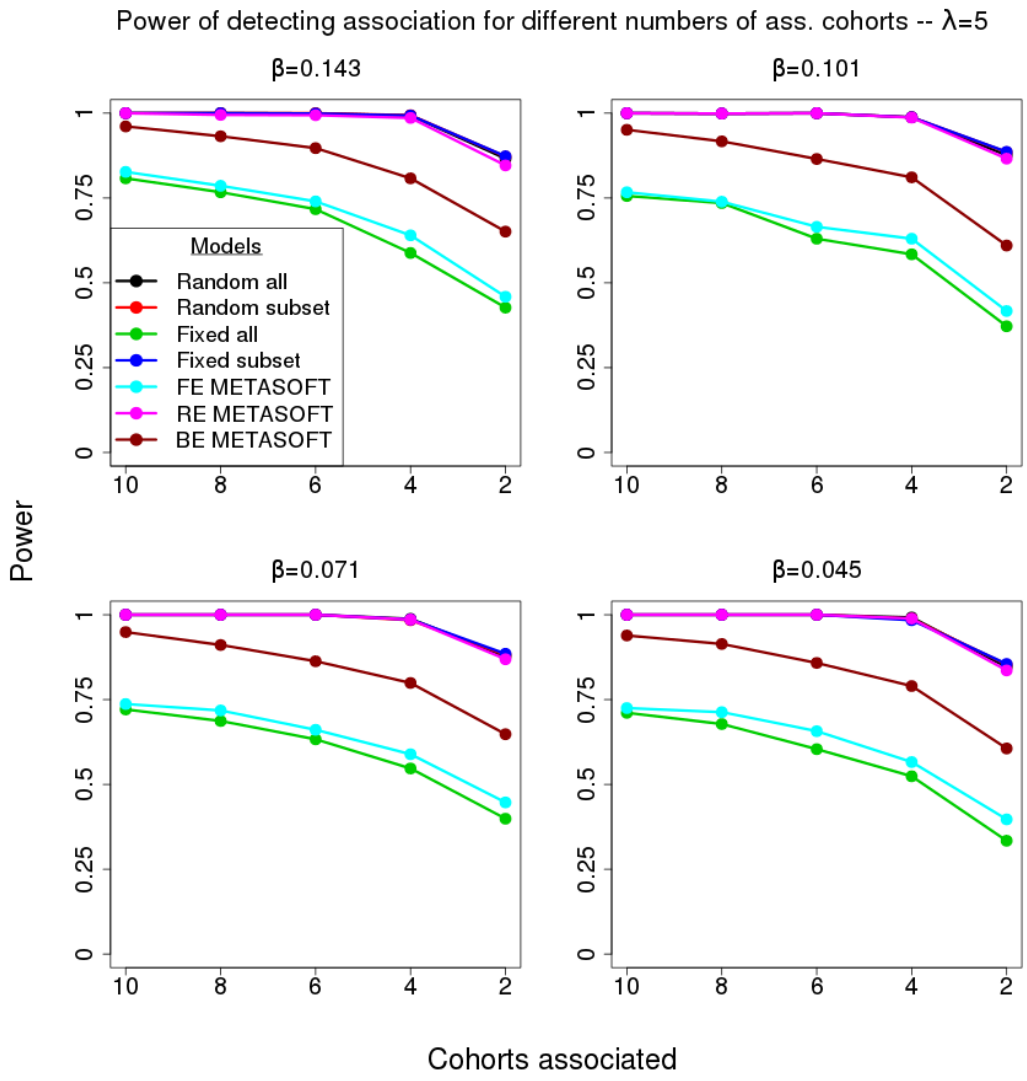

Figure 5.5: Power to detect assocation for the different meta analytic models examined at a significance level equal to 0.0001, all the values of β and $\lambda = 5$. The horizontal axis displays the number of associated cohorts while the vertical axis shows the power. The colours which represent each method are as follows: random effects model, all cohorts associated in black, random effects model, subset of cohorts associated in red, fixed effects model, all cohorts associated in green, fixed effects model, subset of cohorts associated in blue,**METASOFT**'s fixed effects model in cyan, **METASOFT**'s random effects model in pink and **METASOFT**'s binary effects model in brown.

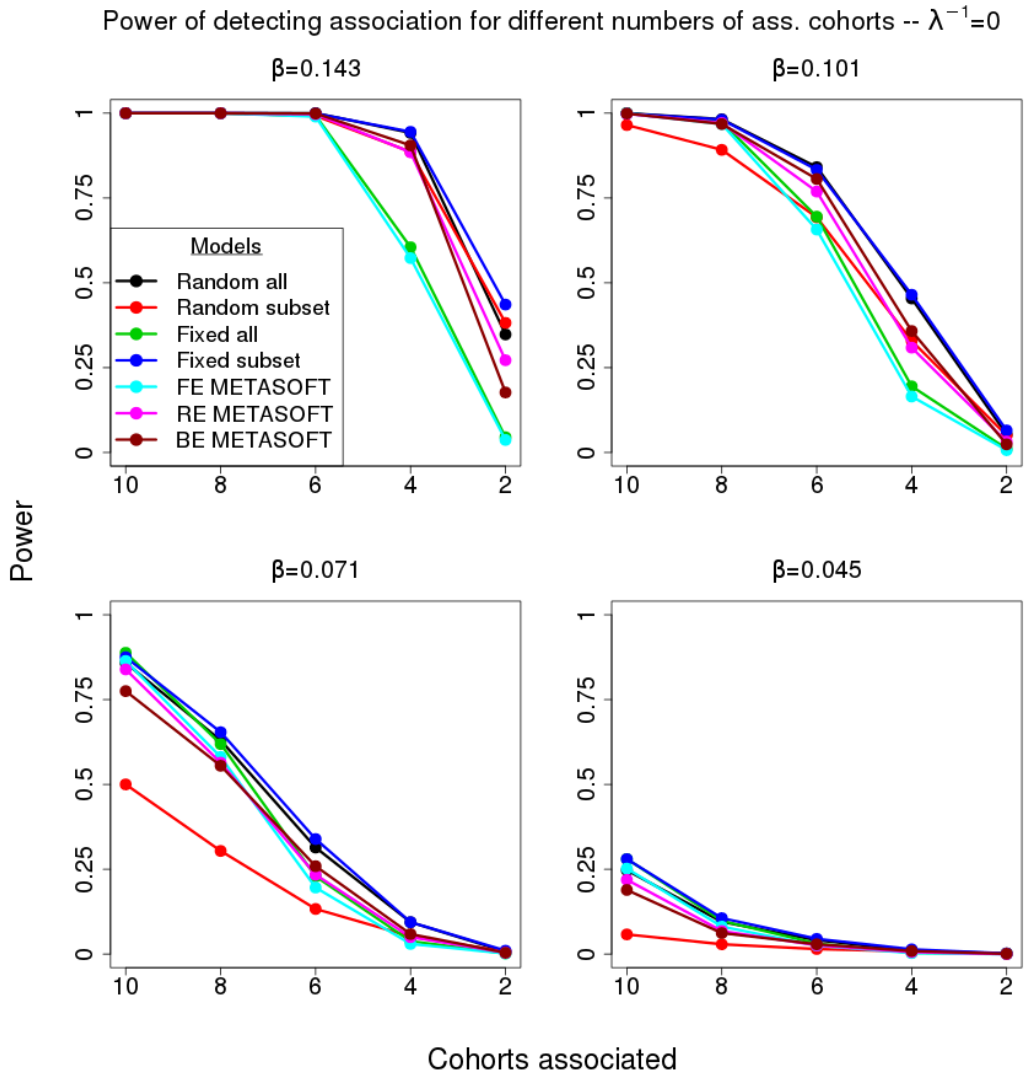

Figure 5.6: Power to detect assocation for the different meta analytic models examined at a significance level equal to 0.0001, all the values of β and $\lambda^{-1} = 0$. The horizontal axis displays the number of associated cohorts while the vertical axis shows the power. The colours which represent each method are as follows: random effects model, all cohorts associated in black, random effects model, subset of cohorts associated in red, fixed effects model, all cohorts associated in green, fixed effects model, subset of cohorts associated in blue,**METASOFT**'s fixed effects model in cyan, **METASOFT**'s random effects model in pink and **METASOFT**'s binary effects model in brown.

5.4 Power to Identify an Effect in Associated Cohorts

We now wish to examine an equally important statistic, the posterior probability of detecting an effect in each examined cohort as it emerges from the two models which account for possible non-association of cohorts. The posterior probability on an effect under the fixed effects model is analogous measure to the *m-value* of **METASOFT** introduced by Han and Eskin [35]. We present the results under $\beta = 0.143$ and three heterogeneity levels, $\lambda = 100, 15$ and $\lambda^{-1} = 0$ in Figures 5.7-5.9. The horizontal axes denote the cohorts while the vertical axes denote the power of detecting a posterior probability of an effect greater than 0.5 across the 1,000 simulated datasets.

Notably, although, the power to detect an effect in associated cohorts does not go over 80% when heterogeneity is present, the three estimates correctly detect the non-association of the relevant cohorts. As the number of associated studies decreases we observe an increase in the power under the random effects model while the power under the fixed effects model and the *m-value* is approximately equal when $\lambda = 100$ and $\lambda^{-1} = 0$. When the heterogeneity is larger, the power to detect an effect under the fixed effects model outperforms the other two estimates. When the data is simulated under no heterogeneity we expect the *m-value* to achieve the highest power since it is modelled to serve this purpose. Along the same lines, we expect equally confident results for our fixed effects model which is transparent in Figure 5.9.

A possible explanation as to why the power to detect an effect under the random effects model is not as high, especially under the presence of heterogeneity, is our choice of hyperparameters which in our attempt to model a broad range of heterogeneity values we are led to loss of power. If we were to use a prior distribution on λ which focusses on the true underlying value, we would have achieved larger power to detect an effect in associated cohorts. Furthermore, especially in the presence of large heterogeneity, effect sizes can be small, very close to zero and it is therefore acceptable that they are much harder to detect.

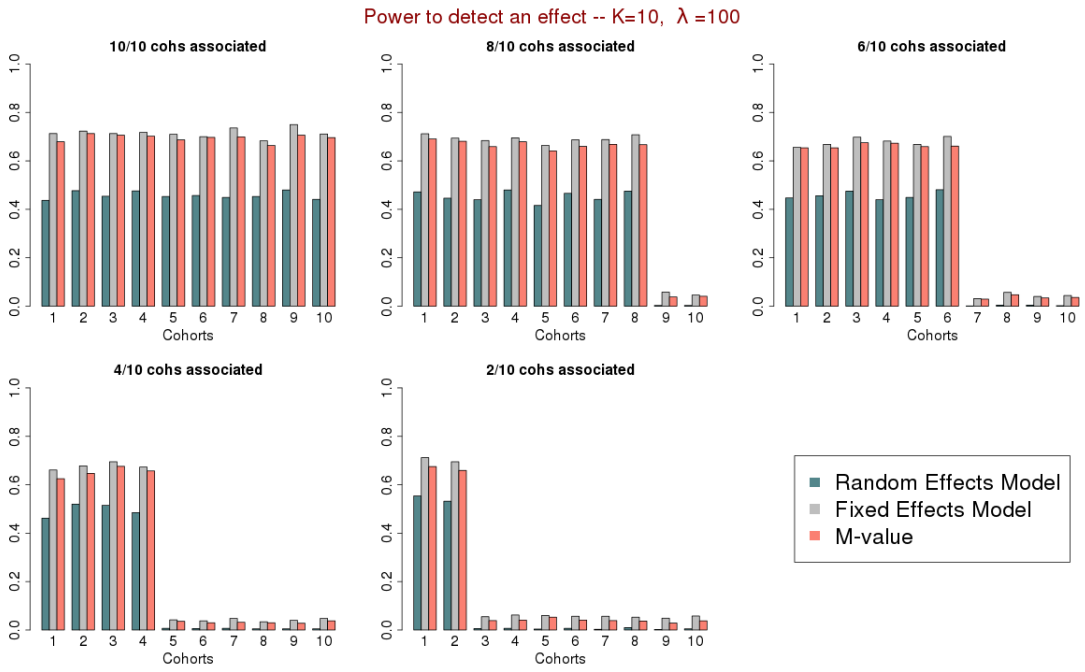

Figure 5.7: Power to detect and effect in each cohort, $\beta = 0.143$ and $\lambda = 100$. The horizontal axes denote the cohorts while the vertical axes denote the power. The posterior probability of detecting an effect under the random effects model is in green, the posterior probability of detecting an effect under the fixed effects model is in grey and the m-value is in red.

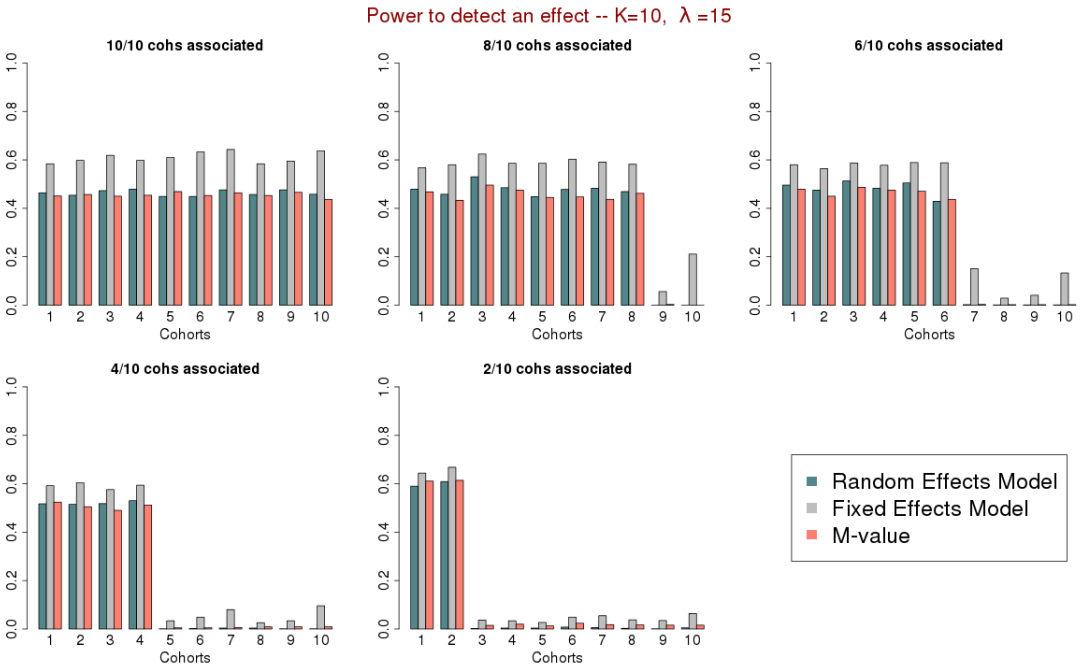

Figure 5.8: Power to detect and effect in each cohort, $\beta = 0.143$ and $\lambda = 15$. The posterior probability of detecting an effect under the random effects model is in green, the posterior probability of detecting an effect under the fixed effects model is in grey and the m-value is in red.

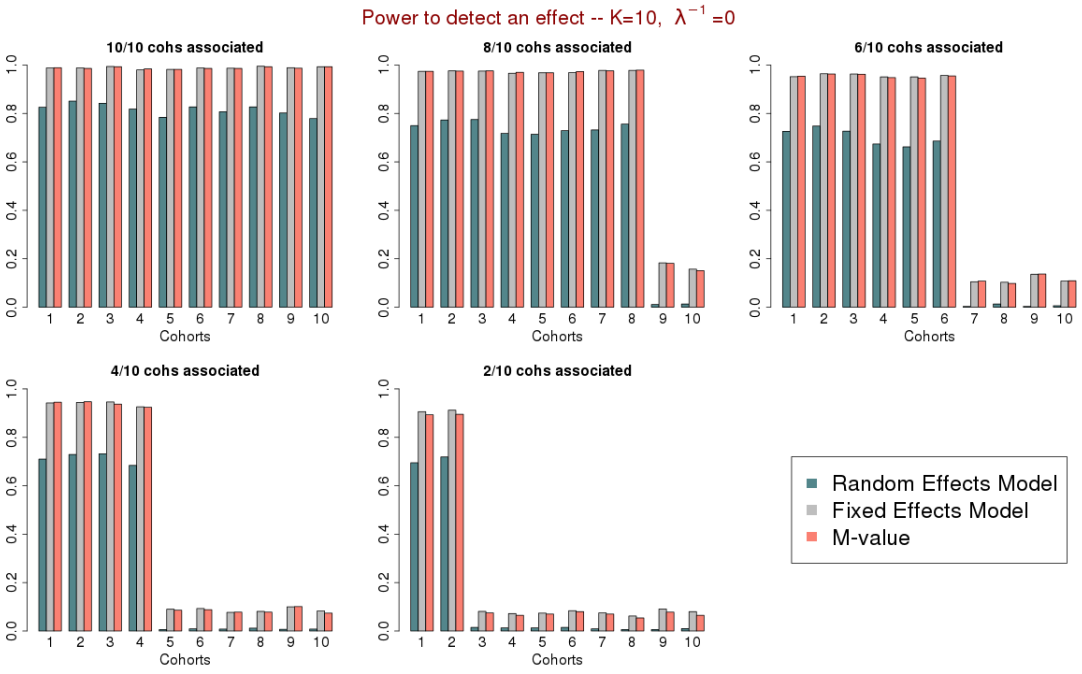

Figure 5.9: Power to detect and effect in each cohort, $\beta = 0.143$ and $\lambda^{-1} = 0$. The posterior probability of detecting an effect under the random effects model is in green, the posterior probability of detecting an effect under the fixed effects model is in grey and the m-value is in red.

5.5 Convergence Analysis

It is always the case, when using an iterative procedure such as a Gibbs sampler to obtain the joint posterior distribution of a collection of parameters $\boldsymbol{\theta}$, say, that certain issues need to be addressed. First, the chains have to be proceeded long enough so that the resulting distributions are representative of the target distributions. Second, the early iterations of a chain are influenced by the starting values of the components of the parameter vector even if the complete chain has reached approximate convergence. Several ways to handle these problems are discussed in [29]. These include monitoring convergence by simulating multiple sequences with starting points dispersed throughout the parameter space and by comparing variation between and within simulated sequences until "within" variation roughly equals "between" variation.

A recommended approach of assessing convergence is based on comparing different simulated sequences with different starting points. Each of the chains is essentially let to run enough in order to forget its starting position. To diminish the effect of the starting values of the parameters, it is, additionally, common to discard the first half of each sequence and focus on the second half, a practice well-known as *burn-in*. Further, each scalar estimand can be monitored when we have parallel sequences and compute the between- and within-sequence variances. Gelman and Rubin [30] along with Brooks and Gelman [11] present a quantitative measure to monitor convergence, named the potential scale reduction factor. This estimates the factor by which the scale of the distribution of a parameter might be reduced if the simulations were continued indefinitely.

To assess the convergence of the Gibbs sampler incorporated in our meta-analysis approach, we examine the posterior marginal distributions for the key parameters of each model. These include $\boldsymbol{\beta}$, the between cohorts variance for the random effects model λ, the variance of $\boldsymbol{\beta}$, ψ and π. We explore the trace plots of the marginal posterior distribution of $\boldsymbol{\beta}$ for chains with different starting points and calculate the potential scale reduction factor for

all the aforementioned parameters. To conclude, we also test the stability of Chib's method for the calculation of the marginal likelihood.

5.5.1 Graphical Analysis

The recommended approach to assessing convergence of iterative simulation is to compare different simulated sequences with overdispersed starting points. For this reason, we examine one dataset simulated under each of the four meta analysis models with the following characteristics: $\boldsymbol{\beta} = 0.143$ and $\lambda = 15$ for the random effects models and for the case of subset associated, we use a dataset with six out of ten cohorts being associated while under the fixed effects models $\boldsymbol{\beta}_F = 0.143$ and, also, six out of ten cohorts are associated when under the subset associated scenario.

Figure 5.10 displays the results of the different chains for $\boldsymbol{\beta}$. The examined starting values are $\{0, 1, -1, 0.5\}$ and the truth is $\boldsymbol{\beta} = \boldsymbol{\beta}_F = 0.143$ which is also shown on each pair of axis by a red horizontal line. Table 5.4 summarizes the posterior mean and standard deviation of each resulting distribution. Despite the distinct starting values, convergence is achieved early on for each chain and the running time of 5,000 iterations proves to be more than adequate for convergence to be achieved. The starting values are forgotten after a small number of iterations therefore dependence on the initial points of a chain does not add any burden in the analysis. The increased posterior standard deviation of the random effects models is explained by the large heterogeneity that the data are simulated under.

5.5.2 The Potential Scale Reduction Factor

Let us denote the parameter under examination with θ. First, we simulate $m >= 2$ chains, each of length $2n$ with starting points drawn from an overdispersed distribution. We discard the first n iterations and focus on the second half of each chain. We label the simulation draws as θ_{ij} where $i = 1, ..., n$ and $j = 1, ..., m$ and proceed with calculating the between-

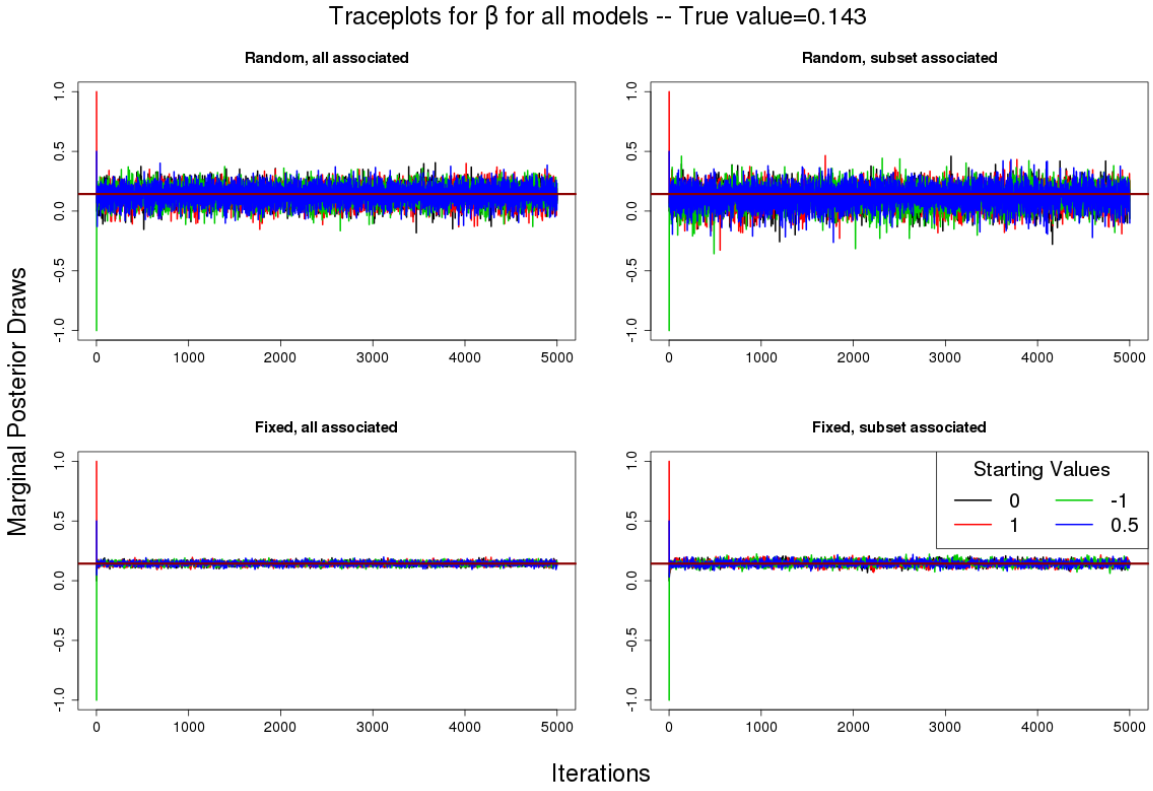

Figure 5.10: Traceplots of the marginal posterior distribution of $\boldsymbol{\beta}$ for all models for chains with overdispersed starting points. The horizontal axes represent the number of iterations performed for each chain and the vertical axes represent the values of $\boldsymbol{\beta}$ drawn from the Gibbs output. Chains with different starting values are coloured as follows: 0 (black), 1(red), -1 (green) and 0.5 (blue). The true value is 0.143 (horizontal line).

and within- sequence variances:

$$B = \frac{n}{m-1}\sum_{j=1}^{m}(\bar{\theta}_{.j} - \bar{\theta}_{..})^2, \qquad \bar{\theta}_{.j} = \frac{1}{n}\sum_{i=1}^{n}\theta_{ij}, \quad \bar{\theta}_{..} = \frac{1}{m}\sum_{j=1}^{m}\bar{\theta}_{.j} \tag{5.6}$$

$$W = \frac{1}{m}\sum_{j=1}^{m}s_j^2, \qquad s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\theta_{ij} - \bar{\theta}_{.j})^2.$$

We can, then, estimate the marginal posterior variance of θ, var$(\theta|y)$ by a weighted average of W and B:

$$\text{var}^{+}(\theta|y) = \hat{\sigma}_{+}^2 = \frac{n-1}{n}W + \frac{B}{n}, \tag{5.7}$$

which overestimates var$(\theta|y)$ assuming the starting distribution is appropriately overdis-

Starting Values	Mean	Standard Deviation
Random Effects Model, All Cohorts Associated		
0	0.138	0.070
1	0.134	0.071
-1	0.136	0.072
0.5	0.135	0.069
Fixed Effects Model, All Cohorts Associated		
0	0.144	0.014
1	0.143	0.018
-1	0.144	0.021
0.5	0.144	0.015
Random Effects Model, Subset of Cohorts Associated		
0	0.121	0.081
1	0.118	0.084
-1	0.118	0.087
0.5	0.121	0.084
Fixed Effects Model, Subset of Cohorts Associated		
0	0.146	0.020
1	0.145	0.023
-1	0.146	0.026
0.5	0.144	0.021

Table 5.4: Posterior mean and standard deviation of $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_F$ for the four meta analysis models for Gibbs run with different starting points.

persed but is unbiased for var$(\theta|y)$ under stationarity. Also, for a finite n, W should be smaller than the marginal posterior variance because the individual sequences have not explored the entire range of the target distribution and, as a result, W will have less variability (in the limit of $n \to \infty$, $E(W) \to$ var$(\theta|y)$).

To estimate what is known about θ we can improve the optimistic $N(\hat{\mu}, \hat{\sigma}^2)$ estimate of the target distibution, where $\hat{\mu} = \bar{\theta}_{..}$, by allowing for the sampling variability of $\hat{\mu}$ and $\hat{\sigma}^2$. This results in an approximate Student's t distribution for θ with centre $\hat{\mu}$, scale $\sqrt{\hat{V}} = \sqrt{\text{var}^+(\theta|y) + B/mn}$ and degrees of freedom $df = 2\hat{V}^2/\hat{\text{var}}(\hat{V})$, where

$$\hat{\text{var}}(\hat{V}) = \left(\frac{n-1}{n}\right)^2 \frac{1}{m}\hat{\text{var}}(s_j^2) + \left(\frac{m+1}{mn}\right)^2 \frac{2}{m-1}B^2 + 2\frac{(m+1)(n-1)}{mn^2}\frac{n}{m}\left\{\hat{\text{cov}}(s_j^2, \bar{\theta}_{.j}^2) - 2\bar{\theta}_{..}\hat{\text{cov}}(s_j^2, \bar{\theta}_{.j})\right\}. \tag{5.8}$$

It follows that as $n \to \infty$ then $df \to \infty$.

Then, the potential scale reduction factor is estimated by:

$$\hat{R} = \frac{\hat{V}}{W} = \frac{m+1}{m}\frac{\text{var}^{+}(\theta|y)}{W} - \frac{n-1}{mn} \tag{5.9}$$

and is interpreted as follows: if $\hat{R}$ is large, either the estimate of the marginal posterior variance can be further decreased by more simulations or that additional simulation will increase W. Alternatively, if the PSRF is close to 1 we can conclude that each of the m sets of n simulated observations is approximately equal to the target distribution.

Further, to account for sampling variability in the variance estimates, $\hat{R}$ can be refined using the estimated degrees of freedom, df. Gelman and Rubin [30] adopted the correction factor $df/(df-2)$. This has proven to be incorrect as it can cause for the corrected scale reduction factor to be infinite or negative in cases where convergence is so slow and $df < 2$. Brooks and Gelman [11] propose that the correct factor should be $(df+3)/(df+1)$. Thus a corrected $\hat{R}$ is defined as:

$$\hat{R}_c = \frac{df+3}{df+1}\hat{R} = \frac{df+3}{df+1}\frac{\hat{V}}{W}. \tag{5.10}$$

Parameter / Iterations	$1,000$	$2,000$	$4,000$
Random Effects Model, All Cohorts Associated			
$\boldsymbol{\beta}$	1.0016	1.0018	0.9997
λ	1.0082	1.0022	1.0004
ψ	1.0023	1.0002	0.9998
Fixed Effects Model, All Cohorts Associated			
$\boldsymbol{\beta}_F$	1.0094	1.0009	1.0120
ψ_F	1.0010	1.0007	0.9997
Random Effects Model, Subset of Cohorts Associated			
$\boldsymbol{\beta}$	1.0652	1.0030	0.9999
λ	1.0030	1.0005	0.9999
ψ	1.0012	1.0008	0.9999
π	1.0854	1.0327	1.0002
Fixed Effects Model, Subset of Cohorts Associated			
$\boldsymbol{\beta}_F$	1.0199	1.0117	1.0045
ψ_F	1.0004	0.9996	1.0000
π	1.0061	1.0067	1.0015

Table 5.5: The corrected potential scale reduction factor for the important parameter of each model based on four chains respectively and different number of iterations.

We calculate $\hat{R}_c$ for the notable parameters of each model based on four chains ($m = 4$) and different number of iterations (burn in iterations equals one half of the iterations in each case): $1,000$, $2,000$ and $4,000$ and present them in Table 5.5. Each $\hat{R}_c$ is very close to the desired value of 1 which leads us to believe that further simulations will not change our inference about the target distribution of each associated scalar estimand. Notably, a smaller number of iterations would not have altered our previous simulation results and would have decreased the running time of each model without any power trade off.

5.5.3 Stability of Chib's Method

To conclude the section dedicated to the convergence diagnostics of the Gibbs algorithm, we examine the stability of Chib's method to calculate the marginal likelihood. In order to do so, we obtain the probability to detect the correct model under two heterogeneity levels twice and expect to see very similar results. Table 5.6 displays these results for the data

simulated under $\boldsymbol{\beta} = 0.143$ and $\lambda = 1000, 15$. The conclusions remain the same for the two runs and particularly the probability estimates vary very little.

		Run 1				Run 2			
		All		Subset		All		Subset	
	Ass.	Random	Fixed	Random	Fixed	Random	Fixed	Random	Fixed
$\lambda = 1000$	10	0.519	0.458	0.013	0.010	0.519	0.458	0.011	0.012
	8	0.642	0.017	0.052	0.289	0.651	0.017	0.057	0.275
	6	0.444	0.005	0.075	0.476	0.445	0.008	0.078	0.469
	4	0.293	0.001	0.152	0.554	0.289	0.003	0.146	0.562
	2	0.147	0.005	0.332	0.516	0.142	0.006	0.330	0.522
$\lambda = 15$	10	0.921	0.000	0.071	0.008	0.931	0	0.060	0.009
	8	0.833	0.000	0.148	0.019	0.833	0.000	0.149	0.018
	6	0.648	0.000	0.288	0.064	0.656	0.000	0.292	0.052
	4	0.419	0.000	0.441	0.140	0.432	0.000	0.437	0.131
	2	0.133	0.001	0.619	0.247	0.146	0.002	0.607	0.245

Table 5.6: Comparison of two runs calculating the probability for the case where $\boldsymbol{\beta} = 0.143$ and $\lambda = 1000, 15$ and different number of associated cohorts using Chib's method to obtain the marginal likelihood (The correct model is highlighted in red while the model which exhibits the highest probability is highlighted in blue).

5.6 Analysis of Real Samples

In this section we present the analysis performed on a set of real samples for which collected data refers to smoking behaviour and smoking-related diseases. Different diseases are associated with smoking with lung cancer being the most identified. Another serious lung disease, chronic obstructive pulmonary disease is largely attributable to smoking.

Recent genetic studies [42, 70, 69, 104, 27] have reported findings for nicotine dependence and smoking related diseases on chromosome 15 and, more specifically, on the $15q25.1$ region which includes the *CHRNA5-CHRNA3-CHRNB4* cluster of cholinergic nicotinic receptor subunit genes. The most significant association genome-wide in recent meta-analyses was found to be the single nucleotide polymorphism rs16969968. It has been replicated for smoking related traits, including cigarettes-per-day and heavy smoking. This variant results in an amino acid change in the alhpa5 receptor subunit protein and has been shown to affect receptor function.

We performed a meta analytic study, using both fixed and random effects models, on a 2Mb region on chromosome 15 in a total sample of 21,952 individuals of European descent across 11 cohorts. The main focus of the analysis was on smoking quantity within current and past smokers defined as a semi-quantitative trait, based on the self-reported variable of cigarettes smoked per day (CPD). Further to the reported SNP, we identified a deletion as the most significant association.

5.6.1 Data Description

The data consisted of 11 cohorts and a total of 21,952 individuals of European descent, sourced from several disease, population and control cohorts. Table 5.7 gives a summary of the cohorts used in the meta-analysis. The last column shows the number of individuals in each cohort with non-missing phenotype information for the CPD phenotype.

			Individuals	
Label	Description	Genotyping	All	CPD
WTCCC-HT	Hypertension cases	Affymetrix500K	1,952	828
GEMS	Dyslipidemia case-control	Affymetrix500K	1,847	860
GSK_COPD	COPD case-control	Illumina 550	1,633	1,630
GSK_Bipolar	Bipolar depression case-control	Illumina 550	1,805	942
GSK_UPD	Unipolar depression case-control	Illumina 550	1,792	897
KORA	Population-based	Affymetrix500K	1,644	251
LOLIPOP	Population-based	Affymetrix500K	1,288	648
MedStar	Coronary artery disease case-control	Affymetrix 6.0	1,322	818
POPGEN	Population-based	Affymetrix 6.0	1,107	571
CoLaus	Population-based	Affymetrix500K	5,636	3,130
WTCCC-CAD	Coronary artery disease cases	Affymetrix500K	1,926	1,235

Table 5.7: Summary information for the cohorts of real samples used in the meta-analysis.

5.6.2 Pre-phasing and Imputation

As the cohorts were genotyped on different SNP arrays, we need to impute the genotypes for all data sets. We use the 1000 Genomes Project March 2012 release where the reference panel consists of 1,092 individuals typed at > 38 million polymorphic sites. The reference panel used by Liu et al. [44] to impute SNPs in the interval 76.4-77.0 Mb (build 36) on chromosome 15, for the 11 cohorts, was the 1000 Genomes Pilot 1 data (April 2009) at 108 estimated CEU haplotypes and 2,189 SNPs in the given interval. The March 2012 collection of haplotypes provided by the 1000 Genomes Project results in direct benefits to genetic studies as it forms a more dense catalogue of variants and increases imputation accuracy and power, especially for low allele frequency variants. The accuracy of using the 1000 Genomes Project data (Phase I), to impute the genotypes of common variants at sites not on the original arrays is typically $90 - 95\%$ in non-African and approximately 90% in African ancestry genomes. For low frequency variants, imputed genotypes have $60\% - 90\%$ accuracy in all populations [95]. Nevertheless, this imposes a high computational cost. In order to overcome this burden we employ the strategy of "pre-phasing" introduced by Howie et al. [37] that achieves to maintain the accuracy and the power while decreasing the computational expense. This approach starts by statistically estimating the haplotypes

underlying the study genotypes and then imputing into these haplotypes as if they were correct.

In order to phase the study genotypes, we use **SHAPEITv1** [18], a phasing algorithm with linear complexity which essentially allows us to use a larger set of conditioning haplotypes in each iteration to estimate an individual's haplotype without any additional cost.

The imputation step is performed with **IMPUTE2** where the reference panel used is the 1000 Genomes Project March 2012 release. This reference set consists of 1,092 individuals from multiple populations across the world, hence capturing a broad cross-section of human genetic diversity. The use of such reference sets becomes more advantageous for low frequency variants (MAF $< 5\%$) because these may be unrepresented in a specific population reference panel but well-represented in a panel from a different population. As imputation error rate increases while the minor allele frequency decreases, we are driven to the calculation of a post-imputation information measure. This helps us to assess the quality of the imputed genotypes which if low, it is advisable to exclude such variants before performing association and subsequently meta-analysis [52].

IMPUTE2 calculates a quantity of the relative statistical information about the variant's allele frequency from the imputed data. A value of 1 indicates that the genotypes are imputed with complete certainty whereas a value of 0 indicates complete uncertainty about the results. In order to observe how this information measure varies with the allele frequency of a variant, we perform two independent imputation runs of a cohort, LOLIPOP. Figure 5.11 shows the post-imputation information measure for each run on the respective axis while different colours refer to different allele frequency bins. We can, clearly, observe that as the allele frequency increases the two information measures lie close or on the 45^o degree line and that the information measure for variants with MAF< 0.01 is highly unstable. The proportion of these rare variants, as we can observe from Figure 5.11, is very large and their instability to impute makes it even harder to analyze them further in downstream analyses. For the subsequent analyses we exclude variants with a post-imputation measure < 0.4.

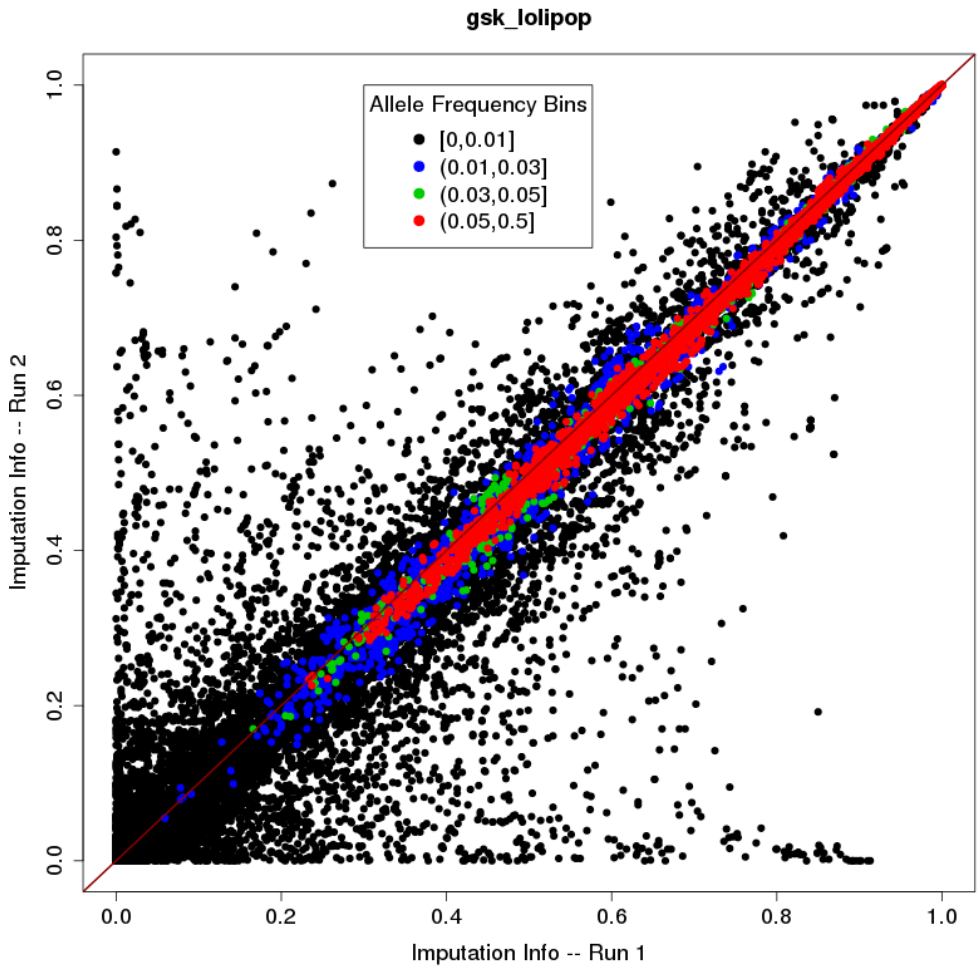

Figure 5.11: Post-imputation information measure from two independent imputation runs for the cohort LOLIPOP. Allele frequencies are divides into the following bins: $[0, 0.01)$ black, $(0.01, 0.03]$ blue, $(0.03 - 0.05]$ green and $(0.05, 0.50]$ red.

5.6.3 Association Analysis

The step that follows pre-phasing and imputation, which is the last one where we analyze each cohort separately, is association testing using the imputed data. Additional care should be taken, though, when handling imputed variants, as they carry uncertainty from the previous step. One approach is to use imputed genotypes that have a posterior probability above a given threshold but this increases the false positive rate and leads to loss of power. An alternative, simple approach is to use the expected allele count (also known as the allele dosage). This strategy has been shown to work just as well as methods which take into account the genotype uncertainty when the effect size of the risk allele is small, both in a frequentist (p-value) and a Bayesian (Bayes factor) framework.

To carry out single-variant association analyses for the eleven cohorts we use **SNPTEST** which implements the expected allele count approach. The levels of the smoking quantity used are 0 (1-10 CPD), 1 (11-20 CPD), 2 (21-30 CPD) and 3 (> 31 CPD). Then, frequentist tests for additive genetic effects are performed on a linear scale using linear regression. Each set of study-specific genetic effect estimates and associated standard errors are used further for the meta analysis.

5.6.4 Meta Analysis

The meta-analysis is accomplished by combining the cohort-specific genetic effect estimates, using a fixed effects model with weights given by the inverse of the variance of the study specific genetic effect estimates. A genomic control inflation factor estimate was obtained by Liu et al. [44] for each cohort and is used to correct for observed inflation (Table 5.8). This quantity is defined as the ratio of the median of the empirically observed distribution of the test statistic to the expected median. This quantifies the extent of the bulk inflation and the plethora false positive rate [17].

The present meta-analysis is conducted using the **META** program which is designed to

Label	Genomic Control Inflation Factor
WTCCC-HT	1.000426
GEMS	1.019739
GSK_COPD	0.9960527
GSK_Bipolar	1.003575
GSK_UPD	1.012388
KORA	1.030356
LOLIPOP	1.008873
MedStar	1.009704
POPGEN	1.035871
CoLaus	1.0131102
WTCCC-CAD	1.010912

Table 5.8: Genomic control inflation factor for each cohort

work with the output of **SNPTEST** [43]. The software allows one to fit both a fixed effects and a random effects model as well as a z-statistic based method. The latter uses the cohort-specific p-values and direction of effect, to form a signed z-statistic which are then summed with weights proportional to the square root of each study's sample size. After fitting the fixed effects model, the resulting combined p-values for the variants with $K \geq 2$ spanning the 2Mb region on chromosome 15 are presented in Figure 5.12. SNPs and indels are plotted by their positions on the chromosome against the negative of the $\log_{10}$ p-value, which assessed association with smoking quantity, on the left y-axis. The recombination rates across the region are plotted against the right y-axis. SNPs are represented as dark blue circles while insertions and deletions are marked as light blue squares. The highest signal, a single-base deletion positioned on an intron on the CHRNA5 gene, is coloured in green while the non-synonymous SNP rs16969968 that has been previously reported to have a strong association with nicotine dependence risk is highlighted in red. The corresponding hit plot obtained from **METASOFT**'s fixed effects model is presented in Figure 5.13. The resulting p-values appear to be analogous as expected since we are fitting the same model in both cases. The Bayesian equivalent results can be found in Figure 5.14. We use the same hyperparameter values as we did in the simulation study (Table 5.2) but decreased the number of iterations(burn-in iterations in brackets) from 5,000(2,000) to 2,000(1,0000)

since we have observed that convergence is achieved at a smaller number of iterations. The same deletion appears to be the hit with the highest Bayes factor, its value being 14.570 on the $\log_{10}$ scale, the marginal likelihood calculated under Chib's method. The $\log_{10}$ Bayes factor of rs16969968 equals 10.600. The next set of hit plots refers to the p-values of **METASOFT**'s binary effects model and the $\log_{10}$ Bayes factors of the fixed effects model with subset of cohorts associated (Figures 5.15 and 5.16). Although the deletion appears to be the top hit in both cases, the $-\log_{10}$ p-value and the corresponding $\log_{10}$ Bayes factor are smaller than in the previous models. The same trend is observed for the complete set of the examined variants.

The successive set of hit plots relates to the results of the random effects models. First, we look at the results of **METASOFT**'s random effects model (Figure 5.17) and our approach's which assumes all cohorts are associated (Figure 5.18). The aftermath of these models is the consistent behaviour of the deletion and SNP of interest, nonetheless we observe negative $\log_{10}$ Bayes factor for some of the SNPs with the lowest being at -6.335 for SNP rs6495348.

We, finally, take a look at the outcome of the general random effects model, allowing for cohorts not to be associated. Given the current specification of model parameters and hyper-parameters we observe a comprehensive decrease in the values of the $\log_{10}$ Bayes factors. Even strong associations, like the top deletion, do not exhibit the large $\log_{10}$ Bayes factors obtained earlier.

Table 5.9 presents the meta-analysis p-values and Bayes factors for the top deletion, 15-78874842, as obtained from the different models. We observe that **META** and **METASOFT**'s fixed effects model p-values are very similar as we would expect. While **METASOFT**'s random effects model produces an equally large $-\log_{10}$ p-value as the fixed effects model, the binary effects model gives a smaller but still highly significant p-value. The $\log_{10}$ Bayes factors which correspond to the three **METASOFT** models have the same ranking as the p-values and the fixed effects model, all cohorts associated appears to be the best model to fit the data. The same conclusion is also drawn by looking at the posterior probabilities

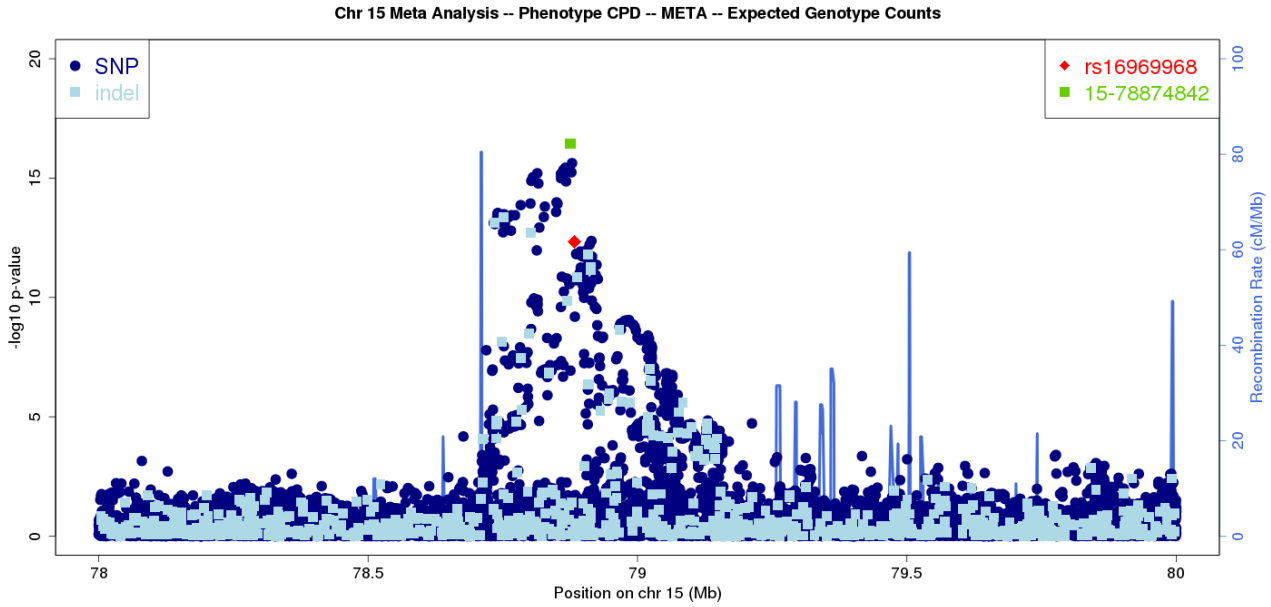

Figure 5.12: Meta-analysis p-values for the SNPs and indels of the 15q25 region obtained from **META** via fitting a fixed effects model. SNPs and indels are plotted by their positions on the chromosome against the negative of the $\log_{10}$ p-value, on the left y-axis. The recombination rates across the region are plotted against the right y-axis. SNPs are represented as dark blue circles while insertions and deletions are marked as light blue squares. The highest signal, a single-base deletion is coloured in green while the non-synonymous SNP rs16969968 is highlighted in red.

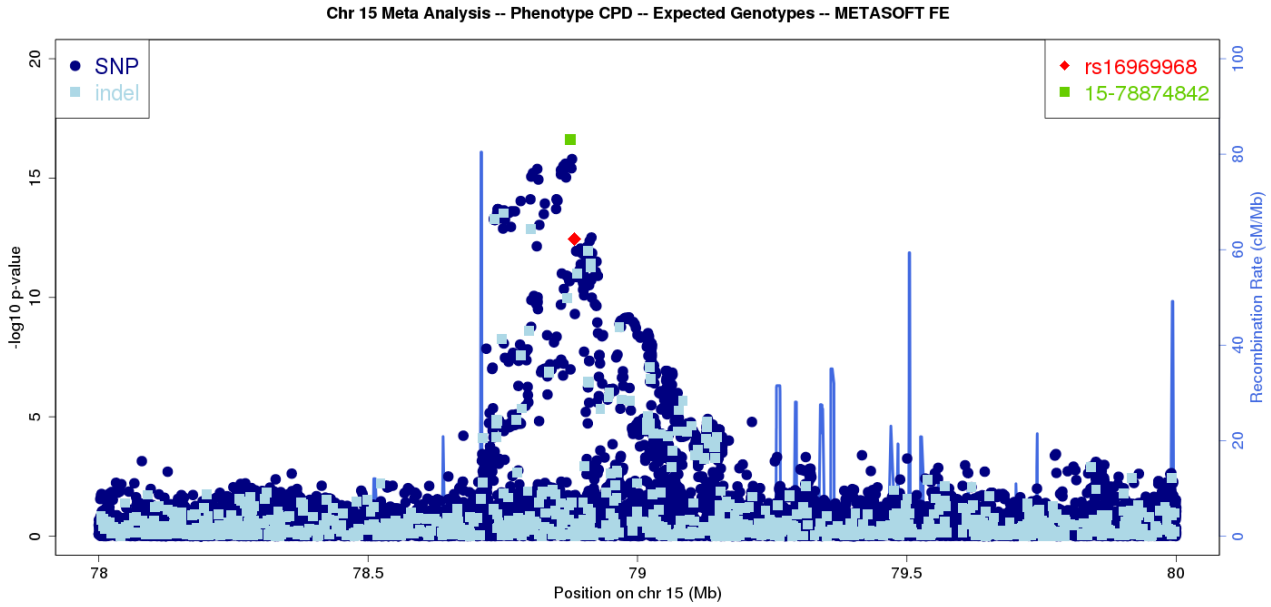

Figure 5.13: Meta-analysis p-values for the SNPs and indels of the 15q25 region obtained from **METASOFT**'s fixed effects model. SNPs and indels are plotted by their positions on the chromosome against the negative of the $\log_{10}$ p-value, on the left y-axis. The recombination rates across the region are plotted against the right y-axis. SNPs are represented as dark blue circles while insertions and deletions are marked as light blue squares. The highest signal, a single-base deletion is coloured in green while the non-synonymous SNP rs16969968 is highlighted in red.

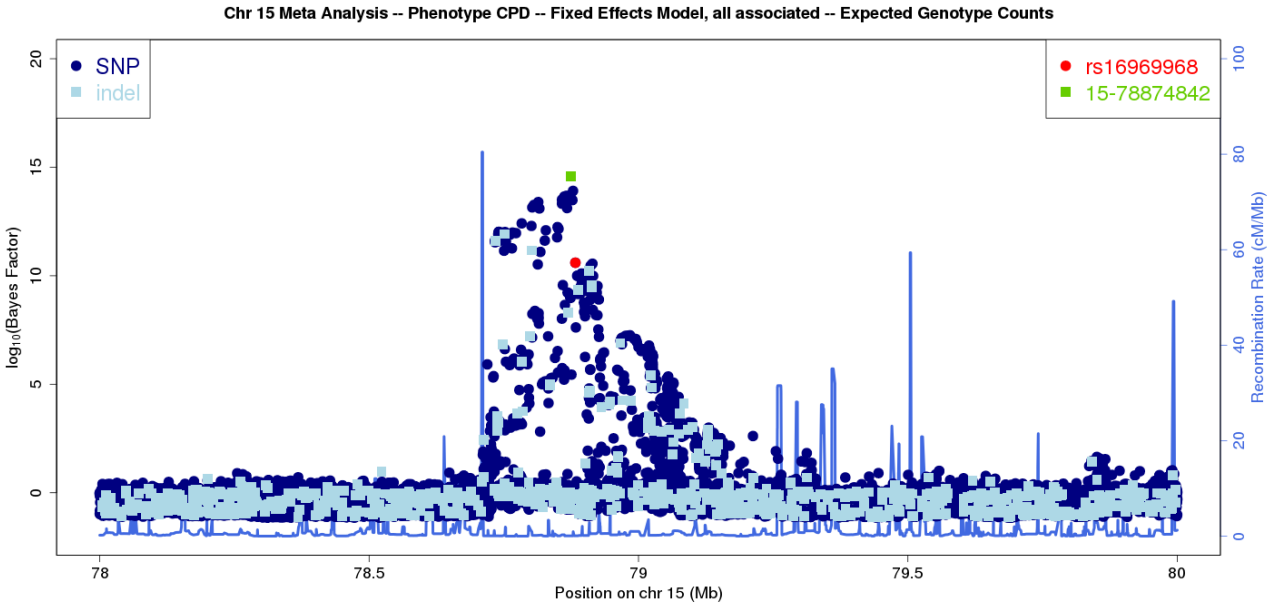

Figure 5.14: Meta-analysis Bayes factors for the SNPs and indels of the 15q25 region obtained under the fixed effects model, all cohorts associated. SNPs and indels are plotted by their positions on the chromosome against against the $\log_{10}$ Bayes factor, on the left y-axis. The recombination rates across the region are plotted against the right y-axis. SNPs are represented as dark blue circles while insertions and deletions are marked as light blue squares. The highest signal, a single-base deletion is coloured in green while the non-synonymous SNP rs16969968 is highlighted in red.

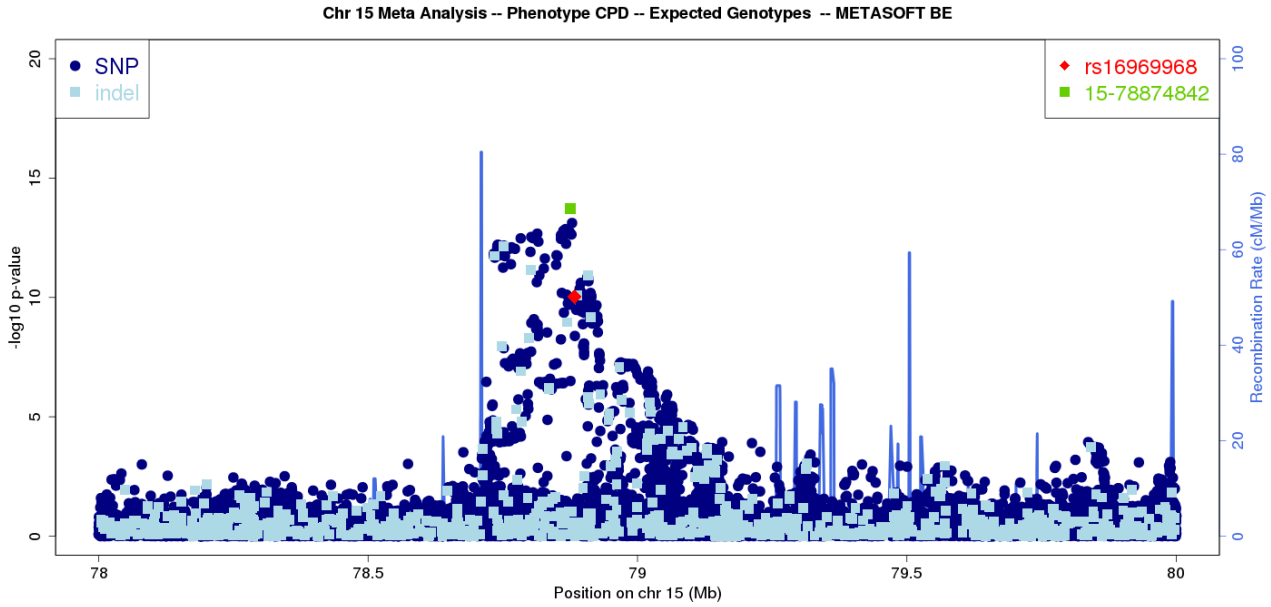

Figure 5.15: Meta-analysis p-values for the SNPs and indels of the 15q25 region obtained from **METASOFT**'s binary effects model. SNPs and indels are plotted by their positions on the chromosome against the negative of the $\log_{10}$ p-value, on the left y-axis. The recombination rates across the region are plotted against the right y-axis. SNPs are represented as dark blue circles while insertions and deletions are marked as light blue squares. The highest signal, a single-base deletion is coloured in green while the non-synonymous SNP rs16969968 is highlighted in red.

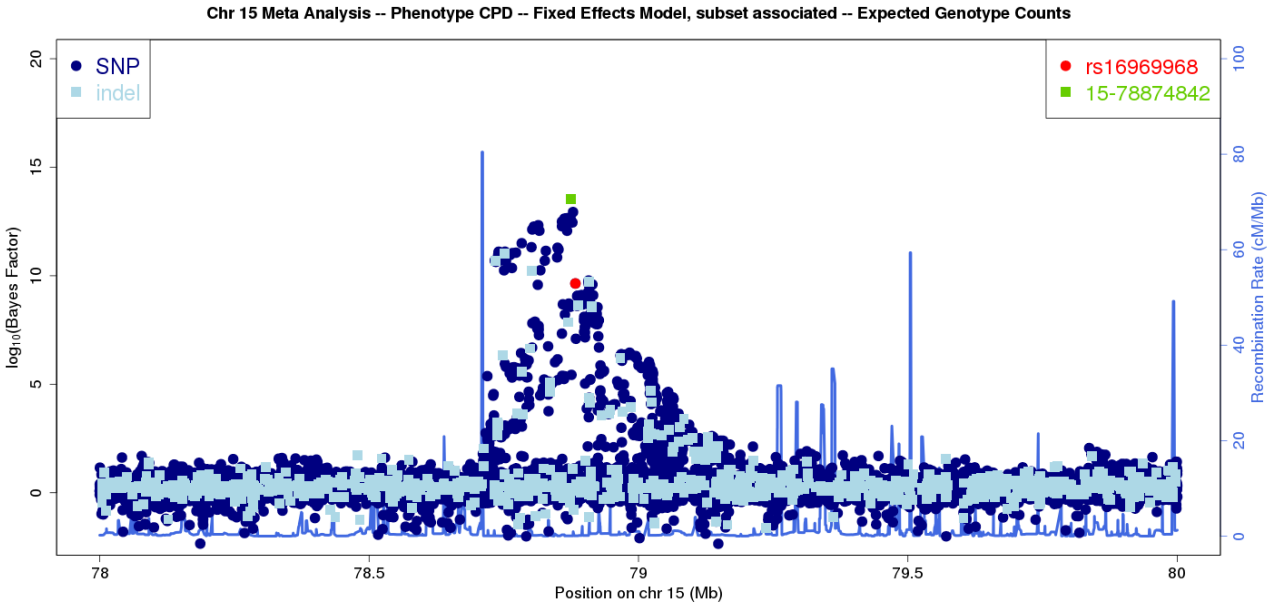

Figure 5.16: Meta-analysis Bayes factors for the SNPs and indels of the 15q25 region obtained under the fixed effects model, subset of cohorts associated. SNPs and indels are plotted by their positions on the chromosome against against the $\log_{10}$ Bayes factor, on the left y-axis. The recombination rates across the region are plotted against the right y-axis. SNPs are represented as dark blue circles while insertions and deletions are marked as light blue squares. The highest signal, a single-base deletion is coloured in green while the non-synonymous SNP rs16969968 is highlighted in red.

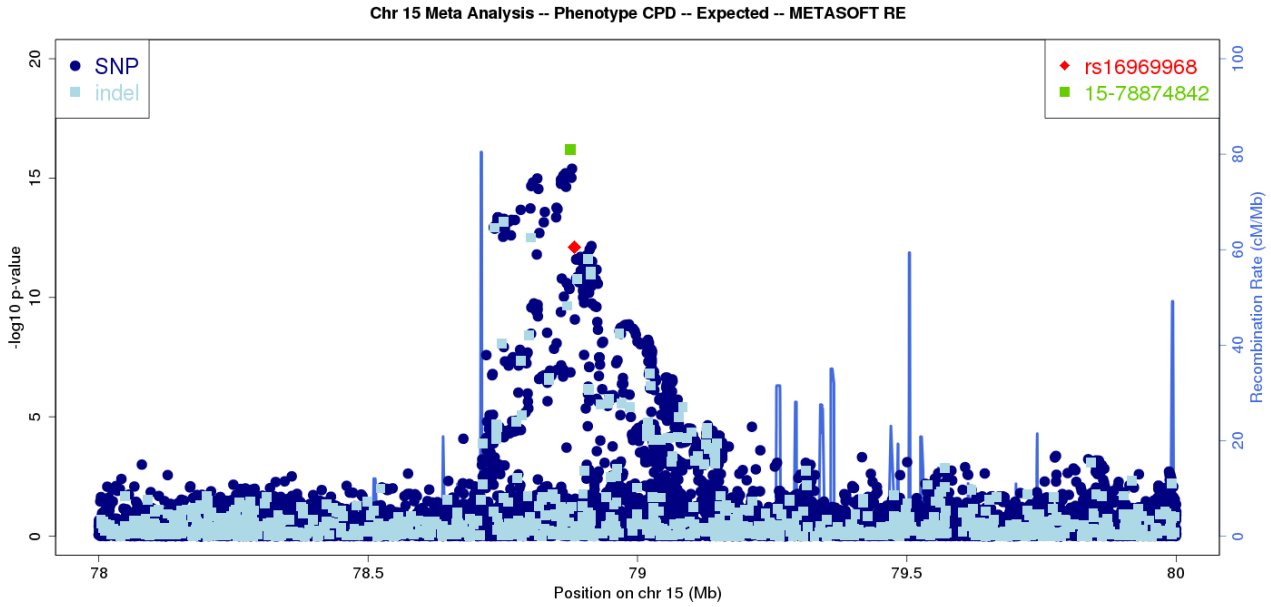

Figure 5.17: Meta-analysis p-values for the SNPs and indels of the 15q25 region obtained from **METASOFT**'s random effects model. SNPs and indels are plotted by their positions on the chromosome against the negative of the $\log_{10}$ p-value, on the left y-axis. The recombination rates across the region are plotted against the right y-axis. SNPs are represented as dark blue circles while insertions and deletions are marked as light blue squares. The highest signal, a single-base deletion is coloured in green while the non-synonymous SNP rs16969968 is highlighted in red.

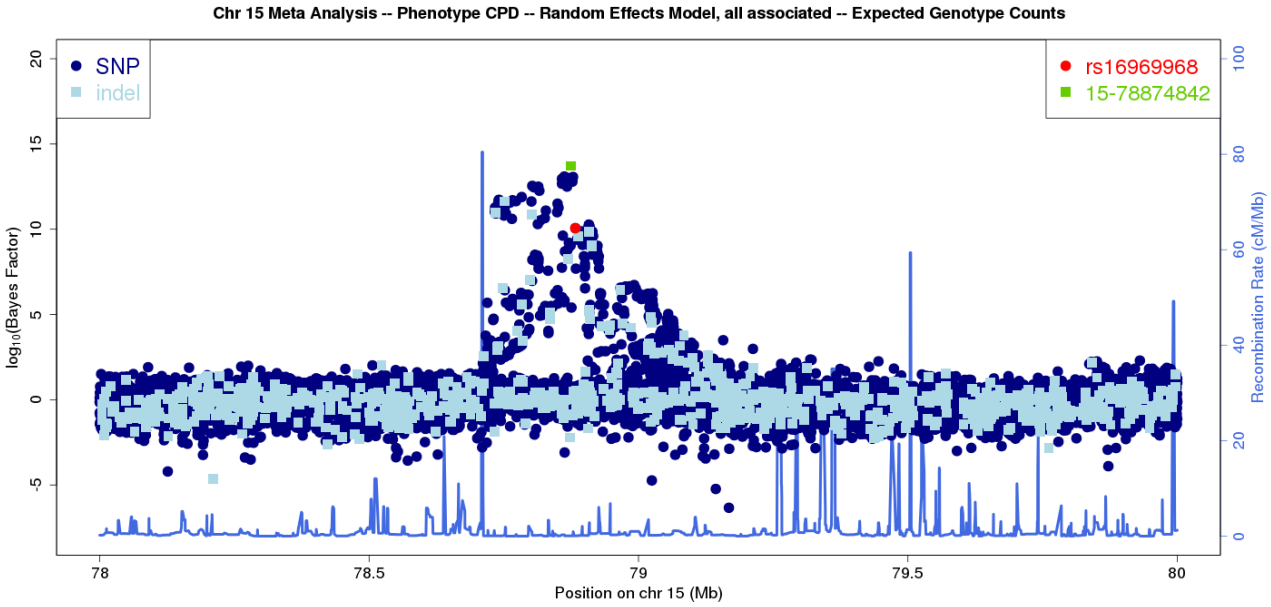

Figure 5.18: Meta-analysis Bayes factors for the SNPs and indels of the 15q25 region obtained under the random effects model, all cohorts associated. SNPs and indels are plotted by their positions on the chromosome against the $\log_{10}$ Bayes factor, on the left y-axis. The recombination rates across the region are plotted against the right y-axis. SNPs are represented as dark blue circles while insertions and deletions are marked as light blue squares. The highest signal, a single-base deletion is coloured in green while the non-synonymous SNP rs16969968 is highlighted in red.

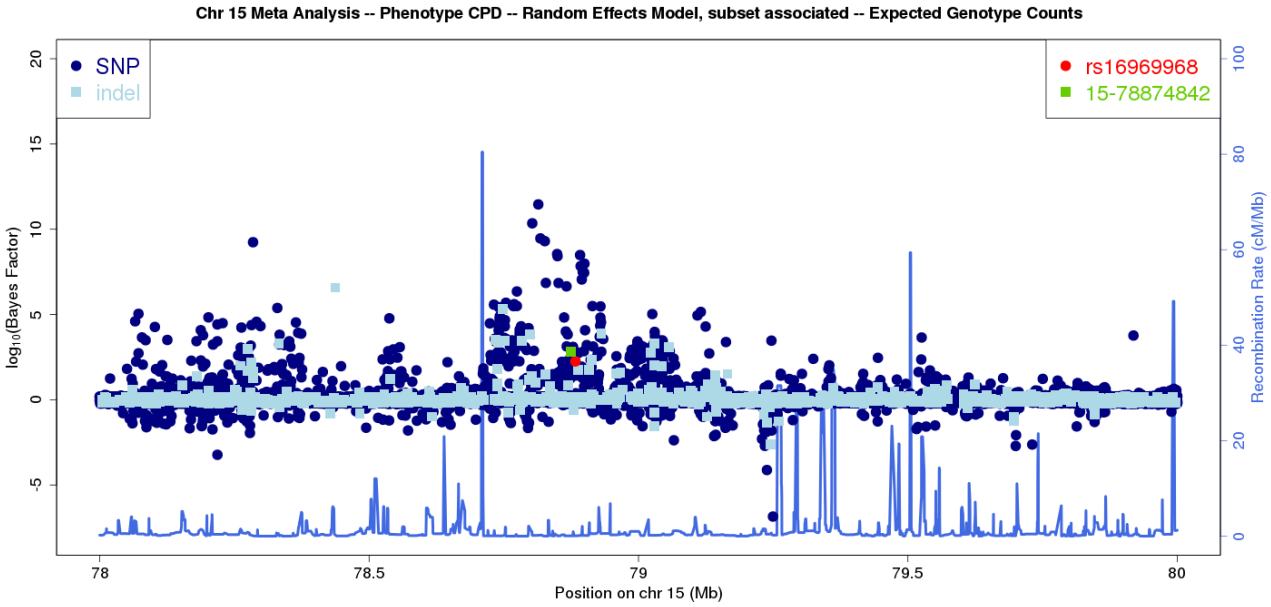

Figure 5.19: Meta-analysis Bayes factors for the SNPs and indels of the 15q25 region obtained under the random effects model, subset of cohorts associated. SNPs and indels are plotted by their positions on the chromosome against the $\log_{10}$ Bayes factor, on the left y-axis. The recombination rates across the region are plotted against the right y-axis. SNPs are represented as dark blue circles while insertions and deletions are marked as light blue squares. The highest signal, a single-base deletion is coloured in green while the non-synonymous SNP rs16969968 is highlighted in red.

of the four models of our approach. On the contrary, the model with the smallest Bayes factor is the general random effects model which allows for cohorts not to be associated.

Model	P-value
META Fixed Effects Model	3.60×10^{-17}
METASOFT Fixed Effects Model	2.41×10^{-17}
METASOFT Random Effects Model	6.27×10^{-17}
METASOFT Binary Effects Model	1.85×10^{-14}
Model	$\log_{10}$ Bayes Factor
Fixed Effects, All Associated	14.57
Random Effects, All Associated	13.83
Fixed Effects, Subset Associated	13.54
Random Effects, Subset Associated	6.06
Model	$P(M_k\|Data)$, $k = 1, 2, 3, 4$
Fixed Effects, All Associated	0.811
Fixed Effects, Subset Associated	0.113
Random Effects, All Associated	0.076
Random Effects, Subset Associated	0

Table 5.9: Meta analysis p-values, Bayes Factors and $P(M_k|Data)$ for $k = 1, ..., 4$ for the deletion at position 78874842 on chromosome 15.

We explore, further, the results of the general random effects model for the deletion which is the top hit for the rest of the models. A possible explanation for the decreased Bayes factors is the small sizes of some cohorts e.g. KORA and POPGEN which are jointly analyzed with very large cohorts, for example, LAUSANNE. The hyperparameters which describe the distribution of the heterogeneity parameter λ also play an important role. So far we have used $c = 4$ and $d = 0.008$, which were also used throughout our simulation study. This choice was made to achieve a prior distribution for λ that would cover a large range of heterogeneity levels, from very little to very large.

We, now, focus on alternative values of these two hyperparameters. Some lead to distributions with small means and increasing variance while we also test one case where the prior mean on λ is large and the standard deviation small (Figure 5.20). We compare some important posterior quantities for these choices for the two random effects model of our approach such as $E(\boldsymbol{\beta}|Data)$ and $E(\lambda|Data)$ which are presented in Table 5.10. Irrespective of the prior assumptions on λ, the posterior mean of $\boldsymbol{\beta}$ is approximately equal to 0.120. On the other hand, the posterior mean and standard deviation of λ vary according to the prior

beliefs. This has an impact on the calculation of the marginal likelihood and the Bayes factor subsequently. While the Bayes factors of the random effects model, all cohorts associated, do not vary greatly, we see a large discrepancy between the smallest and largest Bayes factor under the general random effects model. The conclusions we draw from these results are as follows. When the prior distribution of a parameter focusses on particular values away from the truth, we are led to incorrect marginal likelihood estimates. Moreover, when the prior distribution adds equal weight to all values of λ apart from zero, we obtain more reasonable Bayes factors. We also observe large expected values for λ even if these values are not covered apriori which lead us to determine the fixed effects nature of the genetic effect sizes at the examined cohorts.

We choose to focus on $c = 500$ and $d = 1$ which lead to a larger prior mean of λ and a small prior standard deviation than the default hyperparameters since the large $\log_{10}$ Bayes factors for the fixed effects models suggest the absence of heterogeneity between cohorts. Figure 5.21 presents the hit plot of the examined region where the $\log_{10}$ Bayes factors correspond to the general random effects model for $c = 500$ and $d = 1$. We observe significant differences from the hit plot of Figure 5.19. The deletion 15-78874842 appears to have the largest $\log_{10}$ Bayes factor, equal to 12.74 The corresponding posterior probabilities for the four models when the two random effects models are applied using $c = 500$ and $d = 1$ are as follows: fixed effects model, all cohorts associated 0.827, fixed effects model, subset of cohorts associated 0.075, random effects model, all cohorts associated 0.086 and general random effects 0.012. The $\log_{10}$ Bayes factor for rs16969968 is 9.04. We observe, though, the tendency for larger $\log_{10}$ Bayes factors for variants which the random effects model, all cohorts associated, resulted in smaller estimates.

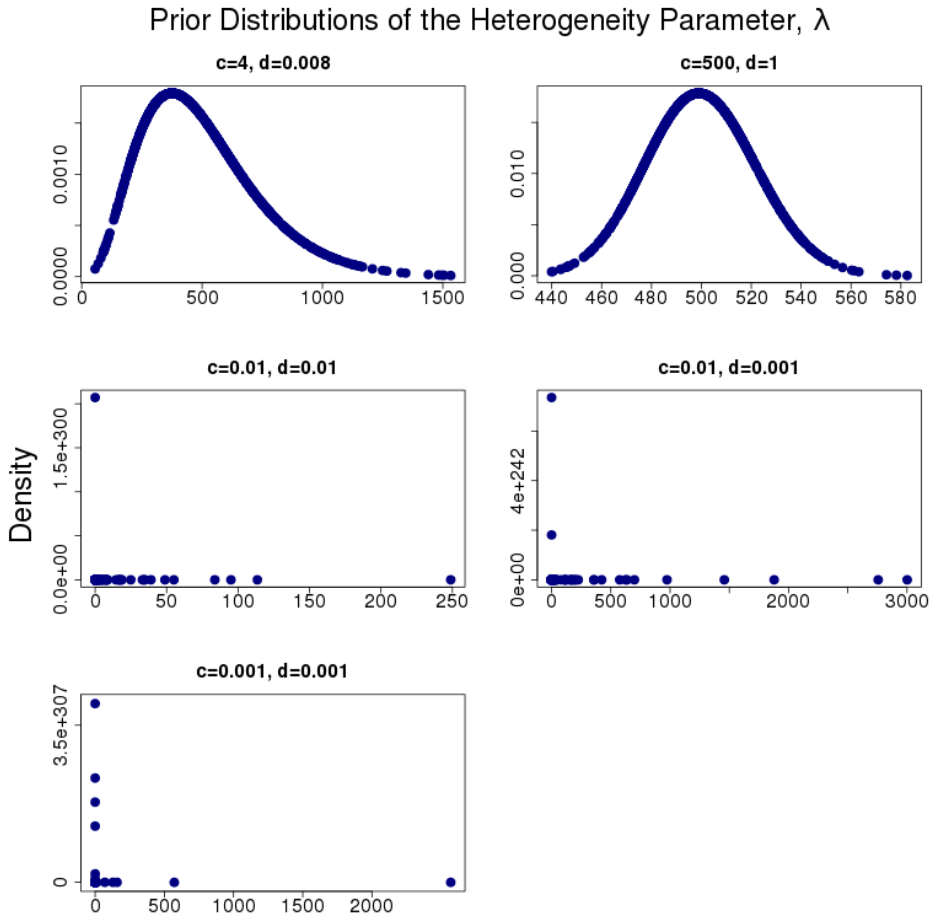

Figure 5.20: Prior distributions of the heterogeneity parameter λ for different combinations of the hyperparameters c and d.

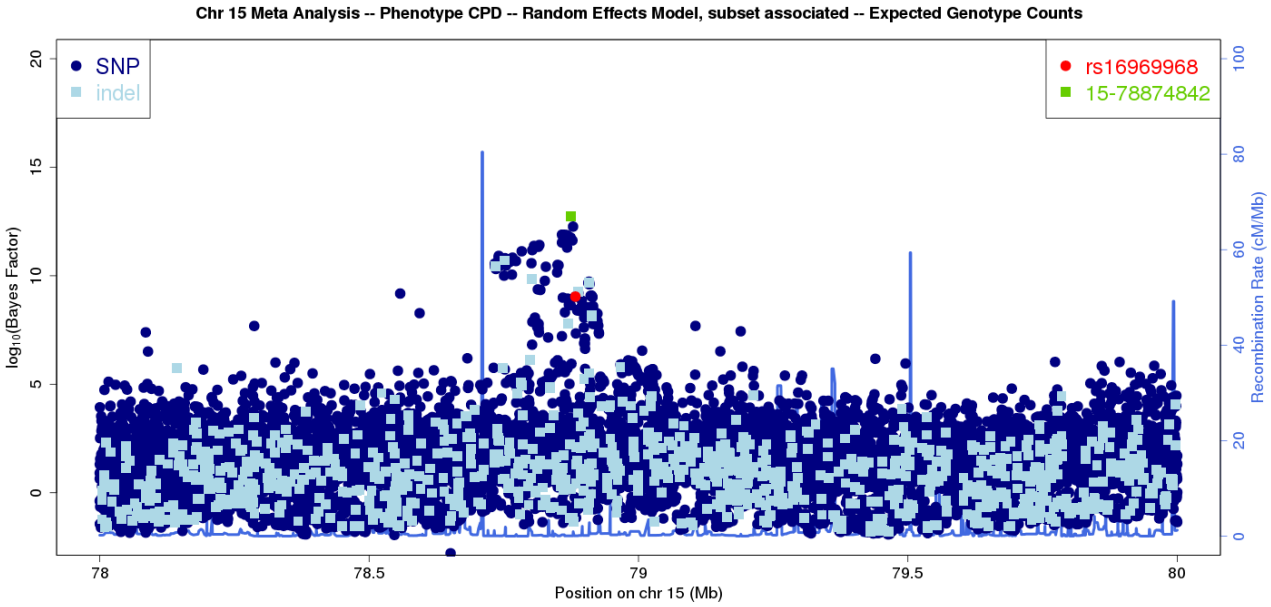

Figure 5.21: Meta-analysis Bayes factors for the SNPs and indels of the 15q25 region obtained under the random effects model, subset of cohorts associated when $c = 500$ and $d = 1$. SNPs and indels are plotted by their positions on the chromosome against the $\log_{10}$ Bayes factor, on the left y-axis. The recombination rates across the region are plotted against the right y-axis. SNPs are represented as dark blue circles while insertions and deletions are marked as light blue squares. The highest signal, a single-base deletion is coloured in green while the non-synonymous SNP rs16969968 is highlighted in red.

Hyperparameters	Posterior Summaries			
	$E(\boldsymbol{\beta}\|Data)$	$E(\lambda\|Data)$	$SD(\lambda\|Data)$	$\log_{10}$BF
	Random Effects Model, All Associated			
$c = 4, d = 0.008$	0.120	703.62	271.97	13.83
$c = 500, d = 1$	0.120	501.38	22.53	13.56
$c = 0.01, d = 0.01$	0.120	266.17	139.49	10.29
$c = 0.01, d = 0.001$	0.120	1284.36	940.81	12.86
$c = 0.001, d = 0.001$	0.118	1300.52	915.71	10.03
	Random Effects Model, Subset Associated			
$c = 4, d = 0.008$	0.112	590.68	262.77	6.06
$c = 500, d = 1$	0.120	502.52	21.58	12.74
$c = 0.01, d = 0.01$	0.122	229.74	134.50	16.42
$c = 0.01, d = 0.001$	0.116	1031.39	785.80	19.54
$c = 0.001, d = 0.001$	0.123	1219.54	846.06	20.46

Table 5.10: Posterior summaries of the top deletion under the two random effects models of our meta-analysis approach for different combinations of c and d.

Table 5.11 summarizes the posterior probability of an effect in each cohort for the top deletion under the random, fixed and binary effects model. We observe that the posterior probability under the fixed effects model and the m-value are very close for each cohort. On the other hand, the results of the random effects model under different combinations of hyperparameters are not so stable except for LAUSANNE which is the largest cohort. KORA, which is the smallest cohort, shows the smallest posterior probability across cohorts for all models. These observations add support to the fact that the performance of the general random effects model is affected by the choice of the hyperparameters describing the prior distribution of the heterogeneity parameter. Hence, testing the performance under different prior conditions will allow us to explore the abilities of the model and increase the confidence in our results.

Cohorts	Random Effects Model					Fixed Effects Model	m-value
	$c = 0.01$ $d = 0.01$	$c = 0.01$ $d = 0.001$	$c = 0.001$ $d = 0.001$	$c = 4$ $d = 0.008$	$c = 500$ $d = 1$		
GEMS	0.940	0.785	1	0.531	1	1	0.997
COPD	0.977	0.961	1	0.632	1	1	0.999
BIPOLAR	0.769	0.725	0.930	0.286	1	0.984	0.981
UNIPOLAR	0.722	0.562	0.937	0.172	1	0.974	0.971
KORA	0.673	0.528	0.842	0.150	1	0.964	0.960
LOLIPOP	0.934	0.859	0.981	0.456	1	1	0.997
MEDSTAR	0.846	0.726	0.964	0.296	1	0.991	0.988
LAUSANNE	1	1	1	1	1	1	1
POPGEN	0.856	0.743	0.952	0.311	1	0.993	0.992
HT	0.867	0.811	0.976	0.345	1	0.992	0.994
CHD	0.919	0.887	0.995	0.460	1	0.994	0.997

Table 5.11: Posterior probability of an effect in each cohort of the top deletion under the random effects model for different combinations of c and d, the fixed effects model and the binary effects model (m-value).

5.7 Summary

Over the course of this chapter we have performed thorough simulations to test the performance of our meta analysis approach and compared its efficiency to an existing method, **METASOFT**. In the end, we applied it to real samples for which collected data referred to smoking behaviour and smoking-related diseases. We used the same haplotype data as in Chapter 3 and constructed phenotypes under different characteristics. We considered 1,000 large meta analyses each consisting of ten studies with different values of the heritability estimate h^2 and $\boldsymbol{\beta}$, a range of heterogeneity levels and different number of associated cohorts.

Our first task was to assess the power of the approach to detect the correct model that the data have come from. We concluded that when all cohorts are associated with the phenotype under all considered values of the heterogeneity parameter and the mean of the random effects distribution, the true model is identified with large power. As the heterogeneity in the effect sizes increased, we observed an increase in the power of the random effects models with the one allowing for cohorts not to be associated to exhibit larger power for decreasing

number of associated studies. When we examined values of $\boldsymbol{\beta}$ closer to zero with little or no heterogeneity, power estimates were not as high since effect sizes were very small. Once λ got bigger and simulated effect sizes included both positive and negative values, power escalated to the same levels as when only positive effect sizes were considered.

We assessed the power of the models to detect association via the means of the Bayes factor and compared the results to the p-values of the three **METASOFT**'s statistics. In order to directly compare the two approaches we obtained the empirical null distributions of each Bayesian model and subsequently obtained thresholds corresponding to level of significance used to calculate the power of the **METASOFT** p-values. Ultimate power was achieved for all models when $\boldsymbol{\beta} = 0.143$ and heterogeneity was not very large. As the number of associated cohorts decreased, so did the power. For smaller values of λ we detected a grouping of the methods where the random effects models and the fixed effects model with subset of cohorts associated belonged to the group with the highest power. When the data was simulated under the fixed effects model with different number of cohorts associated in each case, **METASOFT**'s binary effects model failed to exhibit similar power to the analogous Bayesian model. Especially, as the values of $\boldsymbol{\beta}$ increased, the difference between the two models was the greatest.

We also compared the posterior probability of detecting an effect, under both the random and fixed effects models, with **METASOFT**'s m-value and deduced that the m-value produces analogous results to our version of calculating the same quantity as was expected since the main assumption of both models is homogeneity for the cohort specific effect sizes.

Before concluding the chapter with the analysis of the real samples, we performed convergence analysis to determine the stability of the Gibbs algorithm with the aid of trace plots and the potential scale reduction factor.

Chapter 6

Conclusions and Future Prospects

In this thesis we have presented two models for fine mapping in genetic association studies which have been described and assessed over two parts. In Chapter 2, the methodology of a novel Bayesian fine mapping model has been proposed. It is a flexible and efficient model which analyses relatively small sets of haplotype data consisting of up to a few thousand individuals and up to a few hundreds of variants. A Bayesian framework is employed in order to find the most powerful combinations of SNPs and genetic models which explain association with the phenotype. The principal feature of our model is that it analyzes haplotype data thus capturing additional biological information which is not directly observable when considering genotype data. A Metropolis-Hastings algorithm is applied in order to search the model space efficiently. We have assessed the performance of our model over existing approaches to detect compound heterozygous effects and have concluded that our approach outperforms them over different structures of linkage disequilibrium in the data. More specifically, **FineMap MCMC** has exhibited larger power to detect marginal as well as joint SNP effects. Although **BIMBAM** has showed larger power to detect the true pair of causal SNPs, under certain circumstances, it has been beaten by **FineMap MCMC** when highly correlated pairs were taken into account. We have additionally performed converge diagnostics for the Metropolis Hastings algorithm by calculating the Kullback-Leibler di-

vergence between the true and estimated marginal posterior distribution of the SNPs and by examining the acceptance rates of the chains for the different LD levels.

In Chapter 4 we have described a set of Bayesian hierarchical models for meta-analysis. This new approach accommodates both fixed and random effects and accounts for the existence of an effect in each study. The most general model is the random effects model accounting for possible non-association in the cohorts of the meta-analysis. The other three models are special cases of this model. In order to assist the infrence procedure we have used the Gibbs sampler to obtain samples from each of the parameters' marginal posterior distribution. Posterior summaries such as the posterion mean and variance can be used to provide information on the parameters given the data but also, quite importantly, can be used in the approximation of the marginal likelihood which leads to the Bayes factor of association. We primarily use Chib's method to calculate the marginal likelihood which is based on the fact that the joint posterior distribution can be approximated by a product of conditionals [12] while the alternative approach is based on the Bayesian deviance [67]. Detailed simulations have shown that, based on the posterior probability of each model and under different levels of heterogeneity and number of associated cohorts, the correct underlying model has a large probability to be detected. Moreover, our proposed meta-analysis approach has outperformed, in many cases, existing approaches. The fixed effects model assuming that all cohorts are associated has performed almost the same as **METASOFT**'s fixed effects model, for a range of effect sizes. The fixed effects model accounting for possible absence of a genetic effect in each cohort has especially shown elevated power over its analogous **METASOFT**'s binary effects model, even in the presence of heterogeneity. Specifically, the power of the Bayesian model, for decreasing number of associated cohorts at $\boldsymbol{\beta} = 0.143$ and $\lambda = 1000$, has been twice as high while, for smaller values of $\boldsymbol{\beta}$ the power difference gradually diminishes. For larger levels of heterogeneity, the general fixed effects model has still exhibited higher power across the different values of the mean of the random effects distribution. On the other hand, the three examined random effects models

have shown very similar power levels in the presence of substantial heterogeneity while for larger values of λ and decreasing number of associated cohorts our models have outperformed **METASOFT**' random effects model. Finally, we have obtained the $\log_{10}$ Bayes factors for SNPs and indels in the 2Mb region on chromosome 15 in order to detect association with smoking quantity and have concluded that our Bayes factors show comparable results with the frequentist p-values obtained through **META** and **METASOFT**.

Additional considerations which can expand the spectrum of situations that our models can be applied to include the following ideas. The current underlying assumption with regards to the input data of the fine mapping model is that haplotypes are directly available for the set of individuals. Current technologies, genotyping platforms and imputation algorithms, provide genotype data and in order to extract the haploid structure, phasing algorithms need to be applied. In order to boost power we can address the uncertainty in phase assignment by considering an average of possible haplotype configurations. Moreover, parameters accounting for environmental variation and possible gene$\times$environment interactions can be added to the model in an attempt to achieve a higher estimate of heritability. Exposure factors, dietary habits and lifestyle can provide insights into disease complexity. Such elements can shed more light into both public health and personalized medicine. This consideration can provide additional information especially in a meta-analysis when possible environmental distinctions between cohorts can reveal supplementary sources of heterogeneity [49, 99, 48].

The remarkable advances of The 1000 Genomes Project are undeniable and in the latest data releases not only have SNPs been reported but also insertions, deletions and structural variants. These discoveries urge us to develop new methods or to adjust existing ones to account for these different types of variants. We have already used our meta-analysis approach to measure the association of SNPs as well as of insertions and deletions with smoking quantity and found the top hit to be a single base deletion. Integrating different types of variants within a single framework will only improve the power of association stud-

ies to detect further genetic contribution to disease. Moreover, aside from the association of common variants with disease, rare and low-frequency variants are believed to influence the expression and prevalence of traits and diseases [95, 7], both within the same gene and across different genes, leading to the consideration of variant frequency as well as covariates and previously identified genetic factors in association studies. Such extensions to the current setting of our fine mapping and meta-analysis approaches will be advantageous in identifying rare variants which are strongly believed to shed light on the notion of "missing heritability".

Fine mapping for the location of causal variants and subsequent meta-analysis has primarily been conducted using European-descent populations. The high levels of LD which characterize such populations make identifying and localizing the true causal variants extremely demanding due to sufficient correlation with nearby SNPs. This has increased the interest for transethnic studies both at a single and multiple cohort level. These are conducted across a collection of multiple populations with different ethnic backgrounds. They aim to increase the resolution of fine mapping by expanding the haplotypic diversity of the sample. African populations are characterized by low levels of LD leading to high levels of haplotype diversity which in turn are a powerful tool for fine mapping causal variants. As Africa is the ancestral home of all human populations, understanding the biology of disease in this continent can reveal the genetic origins of common diseases worldwide [92]. For example, a recent transethnic fine mapping study searching for associations with triglycerides, high-density lipoprotein cholesterol and low-density lipoprotein cholesterol respectively was performed in individuals of African American, East Asian and European descent. The focus was to identify and refine additional population-specific signals and assess the significance of previously descibed variants. The study supported the evidence that allelic heterogeneity is a frequent feature of polygenic traits and extended the findings to non-European populations. There was also strong support that fine mapping at genome wide loci identifies population-specific signals and, finally, the results were able to narrow the

sets of candidate functional variants using the increased resolution of fine-mapping by expanding the haplotypic diversity of samples [108]. The efficiency of transethnic studies can explored by comparing different population diversity metrics that will be useful in two instances. First, to identify populations or genomic regions where meta-analyses are likely to be more successful for discovering associations and second, to identify populations or loci to perform deep targeted sequencing for carrying out fine mapping. Based on simulation data from 11 populations, the results indicated that the fixation index or the population specificity of haplotypes are helpful in transethnic meta-analyses, while the degree of haplotype sharing and LD variation can assist in transethnic fine-mapping [101]. The approaches we have presented in Chapters 2 and 4 can easily be extend to incorporate the analysis of samples from multiple and diverse populations including both genetic and environmental information. This will establish a statistical framework to investigate how variations in environment and lifestyle affect the results of genetic association studies in different populations. Transethnic studies are believed to be also applicable to detect lower frequency causal variants [92, 58].

A further consideration would be to use our approaches to discover genetic variants which have a functional impact on gene expression facilitating more robust identification of associations between variants and disease. Statistical genetic methods can be used to map the genetic factors that underline variations in quantitative levels of expression of many thousands of transcripts. The plethora of transcripts might be considered as a quantitative trait that can be mapped with considerable power. These expression QTLs (eQTLs) can be mapped using study designs that are used traditionally to map QTLs. Such studies can not only aid in the interpretation the biological control of expression but also support the interpretation of GWA studies results. The use of eQTLs can assist in fine-mapping causal variants for complex traits and provide testable hypotheses for the mechanism underlying significant GWAS associations. It can provide the tools for a systems biology approach to multifactorial disease and a path to identify and biologically validate suscepti-

bility genes [15, 87].

Efficient methodology for the analysis of genetic data shows great promise for future improvements in our understanding of the genetic architecture of complex traits. Methods will never stop emerging as some challenges in the interpretation of genetic variation remain unanswered. Despite the vast identification of polymorphisms, the search for more biological insights provides a valuable path towards the future.

Appendices

Appendix A

Bayesian Linear Model Derivations

This section describes in detail the derivation of the expression of the $P(D|M)$, the probability of the data given a model formulation in (2.14).

The posterior distribution of the model parameters, as discussed in Section 2.1.3, arises from the application of the Bayes' Theorem. Hence,

$$\begin{aligned} P(\boldsymbol{\beta}, \sigma^2|D) &\propto P(D|\boldsymbol{\beta}, \sigma^2) \times P(\boldsymbol{\beta}, \sigma^2) \\ &= \frac{b^a}{(2\pi)^{k/2}|V|^{1/2}\Gamma(a)}(\sigma^2)^{-(a+(k/2)+1)} \\ &\quad \times \exp[-\{(\boldsymbol{\beta}-\boldsymbol{m})^T\mathbf{V}^{-1}(\boldsymbol{\beta}-\boldsymbol{m})+2b\}/(2\sigma^2)] \\ &\quad \times \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-\frac{(Y-\mathbf{X}\boldsymbol{\beta})^T(Y-\mathbf{X}\boldsymbol{\beta})}{2\sigma^2}\right) \\ &= \frac{b^a}{|V|^{1/2}\Gamma(a)(2\pi)^{(k+n)/2}}(\sigma^2)^{-(a+((n+k)/2)+1)} \times \exp(...) \end{aligned} \tag{A.1}$$

What goes in the exponential in (A.1) is consicely described by the following identity [20]:

$$\begin{aligned} &(Y-\mathbf{X}\boldsymbol{\beta})^T(Y-\mathbf{X}\boldsymbol{\beta})+(\boldsymbol{\beta}-\boldsymbol{m})^T\mathbf{V}^{-1}(\boldsymbol{\beta}-\boldsymbol{m})+2b \\ &\qquad \equiv (\boldsymbol{\beta}-\boldsymbol{m}^*)^T(\mathbf{V}^*)^{-1}(\boldsymbol{\beta}-\boldsymbol{m}^*)+2b^*. \end{aligned} \tag{A.2}$$

With regards to the $P(D|M)$ we proceed as follows

$$
\begin{aligned}
P(D|M) &= \int\int P(D|\boldsymbol{\beta},\sigma^2)P(\boldsymbol{\beta},\sigma^2)d\boldsymbol{\beta}d\sigma^2 \\
&= \int\int \frac{b^a}{|V|^{1/2}\Gamma(a)(2\pi)^{(k+n)/2}}(\sigma^2)^{-(a+((n+k)/2)+1)} \\
&\quad \times \exp\left(-\frac{(\boldsymbol{\beta}-\boldsymbol{m}^*)^T(\mathbf{V}^*)^{-1}(\boldsymbol{\beta}-\boldsymbol{m}^*)+2b^*}{2\sigma^2}\right) \\
&= \frac{b^a}{|V|^{1/2}\Gamma(a)(2\pi)^{n/2}(2\pi)^{k/2}}\int\int(\sigma^2)^{-(a+(n/2)+1)}(\sigma^2)^{-k/2} \\
&\quad \exp\left(-\frac{(\boldsymbol{\beta}-\boldsymbol{m}^*)^T(\mathbf{V}^*)^{-1}(\boldsymbol{\beta}-\boldsymbol{m}^*)}{2\sigma^2}\right)\exp\left(\frac{-2b^*}{2\sigma^2}\right)d\boldsymbol{\beta}d\sigma^2 \\
&= \frac{b^a|V^*|^{1/2}}{|V|^{1/2}\Gamma(a)(2\pi)^{n/2}}\int(\sigma^2)^{-(a+(n/2)+1)} \\
&\quad \underbrace{\int\frac{(\sigma^2)^{-k/2}}{(2\pi)^{k/2}|V^*|^{1/2}}\exp\left(-\frac{(\boldsymbol{\beta}-\boldsymbol{m}^*)^T(\mathbf{V}^*)^{-1}(\boldsymbol{\beta}-\boldsymbol{m}^*)}{2\sigma^2}\right)d\boldsymbol{\beta}}_{Normal\quad Density}\exp\left(\frac{-2b^*}{2\sigma^2}\right)d\sigma^2 \\
&= \frac{b^a|V^*|^{1/2}}{|V|^{1/2}\Gamma(a)(2\pi)^{n/2}}\int(\sigma^2)^{-(\underbrace{a+(n/2)}_{a^*}+1)}\exp\left(\frac{-b^*}{\sigma^2}\right)d\sigma^2 \\
&= \frac{b^a|V^*|^{1/2}\Gamma(a^*)}{|V|^{1/2}\Gamma(a)(2\pi)^{n/2}(b^*)^{a^*}}\underbrace{\int\frac{(b^*)^{a^*}}{\Gamma(a^*)}(\sigma^2)^{-(a^*+1)}\exp\left(\frac{-b^*}{\sigma^2}\right)d\sigma^2}_{Inverse\quad Gamma\quad Density} \\
&= \frac{b^a|V^*|^{1/2}\Gamma(a^*)}{(b^*)^{a^*}|V|^{1/2}\Gamma(a)(2\pi)^{n/2}}
\end{aligned}
$$

■

(A.3)

This is the exact form we see in (2.14).

Appendix B

Naming system for genetic models

- **add_1** → Single additive SNP
- **add_2** → Two additive SNPs
- **gen_1** → Single general SNP
- **gen_2** → Two general SNPs
- **hap2_all** → Full additive haplotype model in the case of 2 SNPs
- The single haplotype models when there are 2 SNPs present are described in the table below. Hence, for example, when we have the haplotype configuration {0 1} we refer to it as model **hap2_2**:

Model	SNP 1	SNP 2
hap2_1	0	0
hap2_2	0	1
hap2_3	1	0
hap2_4	1	1

It should be noted that the notion of the closest model only applies to the additive and general models in the current setting:

Model ($\|S\| = 1$)		Model ($\|S\| = 2$)
add_1	$\longrightarrow$	**add_2**
gen_1	$\longrightarrow$	**gen_2**

If we were to increase the maximum size of S then the haplotype models would also have closest models when a SNP would be added or deleted from the set.

- The genetic models for which two different haplotype configurations acting together affect the phenotype are coded as follows:

Name	H^1_{j1} & H^2_{j1}	H^1_{j2} & H^2_{j2}
hap2comb_1	0 0	0 1
hap2comb_2	0 0	1 0
hap2comb_3	0 0	1 1
hap2comb_4	0 1	1 0
hap2comb_5	0 1	1 1
hap2comb_6	1 0	1 1

Table B.1: Haplotype combinations in the presence of 2 SNPs: H^1_{ji} and H^2_{ji} denote the alleles at the i^{th} haplotype ($i = 1, 2$) of the j^{th} individual ($j = 1, ..., n$) at the two SNPs in S.

Appendix C

Sum of two squared forms

Assuming $\boldsymbol{\Sigma}_1^{-1}$ and $\boldsymbol{\Sigma}_2^{-1}$ are symmetric matrices and $\boldsymbol{x}$, $\boldsymbol{m}_1$, $\boldsymbol{m}_2$ are vectors of the same length we have:

$$\begin{aligned} &-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{m}_1)^T\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x}-\boldsymbol{m}_1)-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{m}_2)^T\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x}-\boldsymbol{m}_2) \\ &=-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{m}_c)^T\boldsymbol{\Sigma}_c^{-1}(\boldsymbol{x}-\boldsymbol{m}_c)+c, \end{aligned} \tag{C.1}$$

where

$$\begin{aligned} \boldsymbol{m}_c &= (\boldsymbol{\Sigma}_1^{-1}+\boldsymbol{\Sigma}_2^{-1})^{-1}(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{m}_1+\boldsymbol{\Sigma}_2^{-1}\boldsymbol{m}_2); \\ \Sigma_c &= (\boldsymbol{\Sigma}_1^{-1}+\boldsymbol{\Sigma}_2^{-1})^{-1}; \\ c &= \frac{1}{2}(\boldsymbol{m}_1^T\boldsymbol{\Sigma}_1^{-1}+\boldsymbol{m}_2^T\boldsymbol{\Sigma}_2^{-1})(\boldsymbol{\Sigma}_1^{-1}+\boldsymbol{\Sigma}_2^{-1})^{-1}(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{m}_1+\boldsymbol{\Sigma}_2^{-1}\boldsymbol{m}_2) \\ &\quad -\frac{1}{2}(\boldsymbol{m}_1^T\boldsymbol{\Sigma}_1^{-1}\boldsymbol{m}_1+\boldsymbol{m}_2^T\boldsymbol{\Sigma}_2^{-1}\boldsymbol{m}_2). \end{aligned}$$

Appendix D

Fully Conditional Posterior Distributions of the Meta Analysis Models

In this appendix we present in detail the derivation of the fully conditional posterior distributions for the four meta analysis models that compose our approach. Let $S_1 = \{i : h_i = 1\}$ and $S_0 = \{i : h_i = 0\}$ as defined in Chapter 4.

D.1 Random Effects Model, All Cohorts Associated

$P(\boldsymbol{\beta}_i|\mu_1, ..., \mu_K, \boldsymbol{\beta}_{-i}, \boldsymbol{\beta}, \tau_1, ..., \tau_K, \lambda, \psi, \mathbf{Y}_1, ..., \mathbf{Y}_K)$: For each cohort we calculate the conditional posterior distribution of the parameter vector $\boldsymbol{\beta}_i$ given the rest of the parameters:

$$
\begin{aligned}
P(\boldsymbol{\beta}_i|\mu_1, ..., \mu_K, \boldsymbol{\beta}_{-i}, \boldsymbol{\beta}, \tau_1, ..., \tau_K, \lambda, \psi, \mathbf{Y}_1, ..., \mathbf{Y}_K) \propto & \qquad \text{(D.1)} \\
& \exp\{-\frac{\tau_i}{2}(\mathbf{Y}_i - \mu_i \mathbf{1}_{N_i} - \mathbf{G}_i\boldsymbol{\beta}_i)^T(\mathbf{Y}_i - \mu_i \mathbf{1}_{N_i} - \mathbf{G}_i\boldsymbol{\beta}_i)\} \\
& \times \exp\{-\frac{\lambda}{2}(\boldsymbol{\beta}_i - \boldsymbol{\beta})^T \mathbf{V}_p^{-1}(\boldsymbol{\beta}_i - \boldsymbol{\beta})\} \\
= \exp\{-\frac{1}{2}(\mathbf{Y}_i - \mu_i \mathbf{1}_{N_i} - \mathbf{G}_i\boldsymbol{\beta}_i)^T \mathbf{I}_{\tau_i}(\mathbf{Y}_i - \mu_i \mathbf{1}_{N_i} - \mathbf{G}_i\boldsymbol{\beta}_i) & \\
- \frac{1}{2}(\boldsymbol{\beta}_i - \boldsymbol{\beta})^T \lambda \mathbf{V}_p^{-1}(\boldsymbol{\beta}_i - \boldsymbol{\beta})\}, &
\end{aligned}
$$

where $\mathbf{I}_{\tau_i}$ is a diagonal matrix with non-zero entries equal to τ_i. Now,

$$P(\boldsymbol{\beta}_i|\mu_1,...,\mu_K,\boldsymbol{\beta}_{-i},\boldsymbol{\beta},\tau_1,...,\tau_K,\lambda,\psi,\mathbf{Y}_1,...,\mathbf{Y}_K) \propto \exp\{-\frac{1}{2}(\boldsymbol{\beta}_i-\boldsymbol{m}_{\boldsymbol{\beta_i}})^T\boldsymbol{\Sigma}_{\boldsymbol{\beta}_i}^{-1}(\boldsymbol{\beta}_i-\boldsymbol{m}_{\boldsymbol{\beta_i}})\} \tag{D.2}$$

$$\sim N(\boldsymbol{m}_{\boldsymbol{\beta_i}},\boldsymbol{\Sigma}_{\boldsymbol{\beta}_i}),$$

where

$$\begin{aligned}
\boldsymbol{m}_{\boldsymbol{\beta_i}} &= (\lambda\mathbf{V}_p^{-1}+\mathbf{G}_i^T\mathbf{I}_{\tau_i}\mathbf{G}_i)^{-1}(\lambda\mathbf{V}_p^{-1}\boldsymbol{\beta}+\mathbf{G}_i^T\mathbf{I}_{\tau_i}(\mathbf{Y}_i-\mu_i\mathbf{1}_{N_i}));\\
\boldsymbol{\Sigma}_{\boldsymbol{\beta}_i} &= (\lambda\mathbf{V}_p^{-1}+\mathbf{G}_i^T\mathbf{I}_{\tau_i}\mathbf{G}_i)^{-1};\\
c_{\boldsymbol{\beta}_i} &= \frac{1}{2}\{\boldsymbol{\beta}^T\lambda\mathbf{V}_p^{-1}+(\mathbf{Y}_i-\mu_i\mathbf{1}_{N_i})^T(\mathbf{G}_i^T\mathbf{I}_{\tau_i})^T\}(\lambda\mathbf{V}_p^{-1}+\mathbf{G}_i^T\mathbf{I}_{\tau_i}\mathbf{G}_i)^{-1}(\lambda\mathbf{V}_p^{-1}\boldsymbol{\beta}+\mathbf{G}_i^T\mathbf{I}_{\tau_i}(\mathbf{Y}_i-\mu_i\mathbf{1}_{N_i}))\\
&\quad -\frac{1}{2}\{\lambda\boldsymbol{\beta}^T\mathbf{V}_p^{-1}\boldsymbol{\beta}+\{(\mathbf{G}_i^T\mathbf{I}_{\tau_i}\mathbf{G}_i)^{-1}\mathbf{G}_i^T\mathbf{I}_{\tau_i}(\mathbf{Y}_i-\mu_i\mathbf{1}_{N_i})\}^T\mathbf{G}_i^T\mathbf{I}_{\tau_i}(\mathbf{Y}_i-\mu_i\mathbf{1}_{N_i})\}.
\end{aligned}$$

Note that $\boldsymbol{\beta}_{-i}=(\boldsymbol{\beta}_1,...,\boldsymbol{\beta}_{i-1},\boldsymbol{\beta}_{i+1},...,\boldsymbol{\beta}_K)$, is the set of parameter vectors not including the parameter vector for cohort i.

$\underline{P(\mu_i|\mu_{-i},\boldsymbol{\beta}_1,...,\boldsymbol{\beta}_K,\boldsymbol{\beta},\tau_1,...,\tau_K,\lambda,\psi,\mathbf{Y}_1,...,\mathbf{Y}_K):}$

$$\begin{aligned}
P(\mu_i|\mu_{-i},\boldsymbol{\beta}_1,...,\boldsymbol{\beta}_K,\boldsymbol{\beta},\tau_1,...,\tau_K,\lambda,\psi,\mathbf{Y}_1,...,\mathbf{Y}_K) &\propto \qquad (D.3)\\
&\exp\{-\frac{\tau_i}{2}(\mathbf{Y}_i-\mu_i\mathbf{1}_{N_i}-\mathbf{G}_i\boldsymbol{\beta}_i)^T(\mathbf{Y}_i-\mu_i\mathbf{1}_{N_i}-\mathbf{G}_i\boldsymbol{\beta}_i)\}\times\exp\{-\frac{\nu}{2}\mu_i^T\mu_i\}\\
&=\exp\{-\frac{1}{2}(\mu_i\mathbf{1}_{N_i}-(\mathbf{Y}_i-\mathbf{G}_i\boldsymbol{\beta}_i))^T\mathbf{I}_{\tau_i}(\mu_i\mathbf{1}_{N_i}-(\mathbf{Y}_i-\mathbf{G}_i\boldsymbol{\beta}_i))-\frac{1}{2}(\mu_i)^T\nu(\mu_i)\}.
\end{aligned}$$

Then,

$$P(\mu_i|\mu_{-i},\boldsymbol{\beta}_1,...,\boldsymbol{\beta}_K,\boldsymbol{\beta},\tau_1,...,\tau_K,\lambda,\psi,\mathbf{Y}_1,...,\mathbf{Y}_K) \propto \exp\{-\frac{1}{2}(\mu_i-\boldsymbol{m}_{\mu_i})^T\boldsymbol{\Sigma}_{\mu_i}^{-1}(\mu_i-\boldsymbol{m}_{\mu_i})\} \tag{D.4}$$

$$\sim N(\boldsymbol{m}_{\mu_i},\boldsymbol{\Sigma}_{\mu_i})$$

and

$$\begin{aligned}
\boldsymbol{m}_{\mu_i} &= \frac{\tau_i}{N_i\tau_i + \nu}\sum_{N_i}(\mathbf{Y}_i - \mathbf{G}_i\boldsymbol{\beta}_i);\\
\boldsymbol{\Sigma}_{\mu_i} &= \frac{1}{N_i\tau_i + \nu};\\
c_{\mu_i} &= \frac{1}{2}\left\{\frac{\tau_i^2}{N_i\tau_i + \nu}\left(\sum_{N_i}(\mathbf{Y}_i - \mathbf{G}_i\boldsymbol{\beta}_i)\right)^2\right\} - \frac{1}{2}\left\{\frac{\tau_i}{N_i}\left(\sum_{N_i}(\mathbf{Y}_i - \mathbf{G}_i\boldsymbol{\beta}_i)\right)^2\right\}.
\end{aligned}$$

$\underline{P(\boldsymbol{\beta}|\mu_1, ..., \mu_K, \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_K, \tau_1, ..., \tau_K, \lambda, \psi, \mathbf{Y}_1, ..., \mathbf{Y}_K)}$:

$$\begin{aligned}
P(\boldsymbol{\beta}|\mu_1, ..., \mu_K, \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_K, \tau_1, ..., \tau_K, \lambda, \psi, \mathbf{Y}_1, ..., \mathbf{Y}_K) &\propto \prod_{i=1}^{K}\left(\exp\{-\frac{\lambda}{2}(\boldsymbol{\beta}_i - \boldsymbol{\beta})^T\mathbf{V}_p^{-1}(\boldsymbol{\beta}_i - \boldsymbol{\beta})\}\right)\\
&\quad \times \exp\{-\frac{\psi}{2}\boldsymbol{\beta}^T\mathbf{W}_p^{-1}\boldsymbol{\beta}\}.
\end{aligned} \tag{D.5}$$

As above we make use of the sum of two squared forms leading to:

$$\begin{aligned}
P(\boldsymbol{\beta}|\mu_1, ..., \mu_K, \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_K, \tau_1, ..., \tau_K, \lambda, \psi, \mathbf{Y}_1, ..., \mathbf{Y}_K) &\propto \exp\{-\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{m}_{\boldsymbol{\beta}})^T\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}(\boldsymbol{\beta} - \boldsymbol{m}_{\boldsymbol{\beta}})\}\\
&\sim N(\boldsymbol{m}_{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}),
\end{aligned} \tag{D.6}$$

where

$$\begin{aligned}
\boldsymbol{m}_{\boldsymbol{\beta}} &= (\sum_{i=1}^{K}\lambda\mathbf{V_p}^{-1} + \psi\mathbf{W}_p^{-1})^{-1}\sum_{i=1}^{K}\lambda\mathbf{V_p}^{-1}\boldsymbol{\beta}_i;\\
\boldsymbol{\Sigma}_{\boldsymbol{\beta}} &= (\sum_{i=1}^{K}\lambda\mathbf{V_p}^{-1} + \psi\mathbf{W}_p^{-1})^{-1};\\
c_{\boldsymbol{\beta}} &= \frac{1}{2}\sum_{i=1}^{K}\boldsymbol{\beta}_i^T\lambda\mathbf{V_p}^{-1}(\sum_{i=1}^{K}\lambda\mathbf{V_p}^{-1} + \psi\mathbf{W}_p^{-1})^{-1}\left(\sum_{i=1}^{K}\lambda\mathbf{V_p}^{-1}\boldsymbol{\beta}_i\right) - \frac{1}{2}\sum_{i=1}^{K}\boldsymbol{\beta}_i^T\lambda\mathbf{V_p}^{-1}\boldsymbol{\beta}_i.
\end{aligned}$$

$\underline{P(\tau_i|\mu_1,...,\mu_K,\boldsymbol{\beta}_1,...,\boldsymbol{\beta}_K,\boldsymbol{\beta},\tau_{-i},\lambda,\psi,\mathbf{Y}_1,...,\mathbf{Y}_K)}$:

$$\begin{aligned} P(\tau_i|\mu_1,...,\mu_K,\boldsymbol{\beta}_1,...,\boldsymbol{\beta}_K,\boldsymbol{\beta},\tau_{-i},\lambda,\psi,\mathbf{Y}_1,...,\mathbf{Y}_K) &\propto \tau_i^{(N_i/2)+j-1}\times \\ &\exp\{-\frac{\tau_i}{2}\{(\mathbf{Y}_i-\mu_i\mathbf{1}_{N_i}-\mathbf{G}_i\boldsymbol{\beta}_i)^T(\mathbf{Y}_i-\mu_i\mathbf{1}_{N_i}-\mathbf{G}_i\boldsymbol{\beta}_i)+2l\}\} \\ &\propto \Gamma(j^*,l^*), \end{aligned} \tag{D.7}$$

where

$$\begin{aligned} j^* &= \frac{N_i}{2}+j; \\ l^* &= \frac{1}{2}(\mathbf{Y}_i-\mu_i\mathbf{1}_{N_i}-\mathbf{G}_i\boldsymbol{\beta}_i)^T(\mathbf{Y}_i-\mu_i\mathbf{1}_{N_i}-\mathbf{G}_i\boldsymbol{\beta}_i)+l. \end{aligned}$$

Note that $\tau_{-i}=(\tau_1,...,\tau_{i-1},\tau_{i+1},...,\tau_K)$, is the set of precisions not including the precision for cohort i.

$\underline{P(\lambda|\mu_1,...,\mu_K,\boldsymbol{\beta}_1,...,\boldsymbol{\beta}_K,\boldsymbol{\beta},\tau_1,...,\tau_K,\psi,\mathbf{Y}_1,...,\mathbf{Y}_K)}$:

$$\begin{aligned} P(\lambda|\mu_1,...,\mu_K,\boldsymbol{\beta}_1,...,\boldsymbol{\beta}_K,\boldsymbol{\beta},\tau_1,...,\tau_K,\psi,\mathbf{Y}_1,...,\mathbf{Y}_K) &\propto \prod_{i=1}^{K}\lambda^{p/2}\exp\{-\frac{\lambda}{2}(\boldsymbol{\beta}_i-\boldsymbol{\beta})^T\mathbf{V}_p^{-1}(\boldsymbol{\beta}_i-\boldsymbol{\beta})\} \\ &\times \lambda^{c-1}\exp\{-d\lambda\} \\ &= \lambda^{K(p/2)+c-1}\exp\{-\frac{\lambda}{2}\times \\ &\left(\sum_{i=1}^{K}(\boldsymbol{\beta}_i-\boldsymbol{\beta})^T\mathbf{V}_p^{-1}(\boldsymbol{\beta}_i-\boldsymbol{\beta})+2d\right)\} \\ &\propto \Gamma(c^*,d^*), \end{aligned} \tag{D.8}$$

where

$$c^* = K\frac{p}{2} + c;$$
$$d^* = \frac{1}{2}\sum_{i=1}^{K}(\boldsymbol{\beta}_i - \boldsymbol{\beta})^T\mathbf{V}_p^{-1}(\boldsymbol{\beta}_i - \boldsymbol{\beta}) + d.$$

$\underline{P(\psi|\mu_1, ..., \mu_K, \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_K, \boldsymbol{\beta}, \tau_1, ..., \tau_K, \lambda, \mathbf{Y}_1, ..., \mathbf{Y}_K)}$:

$$P(\psi|\mu_1, ..., \mu_K, \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_K, \boldsymbol{\beta}, \tau_1, ..., \tau_K, \lambda, \mathbf{Y}_1, ..., \mathbf{Y}_K) \propto \psi^{(p/2)+a-1}\exp\{-\frac{\psi}{2}(\boldsymbol{\beta}^T\mathbf{W}_p^{-1}\boldsymbol{\beta} + 2b)\} \tag{D.9}$$
$$\propto \Gamma(a^*, b^*),$$

where

$$a^* = \frac{p}{2} + a;$$
$$b^* = \frac{1}{2}(\boldsymbol{\beta}^T\mathbf{W}_p^{-1}\boldsymbol{\beta}) + b.$$

D.2 Fixed Effects Model, All Cohorts Associated

$\underline{P(\boldsymbol{\beta}_F|\mu_{F,1}, ..., \mu_{F,K}, \tau_{F,1}, ..., \tau_{F,K}, \psi_F, \mathbf{Y}_1, ..., \mathbf{Y}_K)}$:

$$P(\boldsymbol{\beta}_F|\mu_{F,1}, ..., \mu_{F,K}, \tau_{F,1}, ..., \tau_{F,K}, \psi, \mathbf{Y}_1, ..., \mathbf{Y}_K) \propto \tag{D.10}$$
$$\prod_{i=1}^{K}\exp\{-\frac{\tau_{F,i}}{2}(\mathbf{Y}_i - \mu_{F,i}\mathbf{1}_{N_i} - \mathbf{G}_i\boldsymbol{\beta}_F)^T(\mathbf{Y}_i - \mu_{F,i}\mathbf{1}_{N_i} - \mathbf{G}_i\boldsymbol{\beta}_F)\}$$
$$\times \exp\{-\frac{\psi_F}{2}\boldsymbol{\beta}_F^T\mathbf{W}_p^{-1}\boldsymbol{\beta}_F\}$$
$$= \exp\{-\frac{1}{2}\sum_{i=1}^{K}(\mathbf{Y}_i - \mu_{F,i}\mathbf{1}_{N_i} - \mathbf{G}_i\boldsymbol{\beta}_F)^T(\mathbf{Y}_i - \mu_{F,i}\mathbf{1}_{N_i} - \mathbf{G}_i\boldsymbol{\beta}_F)$$
$$- \frac{1}{2}\boldsymbol{\beta}_F^T(\psi_F\mathbf{W}_p^{-1})\boldsymbol{\beta}_F\}.$$

Then

$$P(\boldsymbol{\beta}_F|\mu_{F,1},...,\mu_{F,K},\tau_{F,1},...,\tau_{F,K},\psi,\mathbf{Y}_1,...,\mathbf{Y}_K) = N(m_{\boldsymbol{\beta}_F},\boldsymbol{\Sigma}_{\boldsymbol{\beta}_F}), \tag{D.11}$$

where

$$\begin{aligned}
m_{\boldsymbol{\beta}_F} &= \left(\sum_{i=1}^{K}\mathbf{G}_i^T\mathbf{I}_{\tau_{F,i}}\mathbf{G}_i + \psi_F\mathbf{W}_p^{-1}\right)^{-1}\left(\sum_{i=1}^{K}\mathbf{G}_i^T\mathbf{I}_{\tau_{F,i}}\left(\mathbf{Y}_i - \mu_{F,i}\mathbf{1}_{N_i}\right)\right)\\
\boldsymbol{\Sigma}_{\boldsymbol{\beta}_F} &= \left(\sum_{i=1}^{K}\mathbf{G}_i^T\mathbf{I}_{\tau_{F,i}}\mathbf{G}_i + \psi_F\mathbf{W}_p^{-1}\right)^{-1}\\
c_{\boldsymbol{\beta}_F} &= \frac{1}{2}\left(\sum_{i=1}^{K}\left(\mathbf{Y}_i - \mu_{F,i}\mathbf{1}_{N_i}\right)^T\left(\mathbf{G}_i^T\mathbf{I}_{\tau_{F,i}}\right)^T\right)\left(\sum_{i=1}^{K}\mathbf{G}_i^T\mathbf{I}_{\tau_{F,i}}\mathbf{G}_i + \psi_F\mathbf{W}_p^{-1}\right)^{-1}\\
&\quad\times\left(\sum_{i=1}^{K}\mathbf{G}_i^T\mathbf{I}_{\tau_{F,i}}\left(\mathbf{Y}_i - \mu_{F,i}\mathbf{1}_{N_i}\right)\right)\\
&\quad-\frac{1}{2}\left\{\sum_{i=1}^{K}\left((\mathbf{G}_i^T\mathbf{I}_{\tau_{F,i}}\mathbf{G}_i)^{-1}\mathbf{G}_i^T\mathbf{I}_{\tau_{F,i}}\left(\mathbf{Y}_i - \mu_{F,i}\mathbf{1}_{N_i}\right)\right)^T\mathbf{G}_i^T\mathbf{I}_{\tau_{F,i}}\left(\mathbf{Y}_i - \mu_{F,i}\mathbf{1}_{N_i}\right)\right\}.
\end{aligned}$$

$\underline{P(\mu_{F,i}|\mu_{F,-i},\boldsymbol{\beta}_F,\tau_{F,1},...,\tau_{F,K},\psi_F,\mathbf{Y}_1,...,\mathbf{Y}_K)}$:

$$\begin{aligned}
P(\mu_{F,i}|\mu_{F,-i},\boldsymbol{\beta}_F,\tau_{F,1},...,\tau_{F,K},\psi_F,\mathbf{Y}_1,...,\mathbf{Y}_K) &\propto \\
\exp\{-\frac{\tau_{F,i}}{2}(\mathbf{Y}_i - \mu_{F,i}\mathbf{1}_{N_i} - \mathbf{G}_i\boldsymbol{\beta}_F)^T(\mathbf{Y}_i - \mu_{F,i}\mathbf{1}_{N_i} - \mathbf{G}_i\boldsymbol{\beta}_F)\}&\\
\times\nu^{1/2}\exp\{-\frac{\nu}{2}\mu_{F,i}^2\}.&
\end{aligned} \tag{D.12}$$

Hence,

$$P(\mu_{F,i}|\mu_{F,-i},\boldsymbol{\beta}_F,\tau_{F,1},...,\tau_{F,K},\psi_F,\mathbf{Y}_1,...,\mathbf{Y}_K) = N(m_{\mu_{F,i}},\boldsymbol{\Sigma}_{\mu_{F,i}}), \tag{D.13}$$

$$\boldsymbol{m}_{\mu_{F,i}} = \frac{\tau_{F,i}}{N_i\tau_{F,i}+\nu}\sum_{N_i}(\mathbf{Y}_i - \mathbf{G}_i\boldsymbol{\beta}_F);$$
$$\boldsymbol{\Sigma}_{\mu_{F,i}} = \frac{1}{N_i\tau_{F,i}+\nu};$$
$$c_{\mu_{F,i}} = \frac{1}{2}\left\{\frac{\tau_{F,i}^2}{N_i\tau_{F,i}+\nu}\left(\sum_{N_i}(\mathbf{Y}_i - \mathbf{G}_i\boldsymbol{\beta}_F)\right)^2\right\} - \frac{1}{2}\left\{\frac{\tau_{F,i}}{N_i}\left(\sum_{N_i}(\mathbf{Y}_i - \mathbf{G}_i\boldsymbol{\beta}_F)\right)^2\right\}.$$

$\underline{P(\tau_{F,i}|\mu_{F,1},...,\mu_{F,K},\boldsymbol{\beta}_F,\tau_{F,-i},\psi_F,\mathbf{Y}_1,...,\mathbf{Y}_K)}$:

$$P(\tau_{F,i}|\mu_{F,1},...,\mu_{F,K},\boldsymbol{\beta}_F,\tau_{F,-i},\psi_F,\mathbf{Y}_1,...,\mathbf{Y}_K) \propto \tau_{F,i}^{N_i/2+j_F-1} \times \exp\{-\frac{\tau_{F,i}}{2}(\mathbf{Y}_i - \mu_{F,i}\mathbf{1}_{N_i} - \mathbf{G}_i\boldsymbol{\beta}_F)^T(\mathbf{Y}_i - \mu_{F,i}\mathbf{1}_{N_i} - \mathbf{G}_i\boldsymbol{\beta}_F) + 2l_F\} \propto \Gamma(j_F^*, l_F^*), \tag{D.14}$$

with

$$j_F^* = \frac{N_i}{2} + j_F;$$
$$l_F = \frac{1}{2}(\mathbf{Y}_i - \mu_{F,i}\mathbf{1}_{N_i} - \mathbf{G}_i\boldsymbol{\beta}_F)^T(\mathbf{Y}_i - \mu_{F,i}\mathbf{1}_{N_i} - \mathbf{G}_i\boldsymbol{\beta}_F) + l_F.$$

$\underline{P(\psi_F|\mu_{F,1},...,\mu_{F,K},\boldsymbol{\beta}_F,\tau_{F,1},...,\tau_{F,K},\mathbf{Y}_1,...,\mathbf{Y}_K)}$:

$$P(\psi_F|\mu_{F,1},...,\mu_{F,K},\boldsymbol{\beta}_F,\tau_{F,1},...,\tau_{F,K},\mathbf{Y}_1,...,\mathbf{Y}_K) \propto \psi_F^{(p/2)+a_F-1} \exp\{-\frac{\psi_F}{2}(\boldsymbol{\beta}_F^T\mathbf{W}_p^{-1}\boldsymbol{\beta}_F + 2b_F)\} \propto \Gamma(a_F^*, b_F^*), \tag{D.15}$$

with

$$a_F^* = \frac{p}{2} + a_F;$$
$$b_F^* = \frac{1}{2}\boldsymbol{\beta}_F^T \mathbf{W}_p^{-1} \boldsymbol{\beta}_F + b_F.$$

D.3 Random Effect Model, Subset of Cohorts Associated

$\underline{P(\boldsymbol{\beta}_i|\mu_1, ..., \mu_K, \boldsymbol{\beta}_{-i}, \boldsymbol{\beta}, \tau_1, ..., \tau_K, \lambda, \psi, h_1, ..., h_K, \pi, \mathbf{Y}_1, ..., \mathbf{Y}_K)}$:

$$\begin{aligned} P(\boldsymbol{\beta}_i|\mu_1, ..., \mu_K, \boldsymbol{\beta}_{-i}, \boldsymbol{\beta}, \tau_1, ..., \tau_K, \lambda, \psi, h_1, ..., h_K, \pi, \mathbf{Y}_1, ..., \mathbf{Y}_K) \propto & \\ \exp\{-\frac{\tau_i}{2}(\mathbf{Y}_i - \mu_i \mathbf{1}_{N_i} - \mathbf{G}_i\boldsymbol{\beta}_i)^T(\mathbf{Y}_i - \mu_i \mathbf{1}_{N_i} - \mathbf{G}_i\boldsymbol{\beta}_i)\} & \\ \times \exp\{-\frac{\lambda}{2}(\boldsymbol{\beta}_i - \boldsymbol{\beta})^T \mathbf{V}_p^{-1}(\boldsymbol{\beta}_i - \boldsymbol{\beta})\} & \\ = \exp\{-\frac{1}{2}(\mathbf{Y}_i - \mu_i \mathbf{1}_{N_i} - \mathbf{G}_i\boldsymbol{\beta}_i)^T \mathbf{I}_{\tau_i}(\mathbf{Y}_i - \mu_i \mathbf{1}_{N_i} - \mathbf{G}_i\boldsymbol{\beta}_i) & \\ -\frac{1}{2}(\boldsymbol{\beta}_i - \boldsymbol{\beta})^T \lambda \mathbf{V}_p^{-1}(\boldsymbol{\beta}_i - \boldsymbol{\beta})\}, & \end{aligned} \tag{D.16}$$

where $\mathbf{I}_{\tau_i}$ is a diagonal matrix with non-zero entries equal to τ_i. Now,

$$\begin{aligned} P(\boldsymbol{\beta}_i|\mu_1, ..., \mu_K, \boldsymbol{\beta}_{-i}, \boldsymbol{\beta}, \tau_1, ..., \tau_K, \lambda, \psi, h_1, ..., h_K, \pi, \mathbf{Y}_1, ..., \mathbf{Y}_K) \propto & \\ \exp\{-\frac{1}{2}(\boldsymbol{\beta}_i - \boldsymbol{m}_{\boldsymbol{\beta}_i})^T \boldsymbol{\Sigma}_{\boldsymbol{\beta}_i}^{-1}(\boldsymbol{\beta}_i - \boldsymbol{m}_{\boldsymbol{\beta}_i})\} & \\ \sim N(\boldsymbol{m}_{\boldsymbol{\beta}_i}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}_i}), & \end{aligned} \tag{D.17}$$

where

$$\begin{aligned}
\boldsymbol{m}_{\boldsymbol{\beta_i}} &= (\lambda \mathbf{V}_p^{-1} + \mathbf{G}_i^T \mathbf{I}_{\tau_i} \mathbf{G}_i)^{-1} (\lambda \mathbf{V}_p^{-1} \boldsymbol{\beta} + \mathbf{G}_i^T \mathbf{I}_{\tau_i} (\mathbf{Y}_i - \mu_i \mathbf{1}_{N_i})); \\
\boldsymbol{\Sigma}_{\boldsymbol{\beta}_i} &= (\lambda \mathbf{V}_p^{-1} + \mathbf{G}_i^T \mathbf{I}_{\tau_i} \mathbf{G}_i)^{-1}; \\
c_{\boldsymbol{\beta}_i} &= \frac{1}{2} \{ \boldsymbol{\beta}^T \lambda \mathbf{V}_p^{-1} + (\mathbf{Y}_i - \mu_i \mathbf{1}_{N_i})^T (\mathbf{G}_i^T \mathbf{I}_{\tau_i})^T \} (\lambda \mathbf{V}_p^{-1} + \mathbf{G}_i^T \mathbf{I}_{\tau_i} \mathbf{G}_i)^{-1} \\
&\quad \times (\lambda \mathbf{V}_p^{-1} \boldsymbol{\beta} + \mathbf{G}_i^T \mathbf{I}_{\tau_i} (\mathbf{Y}_i - \mu_i \mathbf{1}_{N_i})) \\
&\quad - \frac{1}{2} \{ \lambda \boldsymbol{\beta}^T \mathbf{V}_p^{-1} \boldsymbol{\beta} + \{ (\mathbf{G}_i^T \mathbf{I}_{\tau_i} \mathbf{G}_i)^{-1} \mathbf{G}_i^T \mathbf{I}_{\tau_i} (\mathbf{Y}_i - \mu_i \mathbf{1}_{N_i}) \}^T \mathbf{G}_i^T \mathbf{I}_{\tau_i} (\mathbf{Y}_i - \mu_i \mathbf{1}_{N_i}) \}.
\end{aligned}$$

$P(\mu_i | \mu_{-i}, \boldsymbol{\beta}_{i:i \in S_1}, \boldsymbol{\beta}, \tau_1, ..., \tau_K, \lambda, \psi, h_1, ..., h_K, \pi, \mathbf{Y}_1, ..., \mathbf{Y}_K)$, where $i \in S_1$:

$$\begin{aligned}
&P(\mu_i | \mu_{-i}, \boldsymbol{\beta}_{i:i \in S_1}, \boldsymbol{\beta}, \tau_1, ..., \tau_K, \lambda, \psi, h_1, ..., h_K, \pi, \mathbf{Y}_1, ..., \mathbf{Y}_K) \propto \qquad \text{(D.18)} \\
&\exp\{ -\frac{\tau_i}{2} (\mathbf{Y}_i - \mu_i \mathbf{1}_{N_i} - \mathbf{G}_i \boldsymbol{\beta}_i)^T (\mathbf{Y}_i - \mu_i \mathbf{1}_{N_i} - \mathbf{G}_i \boldsymbol{\beta}_i) \} \times \exp\{ -\frac{\nu}{2} \mu_i^T \mu_i \} \\
&= \exp\{ -\frac{1}{2} (\mathbf{Y}_i - \mu_i \mathbf{1}_{N_i} - \mathbf{G}_i \boldsymbol{\beta}_i)^T \mathbf{I}_{\tau_i} (\mathbf{Y}_i - \mu_i \mathbf{1}_{N_i} - \mathbf{G}_i \boldsymbol{\beta}_i) - \frac{1}{2} (\mu_i)^T \nu (\mu_i) \} \\
&\propto N(\boldsymbol{m}_{\mu_i}, \boldsymbol{\Sigma}_{\mu_i}).
\end{aligned}$$

The values of $\boldsymbol{m}_{\mu_i}$ and $\boldsymbol{\Sigma}_{\mu_i}$ are as follows:

$$\begin{aligned}
\boldsymbol{m}_{\mu_i} &= \frac{\tau_i}{N_i \tau_i + \nu} \sum_{N_i} (\mathbf{Y}_i - \mathbf{G}_i \boldsymbol{\beta}_i); \\
\boldsymbol{\Sigma}_{\mu_i} &= \frac{1}{N_i \tau_i + \nu}; \\
c_{\mu_i} &= \frac{1}{2} \left\{ \frac{\tau_i^2}{N_i \tau_i + \nu} \left(\sum_{N_i} (\mathbf{Y}_i - \mathbf{G}_i \boldsymbol{\beta}_i) \right)^2 \right\} - \frac{1}{2} \left\{ \frac{\tau_i}{N_i} \left(\sum_{N_i} (\mathbf{Y}_i - \mathbf{G}_i \boldsymbol{\beta}_i) \right)^2 \right\}.
\end{aligned}$$

$\underline{P(\mu_i|\mu_{-i}, \boldsymbol{\beta}_{i:i\in S_1}, \boldsymbol{\beta}, \tau_1, ..., \tau_K, \lambda, \psi, h_1, ..., h_K, \pi, \mathbf{Y}_1, ..., \mathbf{Y}_K)}$, where $i \in S_0$:

$$\begin{aligned} P(\mu_i|\mu_{-i}, \boldsymbol{\beta}_{i:i\in S_1}, \boldsymbol{\beta}, \tau_1, ..., \tau_K, \lambda, \psi, h_1, ..., h_K, \pi, \mathbf{Y}_1, ..., \mathbf{Y}_K) &\propto \\ \exp\{-\frac{\tau_i}{2}(\mathbf{Y}_i - \mu_i \mathbf{1}_{N_i})^T(\mathbf{Y}_i - \mu_i \mathbf{1}_{N_i})\} &\times \exp\{-\frac{\nu}{2}\mu_i^T \mu_i\} \\ = \exp\{-\frac{1}{2}(\mathbf{Y}_i - \mu_i \mathbf{1}_{N_i})^T \mathbf{I}_{\tau_i}(\mathbf{Y}_i - \mu_i \mathbf{1}_{N_i}) - \frac{1}{2}(\mu_i)^T \nu(\mu_i)\} & \\ \propto N(\boldsymbol{m}_{\mu_i}, \boldsymbol{\Sigma}_{\mu_i}), & \end{aligned} \tag{D.19}$$

with

$$\begin{aligned} \boldsymbol{m}_{\mu_i} &= \frac{\tau_i}{N_i\tau_i + \nu} \sum_{N_i} \mathbf{Y}_i; \\ \boldsymbol{\Sigma}_{\mu_i} &= \frac{1}{N_i\tau_i + \nu}; \\ c_{\mu_i} &= \frac{1}{2}\left\{ \frac{\tau_i^2}{N_i\tau_i + \nu} \left(\sum_{N_i} \mathbf{Y}_i \right)^2 \right\} - \frac{1}{2}\left\{ \frac{\tau_i}{N_i} \left(\sum_{N_i} \mathbf{Y}_i \right)^2 \right\}. \end{aligned}$$

$\underline{P(\boldsymbol{\beta}|\mu_1, ..., \mu_K, \boldsymbol{\beta}_{i:i\in S_1}, \tau_1, ..., \tau_K, \lambda, \psi, h_1, ..., h_K, \pi, \mathbf{Y}_1, ..., \mathbf{Y}_K)}$:

$$\begin{aligned} P(\boldsymbol{\beta}|\mu_1, ..., \mu_K, \boldsymbol{\beta}_{i:i\in S_1}, \tau_1, ..., \tau_K, \lambda, \psi, h_1, ..., h_K, \pi, \mathbf{Y}_1, ..., \mathbf{Y}_K) &\propto \\ \prod_{i\in S_1} \left(\exp\{-\frac{\lambda}{2}(\boldsymbol{\beta}_i - \boldsymbol{\beta})^T \mathbf{V}_p^{-1}(\boldsymbol{\beta}_i - \boldsymbol{\beta})\} \right) &\times \exp\{-\frac{\psi}{2}\boldsymbol{\beta}^T \mathbf{W}_p^{-1}\boldsymbol{\beta}\} \\ \sim N(\boldsymbol{m}_{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}), & \end{aligned} \tag{D.20}$$

where

$$\begin{aligned} \boldsymbol{m}_{\boldsymbol{\beta}} &= (\sum_{i\in S_1} \lambda \mathbf{V_p}^{-1} + \psi \mathbf{W}_p^{-1})^{-1} \sum_{i\in S_1} \lambda \mathbf{V_p}^{-1} \boldsymbol{\beta}_i; \\ \boldsymbol{\Sigma}_{\boldsymbol{\beta}} &= (\sum_{i\in S_1} \lambda \mathbf{V_p}^{-1} + \psi \mathbf{W}_p^{-1})^{-1}; \\ c_{\boldsymbol{\beta}} &= \frac{1}{2} \sum_{i\in S_1} \boldsymbol{\beta}_i^T \lambda \mathbf{V_p}^{-1} (\sum_{i\in S_1} \lambda \mathbf{V_p}^{-1} + \psi \mathbf{W}_p^{-1})^{-1} \left(\sum_{i\in S_1} \lambda \mathbf{V_p}^{-1} \boldsymbol{\beta}_i \right) - \frac{1}{2} \sum_{i\in S_1} \boldsymbol{\beta}_i^T \lambda \mathbf{V_p}^{-1} \boldsymbol{\beta}_i. \end{aligned}$$

$\underline{P(\tau_i|\mu_1,...,\mu_K,\boldsymbol{\beta}_{i:i\in S_1},\boldsymbol{\beta},\tau_{-i},\lambda,\psi,h_1,...,h_K,\pi,\mathbf{Y}_1,...,\mathbf{Y}_K)}$, where $i \in S_1$:

$$P(\tau_i|\mu_1,...,\mu_K,\boldsymbol{\beta}_{i:i\in S_1},\boldsymbol{\beta},\tau_{-i},\lambda,\psi,h_1,...,h_K,\pi,\mathbf{Y}_1,...,\mathbf{Y}_K) \propto \tag{D.21}$$
$$\tau_i^{(N_i/2)+j-1} \times \exp\{-\frac{\tau_i}{2}\{(\mathbf{Y}_i - \mu_i\mathbf{1}_{N_i} - \mathbf{G}_i\boldsymbol{\beta}_i)^T(\mathbf{Y}_i - \mu_i\mathbf{1}_{N_i} - \mathbf{G}_i\boldsymbol{\beta}_i) + 2l\}\}$$
$$\propto \Gamma(j^*, l^*),$$

where

$$j^* = \frac{N_i}{2} + j;$$
$$l^* = \frac{1}{2}(\mathbf{Y}_i - \mu_i\mathbf{1}_{N_i} - \mathbf{G}_i\boldsymbol{\beta}_i)^T(\mathbf{Y}_i - \mu_i\mathbf{1}_{N_i} - \mathbf{G}_i\boldsymbol{\beta}_i) + l.$$

$\underline{P(\tau_i|\mu_1,...,\mu_K,\boldsymbol{\beta}_{i:i\in S_1},\boldsymbol{\beta},\tau_{-i},\lambda,\psi,h_1,...,h_K,\pi,\mathbf{Y}_1,...,\mathbf{Y}_K)}$, where $i \in S_0$:

$$P(\tau_i|\mu_1,...,\mu_K,\boldsymbol{\beta}_{i:i\in S_1},\boldsymbol{\beta},\tau_{-i},\lambda,\psi,h_1,...,h_K,\pi,\mathbf{Y}_1,...,\mathbf{Y}_K) \propto \tag{D.22}$$
$$\tau_i^{(N_i/2)+j-1} \times \exp\{-\frac{\tau_i}{2}\{(\mathbf{Y}_i - \mu_i\mathbf{1}_{N_i})^T(\mathbf{Y}_i - \mu_i\mathbf{1}_{N_i}) + 2l\}\}$$
$$\propto \Gamma(j^*, l^*),$$

where

$$j^* = \frac{N_i}{2} + j;$$
$$l^* = \frac{1}{2}(\mathbf{Y}_i - \mu_i\mathbf{1}_{N_i})^T(\mathbf{Y}_i - \mu_i\mathbf{1}_{N_i}) + l.$$

$\underline{P(\lambda|\mu_1,...,\mu_K,\boldsymbol{\beta}_{i:i\in S_1},\boldsymbol{\beta},\tau_1,...,\tau_K,\psi,h_1,...,h_K,\pi,\mathbf{Y}_1,...,\mathbf{Y}_K)}$:

$$P(\lambda|\mu_1,...,\mu_K,\boldsymbol{\beta}_{i:i\in S_1},\boldsymbol{\beta},\tau_1,...,\tau_K,\psi,h_1,...,h_K,\pi,\mathbf{Y}_1,...,\mathbf{Y}_K) \propto \tag{D.23}$$
$$\prod_{i\in S_1}\lambda^{p/2}\exp\{-\frac{\lambda}{2}(\boldsymbol{\beta}_i-\boldsymbol{\beta})^T\mathbf{V}_p^{-1}(\boldsymbol{\beta}_i-\boldsymbol{\beta})\}\times\lambda^{c-1}\exp\{-d\lambda\}$$
$$=\lambda^{K_1(p/2)+c-1}\exp\{-\frac{\lambda}{2}\times(\sum_{i\in S_1}(\boldsymbol{\beta}_i-\boldsymbol{\beta})^T\mathbf{V}_p^{-1}(\boldsymbol{\beta}_i-\boldsymbol{\beta})+2d)\}$$
$$\propto\Gamma(c^*,d^*),$$

where

$$c^*=K_1\frac{p}{2}+c;$$
$$d^*=\frac{1}{2}\sum_{i\in S_1}(\boldsymbol{\beta}_i-\boldsymbol{\beta})^T\mathbf{V}_p^{-1}(\boldsymbol{\beta}_i-\boldsymbol{\beta})+d.$$

$P(\psi|\mu_1,...,\mu_K,\boldsymbol{\beta}_{i:i\in S_1},\boldsymbol{\beta},\tau_1,...,\tau_K,\lambda,h_1,...,h_K,\pi,\mathbf{Y}_1,...,\mathbf{Y}_K)$:

$$P(\psi|\mu_1,...,\mu_K,\boldsymbol{\beta}_{i:i\in S_1},\boldsymbol{\beta},\tau_1,...,\tau_K,\lambda,h_1,...,h_K,\pi,\mathbf{Y}_1,...,\mathbf{Y}_K) \propto \tag{D.24}$$
$$\psi^{(p/2)+a-1}\exp\{-\frac{\psi}{2}(\boldsymbol{\beta}^T\mathbf{W}_p^{-1}\boldsymbol{\beta}+2b)\}$$
$$\propto\Gamma(a^*,b^*),$$

where

$$a^*=\frac{p}{2}+a;$$
$$b^*=\frac{1}{2}(\boldsymbol{\beta}^T\mathbf{W}_p^{-1}\boldsymbol{\beta})+b.$$

$P(\pi|\mu_1,...,\mu_K,\boldsymbol{\beta}_{i:i\in S_1},\boldsymbol{\beta},\tau_1,...,\tau_K,\lambda,\psi,h_1,...,h_K,\mathbf{Y}_1,...,\mathbf{Y}_K)$:

$$P(\pi|\mu_1,...,\mu_K,\boldsymbol{\beta}_{i:i\in S_1},\boldsymbol{\beta},\tau_1,...,\tau_K,\lambda,\psi,h_1,...,h_K,\mathbf{Y}_1,...,\mathbf{Y}_K) \propto \tag{D.25}$$

$$\prod_{i\in S_1}\pi\prod_{i\in S_0}(1-\pi)\times\pi^{e-1}(1-\pi)^{f-1}$$

$$=\pi^{K_1+e-1}(1-\pi)^{K_0+f-1}$$

$$\propto Beta(e^*,f^*),$$

where

$$e^*=K_1+e;$$

$$f^*=K_0+f.$$

$\underline{P(h_i|\mu_1,...,\mu_K,\boldsymbol{\beta}_{i:i\in S_1},\boldsymbol{\beta},\tau_1,...,\tau_K,\lambda,\psi,h_{-i},\mathbf{Y}_1,...,\mathbf{Y}_K)}$:

$$P(h_i|\mu_1,...,\mu_K,\boldsymbol{\beta}_{i:i\in S_1},\boldsymbol{\beta},\tau_1,...,\tau_K,\lambda,\psi,h_{-i},\mathbf{Y}_1,...,\mathbf{Y}_K)\propto \tag{D.26}$$

$$\left(N(\mu_i+\mathbf{G}_i\boldsymbol{\beta}_i,\tau_i^{-1}\mathbf{I}_{N_i})N(\boldsymbol{\beta}_i;\boldsymbol{\beta},\lambda^{-1}\mathbf{V}_p)N(\boldsymbol{\beta};\mathbf{0},\psi^{-1}\mathbf{W}_p)\Gamma(\lambda;c,d)\right)^{h_i}\left(N(\mu_i,\tau_i^{-1}\mathbf{I}_{N_i})\right)^{1-h_i}$$

$$\times\pi^{h_i}(1-\pi)^{1-h_i}$$

$$\propto\left(\pi N(\mu_i+\mathbf{G}_i\boldsymbol{\beta}_i,\tau_i^{-1}\mathbf{I}_{N_i})N(\boldsymbol{\beta},\lambda^{-1}\mathbf{V}_p)N(\mathbf{0},\psi^{-1}\mathbf{W}_p)\Gamma(c,d)\right)^{h_i}\left((1-\pi)N(\mu_i,\tau_i^{-1}\mathbf{I}_{N_i})\right)^{1-h_i}$$

$$\propto Bernoulli(\pi^*),$$

where

$$\pi^* = \frac{\tilde{\pi}}{\tilde{\pi} + \tilde{q}};$$
$$\tilde{\pi} = \pi \exp\{-\frac{\tau_i}{2}(\mathbf{Y}_i - \mu_i \mathbf{1}_{N_i} - \mathbf{G}_i\boldsymbol{\beta}_i)^T(\mathbf{Y}_i - \mu_i \mathbf{1}_{N_i} - \mathbf{G}_i\boldsymbol{\beta}_i)\} N(\boldsymbol{\beta}_i; \boldsymbol{\beta}, \lambda^{-1}\mathbf{V}_p)$$
$$N(\boldsymbol{\beta}; \mathbf{0}, \psi^{-1}\mathbf{W}_p)\Gamma(\lambda; c, d);$$
$$\tilde{q} = (1-\pi)\exp\{-\frac{\tau_i}{2}(\mathbf{Y}_i - \mu_i \mathbf{1}_{N_i})^T(\mathbf{Y}_i - \mu_i \mathbf{1}_{N_i})\}.$$

D.4 Fixed Effects Model, Subset of Cohorts Associated

$\underline{P(\boldsymbol{\beta}_F|\mu_{F,1}, ..., \mu_{F,K}, \tau_{F,1}, ..., \tau_{F,K}, \psi_F, \mathbf{Y}_1, ..., \mathbf{Y}_K)}$:

$$P(\boldsymbol{\beta}_F|\mu_{F,1}, ..., \mu_{F,K}, \tau_{F,1}, ..., \tau_{F,K}, \psi, h_1, ..., h_K, \pi, \mathbf{Y}_1, ..., \mathbf{Y}_K) \propto \tag{D.27}$$
$$\prod_{i \in S_1} \exp\{-\frac{\tau_{F,i}}{2}(\mathbf{Y}_i - \mu_{F,i}\mathbf{1}_{N_i} - \mathbf{G}_i\boldsymbol{\beta}_F)^T(\mathbf{Y}_i - \mu_{F,i}\mathbf{1}_{N_i} - \mathbf{G}_i\boldsymbol{\beta}_F)\}$$
$$\times \exp\{-\frac{\psi_F}{2}\boldsymbol{\beta}_F^T\mathbf{W}_p^{-1}\boldsymbol{\beta}_F\}$$
$$= \exp\{-\frac{1}{2}\sum_{i \in S_1}(\mathbf{Y}_i - \mu_{F,i}\mathbf{1}_{N_i} - \mathbf{G}_i\boldsymbol{\beta}_F)^T(\mathbf{Y}_i - \mu_{F,i}\mathbf{1}_{N_i} - \mathbf{G}_i\boldsymbol{\beta}_F)$$
$$- \frac{1}{2}\boldsymbol{\beta}_F^T(\psi_F\mathbf{W}_p^{-1})\boldsymbol{\beta}_F\}$$
$$\propto N(m_{\boldsymbol{\beta}_F}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}_F}),$$

where,

$$
\begin{aligned}
m_{\boldsymbol{\beta}_F} &= \left(\sum_{i\in S_1}\mathbf{G}_i^T\mathbf{I}_{\tau_{F,i}}\mathbf{G}_i + \psi_F\mathbf{W}_p^{-1}\right)^{-1}\left(\sum_{i\in S_1}\mathbf{G}_i^T\mathbf{I}_{\tau_{F,i}}\left(\mathbf{Y}_i - \mu_{F,i}\mathbf{1}_{N_i}\right)\right)\\
\boldsymbol{\Sigma}_{\boldsymbol{\beta}_F} &= \left(\sum_{i\in S_1}\mathbf{G}_i^T\mathbf{I}_{\tau_{F,i}}\mathbf{G}_i + \psi_F\mathbf{W}_p^{-1}\right)^{-1}\\
c_{\boldsymbol{\beta}_F} &= \frac{1}{2}\left(\sum_{i\in S_1}\left(\mathbf{Y}_i - \mu_{F,i}\mathbf{1}_{N_i}\right)^T\left(\mathbf{G}_i^T\mathbf{I}_{\tau_{F,i}}\right)^T\right)\left(\sum_{i\in S_1}\mathbf{G}_i^T\mathbf{I}_{\tau_{F,i}}\mathbf{G}_i + \psi_F\mathbf{W}_p^{-1}\right)^{-1}\\
&\quad\times\left(\sum_{i\in S_1}\mathbf{G}_i^T\mathbf{I}_{\tau_{F,i}}\left(\mathbf{Y}_i - \mu_{F,i}\mathbf{1}_{N_i}\right)\right)\\
&\quad-\frac{1}{2}\left\{\sum_{i\in S_1}\left((\mathbf{G}_i^T\mathbf{I}_{\tau_{F,i}}\mathbf{G}_i)^{-1}\mathbf{G}_i^T\mathbf{I}_{\tau_{F,i}}\left(\mathbf{Y}_i - \mu_{F,i}\mathbf{1}_{N_i}\right)\right)^T\mathbf{G}_i^T\mathbf{I}_{\tau_{F,i}}\left(\mathbf{Y}_i - \mu_{F,i}\mathbf{1}_{N_i}\right)\right\}.
\end{aligned}
$$

$P(\mu_{F,i}|\mu_{F,-i},\boldsymbol{\beta}_F,\tau_{F,1},...,\tau_{F,K},\psi_F,h_1,...,h_K,\pi,\mathbf{Y}_1,...,\mathbf{Y}_K)$, where $i\in S_1$:

$$
P(\mu_{F,i}|\mu_{F,-i},\boldsymbol{\beta}_F,\tau_{F,1},...,\tau_{F,K},\psi_F,h_1,...,h_K,\pi,\mathbf{Y}_1,...,\mathbf{Y}_K)\propto \tag{D.28}
$$

$$
\begin{aligned}
&\exp\{-\frac{\tau_{F,i}}{2}(\mathbf{Y}_i-\mu_{F,i}\mathbf{1}_{N_i}-\mathbf{G}_i\boldsymbol{\beta}_F)^T(\mathbf{Y}_i-\mu_{F,i}\mathbf{1}_{N_i}-\mathbf{G}_i\boldsymbol{\beta}_F)\}\times\exp\{-\frac{\nu}{2}\mu_{F,i}^T\mu_{F,i}\}\\
=&\exp\{-\frac{1}{2}(\mathbf{Y}_i-\mu_{F,i}\mathbf{1}_{N_i}-\mathbf{G}_i\boldsymbol{\beta}_F)^T\mathbf{I}_{\tau_{F,i}}(\mathbf{Y}_i-\mu_{F,i}\mathbf{1}_{N_i}-\mathbf{G}_i\boldsymbol{\beta}_F)-\frac{1}{2}(\mu_{F,i})^T\nu(\mu_{F,i})\}\\
&\propto N(\boldsymbol{m}_{\mu_{F,i}},\boldsymbol{\Sigma}_{\mu_{F,i}}).
\end{aligned}
$$

The values of $\boldsymbol{m}_{\mu_{F,i}}$ and $\boldsymbol{\Sigma}_{\mu_{F,i}}$ are as follows:

$$
\begin{aligned}
\boldsymbol{m}_{\mu_{F,i}} &= \frac{\tau_{F,i}}{N_i\tau_{F,i}+\nu}\sum_{N_i}(\mathbf{Y}_i-\mathbf{G}_i\boldsymbol{\beta}_F);\\
\boldsymbol{\Sigma}_{\mu_{F,i}} &= \frac{1}{N_i\tau_{F,i}+\nu};\\
c_{\mu_{F,i}} &= \frac{1}{2}\left\{\frac{\tau_{F,i}^2}{N_i\tau_{F,i}+\nu}\left(\sum_{N_i}(\mathbf{Y}_i-\mathbf{G}_i\boldsymbol{\beta}_F)\right)^2\right\}-\frac{1}{2}\left\{\frac{\tau_{F,i}}{N_i}\left(\sum_{N_i}(\mathbf{Y}_i-\mathbf{G}_i\boldsymbol{\beta}_F)\right)^2\right\}.
\end{aligned}
$$

$\underline{P(\mu_{F,i}|\mu_{F,-i}, \boldsymbol{\beta}_F, \tau_{F,1}, ..., \tau_{F,K}, \psi_F, h_1, ..., h_K, \pi, \mathbf{Y}_1, ..., \mathbf{Y}_K)}$, where $i \in S_0$:

$$\begin{aligned} P(\mu_{F,i}|\mu_{F,-i}, \boldsymbol{\beta}_F, \tau_{F,1}, ..., \tau_{F,K}, \psi_F, h_1, ..., h_K, \pi, \mathbf{Y}_1, ..., \mathbf{Y}_K) \propto \\ \exp\{-\frac{\tau_{F,i}}{2}(\mathbf{Y}_i - \mu_{F,i}\mathbf{1}_{N_i})^T(\mathbf{Y}_i - \mu_{F,i}\mathbf{1}_{N_i})\} \times \exp\{-\frac{\nu}{2}\mu_{F,i}^T\mu_{F,i}\} \\ = \exp\{-\frac{1}{2}(\mathbf{Y}_i - \mu_{F,i}\mathbf{1}_{N_i})^T\mathbf{I}_{\tau_{F,i}}(\mathbf{Y}_i - \mu_{F,i}\mathbf{1}_{N_i}) - \frac{1}{2}(\mu_{F,i})^T\nu(\mu_{F,i})\} \\ \propto N(\boldsymbol{m}_{\mu_{F,i}}, \boldsymbol{\Sigma}_{\mu_{F,i}}), \end{aligned} \tag{D.29}$$

with

$$\begin{aligned} \boldsymbol{m}_{\mu_{F,i}} &= \frac{\tau_{F,i}}{N_i\tau_{F,i} + \nu}\sum_{N_i}\mathbf{Y}_i; \\ \boldsymbol{\Sigma}_{\mu_{F,i}} &= \frac{1}{N_i\tau_{F,i} + \nu}; \\ c_{\mu_{F,i}} &= \frac{1}{2}\left\{\frac{\tau_{F,i}^2}{N_i\tau_{F,i} + \nu}\left(\sum_{N_i}\mathbf{Y}_i\right)^2\right\} - \frac{1}{2}\left\{\frac{\tau_{F,i}}{N_i}\left(\sum_{N_i}\mathbf{Y}_i\right)^2\right\}. \end{aligned}$$

$\underline{P(\tau_{F,i}|\mu_{F,1}, ..., \mu_{F,K}, \boldsymbol{\beta}_F, \tau_{F,-i}, \psi_F, h_1, ..., h_K, \pi, \mathbf{Y}_1, ..., \mathbf{Y}_K)}$, where $i \in S_1$:

$$\begin{aligned} P(\tau_{F,i}|\mu_{F,1}, ..., \mu_{F,K}, \boldsymbol{\beta}_F, \tau_{F,-i}, \psi_F, h_1, ..., h_K, \pi, \mathbf{Y}_1, ..., \mathbf{Y}_K) \propto \\ \tau_{F,i}^{(N_i/2)+j-1} \times \exp\{-\frac{\tau_{F,i}}{2}\{(\mathbf{Y}_i - \mu_{F,i}\mathbf{1}_{N_i} - \mathbf{G}_i\boldsymbol{\beta}_F)^T(\mathbf{Y}_i - \mu_{F,i}\mathbf{1}_{N_i} - \mathbf{G}_i\boldsymbol{\beta}_F) + 2l_F\}\} \\ \propto \Gamma(j_F^*, l_F^*), \end{aligned} \tag{D.30}$$

where

$$\begin{aligned} j_F^* &= \frac{N_i}{2} + j_F; \\ l_F^* &= \frac{1}{2}(\mathbf{Y}_i - \mu_{F,i}\mathbf{1}_{N_i} - \mathbf{G}_i\boldsymbol{\beta}_F)^T(\mathbf{Y}_i - \mu_{F,i}\mathbf{1}_{N_i} - \mathbf{G}_i\boldsymbol{\beta}_F) + l_F. \end{aligned}$$

$\underline{P(\tau_{F,i}|\mu_{F,1}, ..., \mu_{F,K}, \boldsymbol{\beta}_F, \tau_{F,-i}, \psi_F, h_1, ..., h_K, \pi, \mathbf{Y}_1, ..., \mathbf{Y}_K)}$, where $i \in S_0$:

$$P(\tau_{F,i}|\mu_{F,1},...,\mu_{F,K},\boldsymbol{\beta}_F,\tau_{F,-i},\psi_F,h_1,...,h_K,\pi,\mathbf{Y}_1,...,\mathbf{Y}_K) \propto \tag{D.31}$$
$$\tau_{F,i}^{(N_i/2)+j-1} \times \exp\{-\frac{\tau_{F,i}}{2}\{(\mathbf{Y}_i-\mu_{F,i}\mathbf{1}_{N_i})^T(\mathbf{Y}_i-\mu_{F,i}\mathbf{1}_{N_i})+2l_F\}\}$$
$$\propto \Gamma(j_F^*,l_F^*),$$

where

$$j_F^* = \frac{N_i}{2}+j_F;$$
$$l_F^* = \frac{1}{2}(\mathbf{Y}_i-\mu_{F,i}\mathbf{1}_{N_i})^T(\mathbf{Y}_i-\mu_{F,i}\mathbf{1}_{N_i})+l_F.$$

$\underline{P(\psi_F|\mu_{F,1},...,\mu_{F,K},\boldsymbol{\beta}_F,\tau_{F,1},...,\tau_{F,K},h_1,...,h_K,\pi,\mathbf{Y}_1,...,\mathbf{Y}_K)}$:

$$P(\psi_F|\mu_{F,1},...,\mu_{F,K},\boldsymbol{\beta}_F,\tau_{F,1},...,\tau_{F,K},h_1,...,h_K,\pi,\mathbf{Y}_1,...,\mathbf{Y}_K) \propto (\psi_F)^{(p/2)+a_F-1} \tag{D.32}$$
$$\times \exp\{-\frac{\psi_F}{2}(\boldsymbol{\beta}_F^T\mathbf{W}_p^{-1}\boldsymbol{\beta}_F+2b_F)\}$$
$$\propto \Gamma(a_F^*,b_F^*),$$

where

$$a_F^* = \frac{p}{2}+a_F;$$
$$b_F^* = \frac{1}{2}(\boldsymbol{\beta}_F^T\mathbf{W}_p^{-1}\boldsymbol{\beta}_F)+b_F.$$

$\underline{P(\pi|\mu_{F,1},...,\mu_{F,K},\boldsymbol{\beta}_F,\tau_{F,1},...,\tau_{F,K},\psi_F,h_1,...,h_K,\mathbf{Y}_1,...,\mathbf{Y}_K)}$:

$$P(\pi|\mu_{F,1},...,\mu_{F,K},\boldsymbol{\beta}_F,\tau_{F,1},...,\tau_{F,K},\psi_F,h_1,...,h_K,\mathbf{Y}_1,...,\mathbf{Y}_K) \propto \tag{D.33}$$
$$\prod_{i\in S_1}\pi\prod_{i\in S_0}(1-\pi)\times\pi^{e-1}(1-\pi)^{f-1}$$
$$=\pi^{K_1+e-1}(1-\pi)^{K_0+f-1}$$
$$\propto Beta(e^*,f^*),$$

where

$$e^* = K_1 + e;$$
$$f^* = K_0 + f.$$

$\underline{P(h_i|\mu_{F,1},...,\mu_{F,K},\boldsymbol{\beta}_F,\tau_{F,1},...,\tau_{F,K},\psi_F,h_{-i},\mathbf{Y}_1,...,\mathbf{Y}_K)}$:

$$P(h_i|\mu_{F,1},...,\mu_{F,K},\boldsymbol{\beta}_F,\tau_{F,1},...,\tau_{F,K},\psi_F,h_{-i},\mathbf{Y}_1,...,\mathbf{Y}_K) \propto \tag{D.34}$$
$$\left(N(\mu_{F,i}+\mathbf{G}_i\boldsymbol{\beta}_F,\tau_{F,i}^{-1}\mathbf{I}_{N_i})\right)^{h_i}\left(N(\mu_{F,i},\tau_{F,i}^{-1}\mathbf{I}_{N_i})\right)^{1-h_i}\times\pi^{h_i}(1-\pi)^{1-h_i}$$
$$\propto\left(\pi N(\mu_{F,i}+\mathbf{G}_i\boldsymbol{\beta}_F,\tau_{F,i}^{-1}\mathbf{I}_{N_i})\right)^{h_i}\left((1-\pi)N(\mu_{F,i},\tau_{F,i}^{-1}\mathbf{I}_{N_i})\right)^{1-h_i}$$
$$\propto Bernoulli(\pi^*),$$

where

$$\pi^* = \frac{\tilde{\pi}}{\tilde{\pi}+\tilde{q}};$$
$$\tilde{\pi} = \pi\exp\{-\frac{\tau_{F,i}}{2}(\mathbf{Y}_i-\mu_{F,i}-\mathbf{G}_i\boldsymbol{\beta}_F)^T(\mathbf{Y}_i-\mu_{F,i}\mathbf{G}_i\boldsymbol{\beta}_F)\};$$
$$\tilde{q} = (1-\pi)\exp\{-\frac{\tau_{F,i}}{2}(\mathbf{Y}_i-\mu_{F,i})^T(\mathbf{Y}_i-\mu_{F,i})\}.$$

Appendix E

Additional Simulation Results for the Fine Mapping Model

This appendix presents the simulation results for the detection of compound heterozygosity effects when $\beta_1 = 0.20$ and $\beta_2 = 0$ in Equation 3.1. As a general remark **FineMap MCMC** outperforms the other methods just as we observe in the case of a larger effect size, both in the detection of marginal effects and in the detection of the causal pair.

When the flanking sites LD is low, the only case where the **FineMap MCMC** suffers some power loss is when the causal SNPs are in high LD. Although the true positive rates are generally lower, **BIMBAM** does not manage to surpass **FineMap MCMC** apart from the afforementioned case. When we account for higher levels of correlation in the flanking sites, the power to identify one of the causal SNPs is higher for our fine mapping model. On the other hand, when it comes to the power to detect the causal pair of SNPs, we manage to outperform **BIMBAM** only when we account for pairs in high correlation with the actual pair. This is no surprise since the MCMC algorithm frequently "visits" SNPs in high correlation with true underlying variants as there is no way to distinguish between the two, especially if they are in perfect LD.

Moreover, we observe that the power to detect the true genetic model is larger than for $\beta_1 =$

0.30 although we observe the same ranking in performance of the causal LD structures. The high causal LD cases appear to have the lowest true positive rates out of the three considered levels of between causal SNPs correlation.

Finally, we also examine the speed of the MCMC chain to convergence using the Kullback-Leibler divergence measure and obtain the same results as for $\beta_1 = 0.30$. On the other hand, when examining the sensitivity of the model to prior assumptions we observe smaller KL estimates under the low causal/med flanking scenario than for the other two cases. Nonetheless, all cases perform equally well when considering larger values of p.

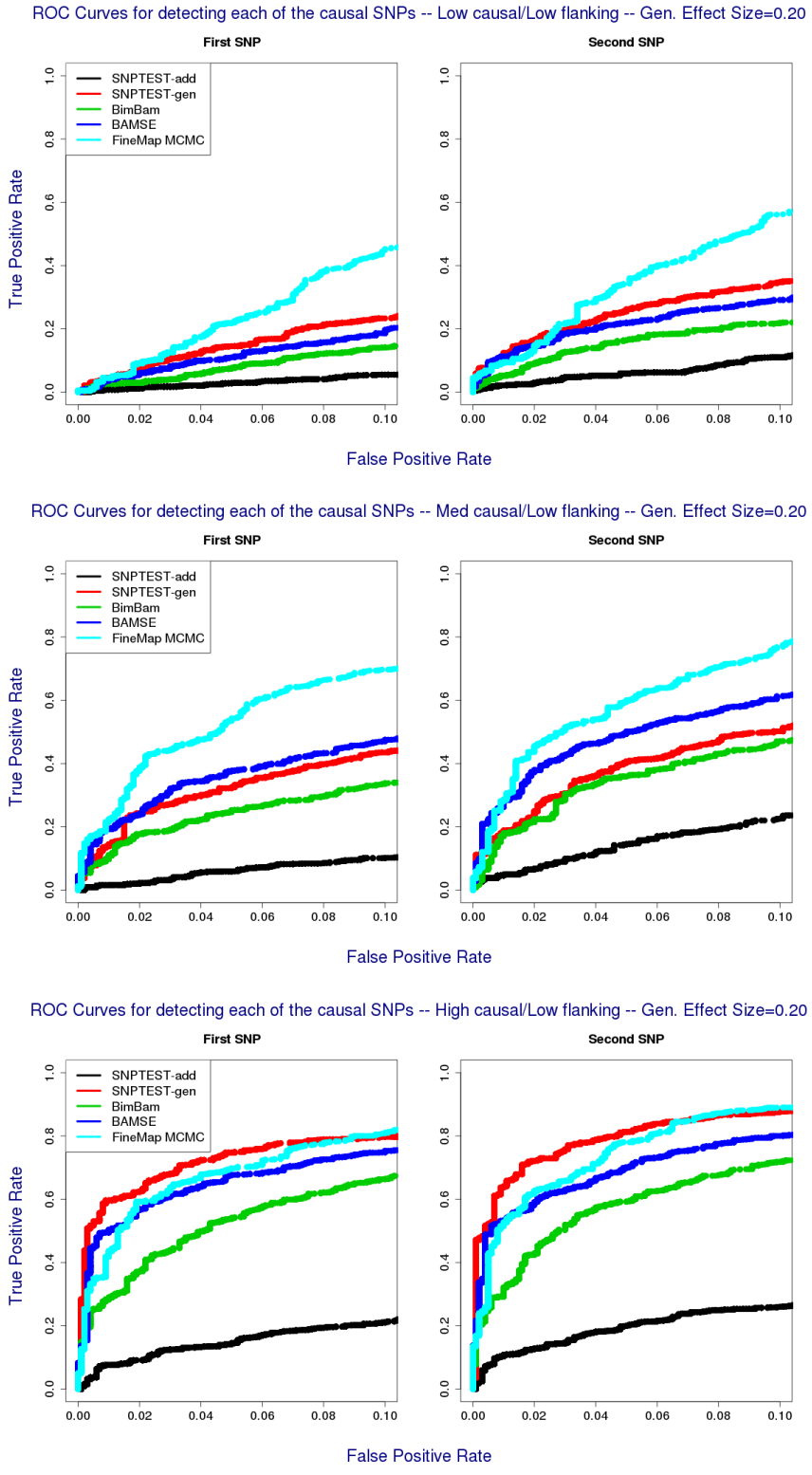

Figure E.1: ROC curves to identify the marginal effects of the true causal variants for the different method examined. The horizontal axes represent the false positive rates while the vertical axes represent the true positive rates. The different methods are coded using the following colours: **SNPTEST-Additive** (black), **SNPTEST-General** (red), **BIMBAM** (green), **BAMSE** (blue), **FineMap MCMC** (cyan) (LD: low, med, high causal/low flanking).

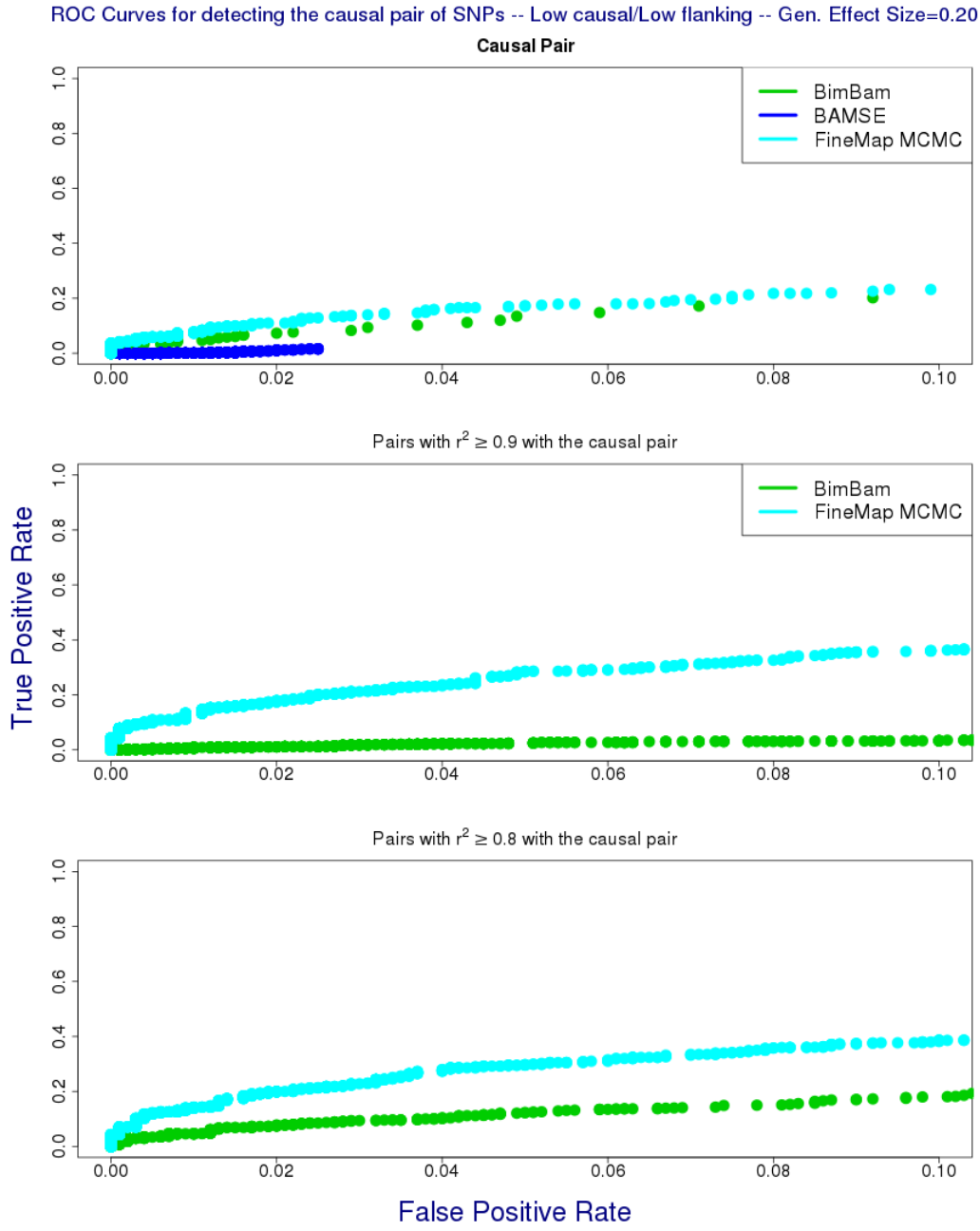

Figure E.2: ROC curves to detect the causal pair of variants (**BAMSE**, **BIMBAM** and **FineMap MCMC**) and to detect the pairs in high LD ($r^2 = 0.80$ and 0.90) with the causal combination of SNPs (**BIMBAM** and **FineMap MCMC**). The horizontal axes represent the false positive rates while the vertical axes represent the true positive rates. The different methods are coded using the following colours: **BIMBAM** (green), **BAMSE** (blue), **FineMap MCMC** (cyan) (LD: low causal/low flanking).

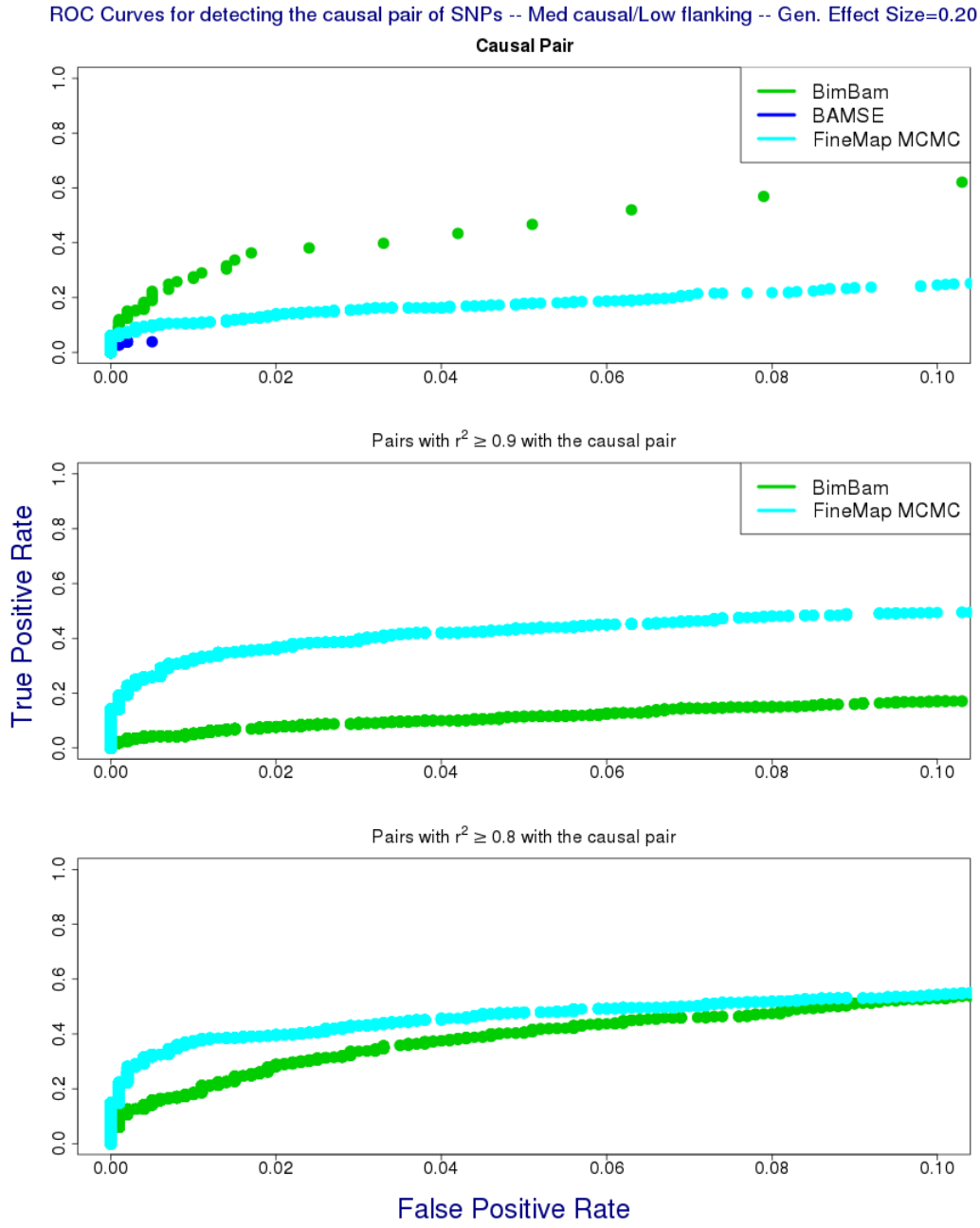

Figure E.3: ROC curves to detect the causal pair of variants (**BAMSE**, **BIMBAM** and **FineMap MCMC**) and to detect the pairs in high LD ($r^2 = 0.80$ and 0.90) with the causal combination of SNPs (**BIMBAM** and **FineMap MCMC**). The horizontal axes represent the false positive rates while the vertical axes represent the true positive rates. The different methods are coded using the following colours: **BIMBAM** (green), **BAMSE** (blue), **FineMap MCMC** (cyan) (LD: med causal/low flanking).

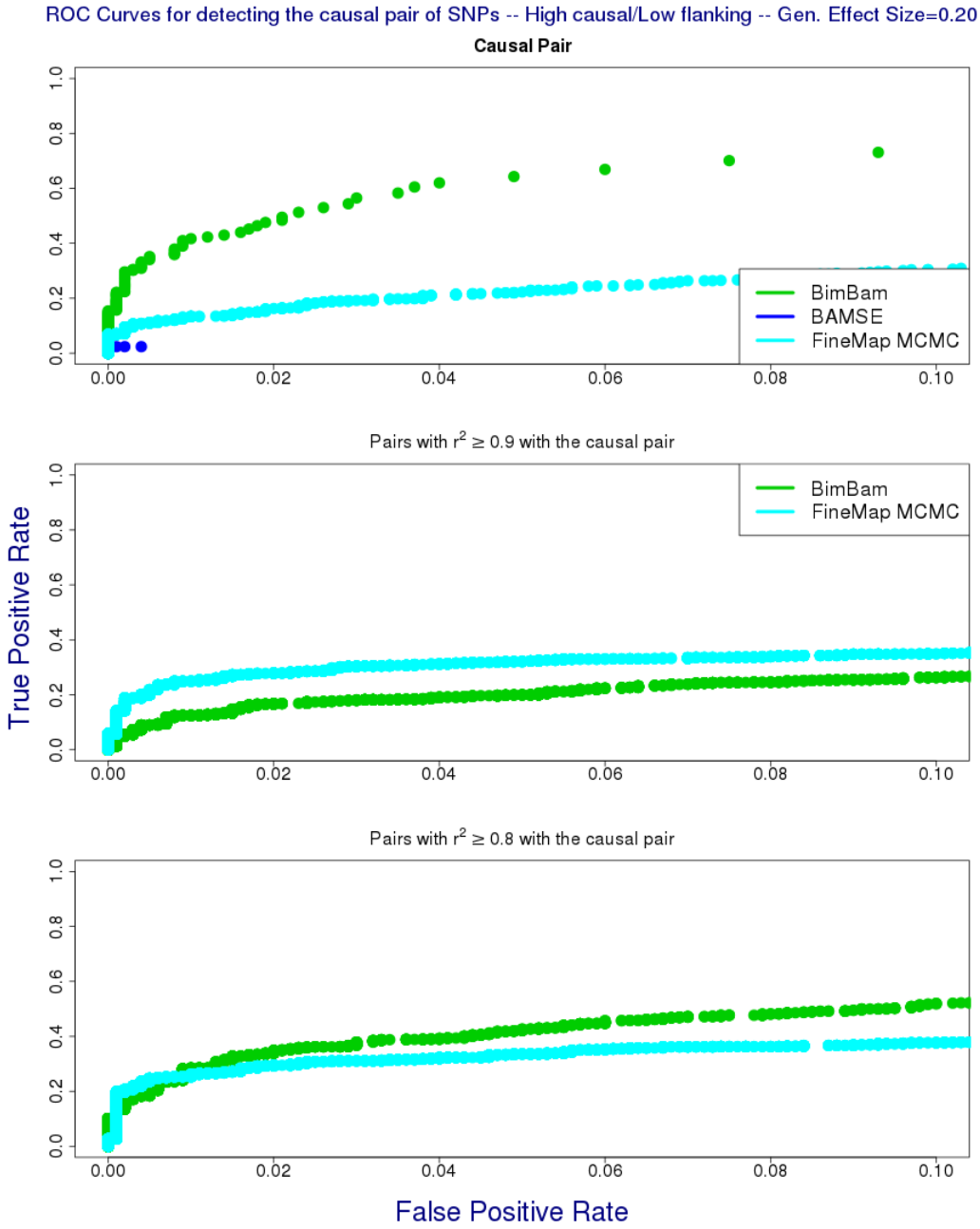

Figure E.4: ROC curves to detect the causal pair of variants (**BAMSE**, **BIMBAM** and **FineMap MCMC**) and to detect the pairs in high LD ($r^2 = 0.80$ and 0.90) with the causal combination of SNPs (**BIMBAM** and **FineMap MCMC**). The horizontal axes represent the false positive rates while the vertical axes represent the true positive rates. The different methods are coded using the following colours: **BIMBAM** (green), **BAMSE** (blue), **FineMap MCMC** (cyan) (LD: high causal/low flanking).

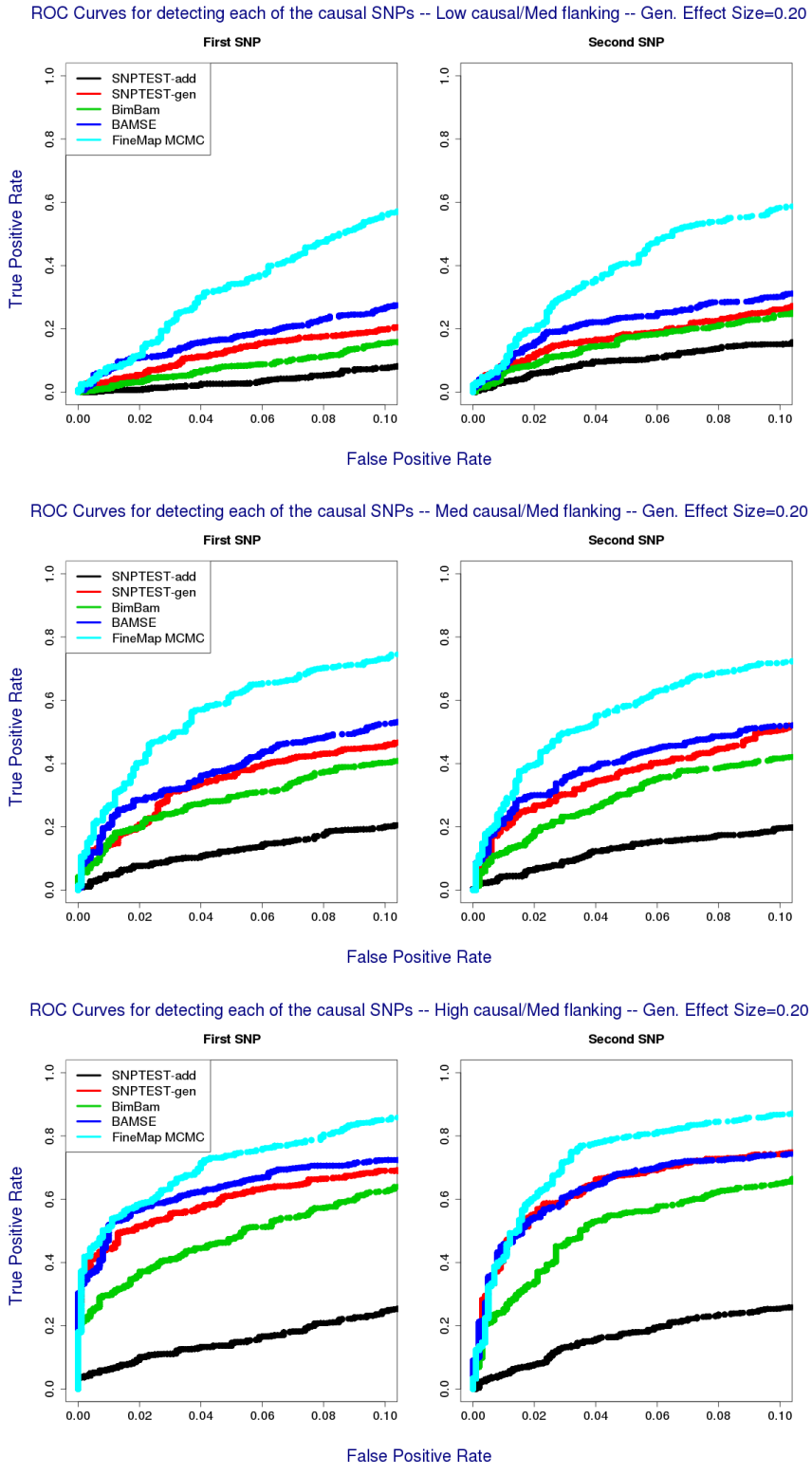

Figure E.5: ROC curves to identify the marginal effects of the true causal variants for the different method examined. The horizontal axes represent the false positive rates while the vertical axes represent the true positive rates. The different methods are coded using the following colours: **SNPTEST-Additive** (black), **SNPTEST-General** (red), **BIMBAM** (green), **BAMSE** (blue), **FineMap MCMC** (cyan) (LD: low, med, high causal/med flanking).

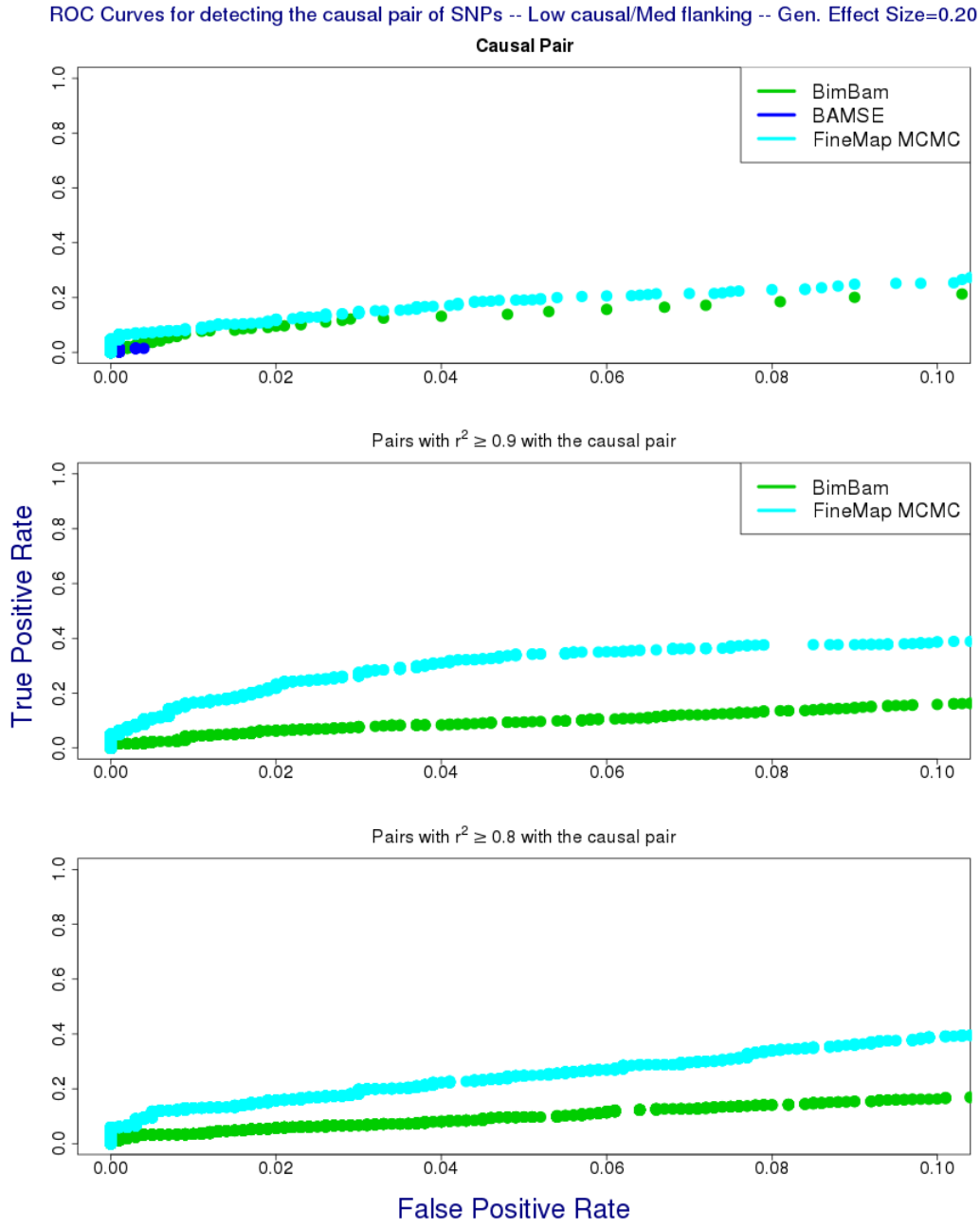

Figure E.6: ROC curves to detect the causal pair of variants (**BAMSE**, **BIMBAM** and **FineMap MCMC**) and to detect the pairs in high LD ($r^2 = 0.80$ and 0.90) with the causal combination of SNPs (**BIMBAM** and **FineMap MCMC**). The horizontal axes represent the false positive rates while the vertical axes represent the true positive rates. The different methods are coded using the following colours: **BIMBAM** (green), **BAMSE** (blue), **FineMap MCMC** (cyan) (LD: low causal/med flanking).

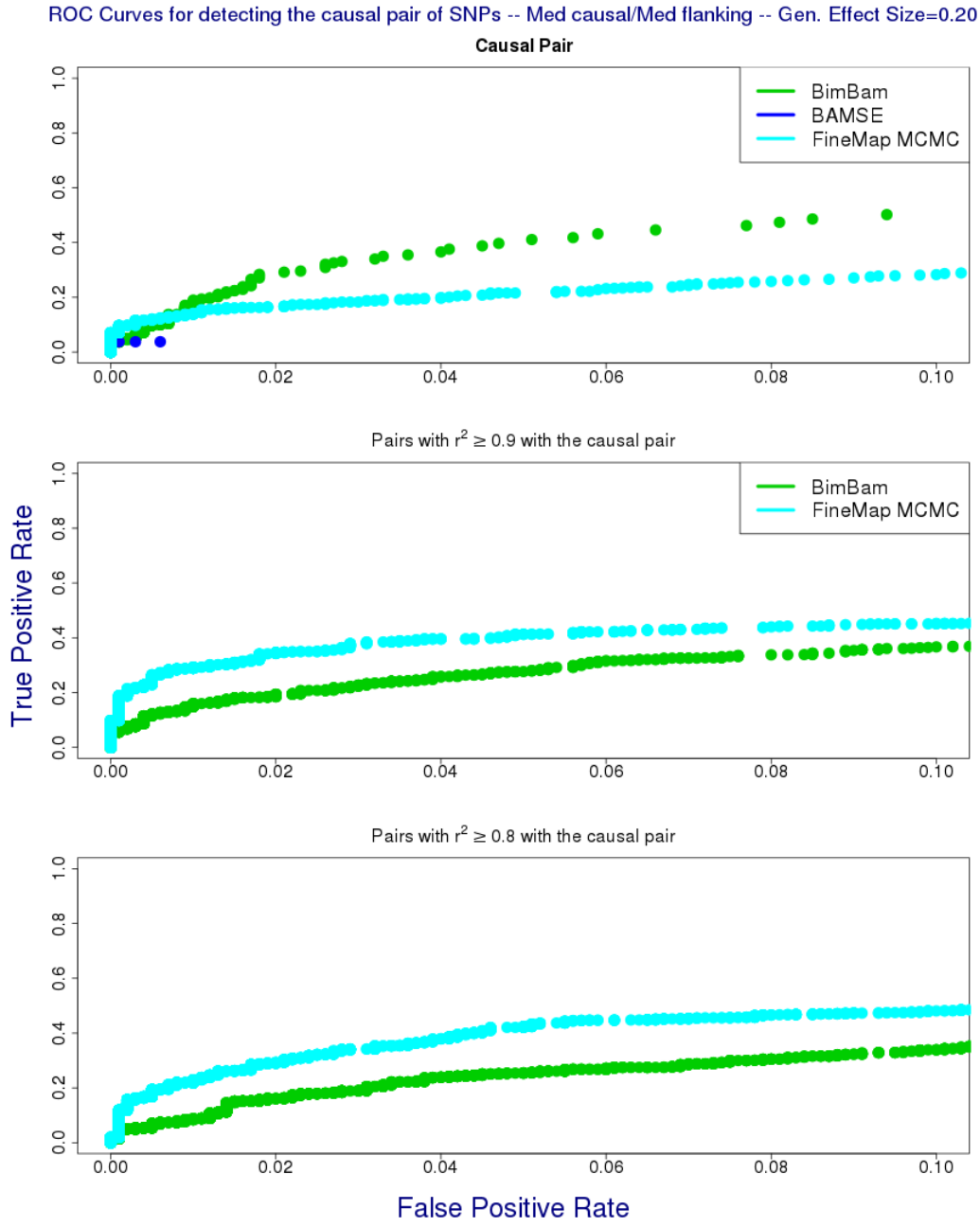

Figure E.7: ROC curves to detect the causal pair of variants (**BAMSE**, **BIMBAM** and **FineMap MCMC**) and to detect the pairs in high LD ($r^2 = 0.80$ and 0.90) with the causal combination of SNPs (**BIMBAM** and **FineMap MCMC**). The horizontal axes represent the false positive rates while the vertical axes represent the true positive rates. The different methods are coded using the following colours: **BIMBAM** (green), **BAMSE** (blue), **FineMap MCMC** (cyan) (LD: med causal/med flanking).

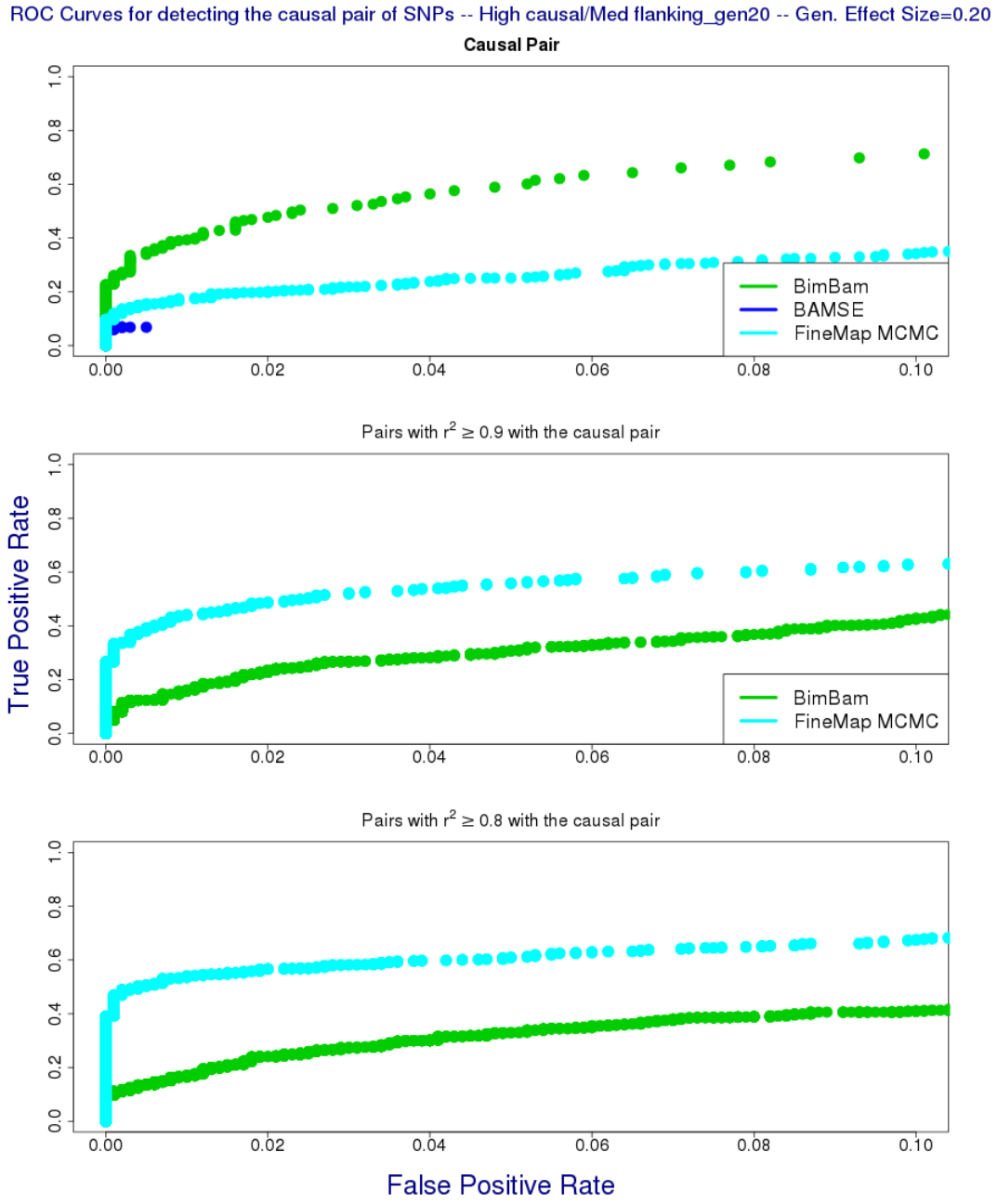

Figure E.8: ROC curves to detect the causal pair of variants (**BAMSE**, **BIMBAM** and **FineMap MCMC**) and to detect the pairs in high LD ($r^2 = 0.80$ and 0.90) with the causal combination of SNPs (**BIMBAM** and **FineMap MCMC**). The horizontal axes represent the false positive rates while the vertical axes represent the true positive rates. The different methods are coded using the following colours: **BIMBAM** (green), **BAMSE** (blue), **FineMap MCMC** (cyan) (LD: high causal/med flanking).

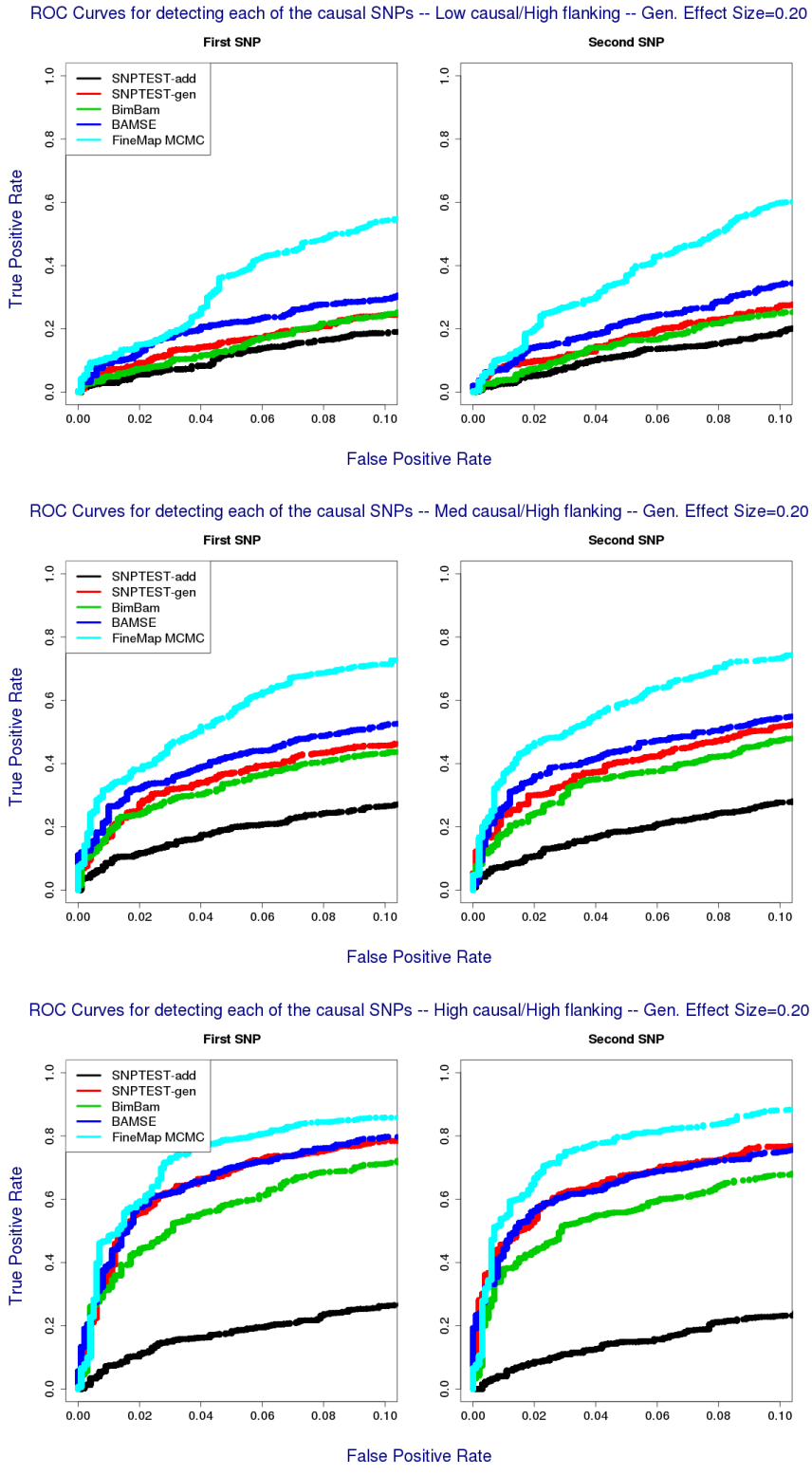

Figure E.9: ROC curves to identify the marginal effects of the true causal variants for the different method examined. The horizontal axes represent the false positive rates while the vertical axes represent the true positive rates. The different methods are coded using the following colours: **SNPTEST-Additive** (black), **SNPTEST-General** (red), **BIMBAM** (green), **BAMSE** (blue), **FineMap MCMC** (cyan) (LD: low, med, high causal/high flanking).

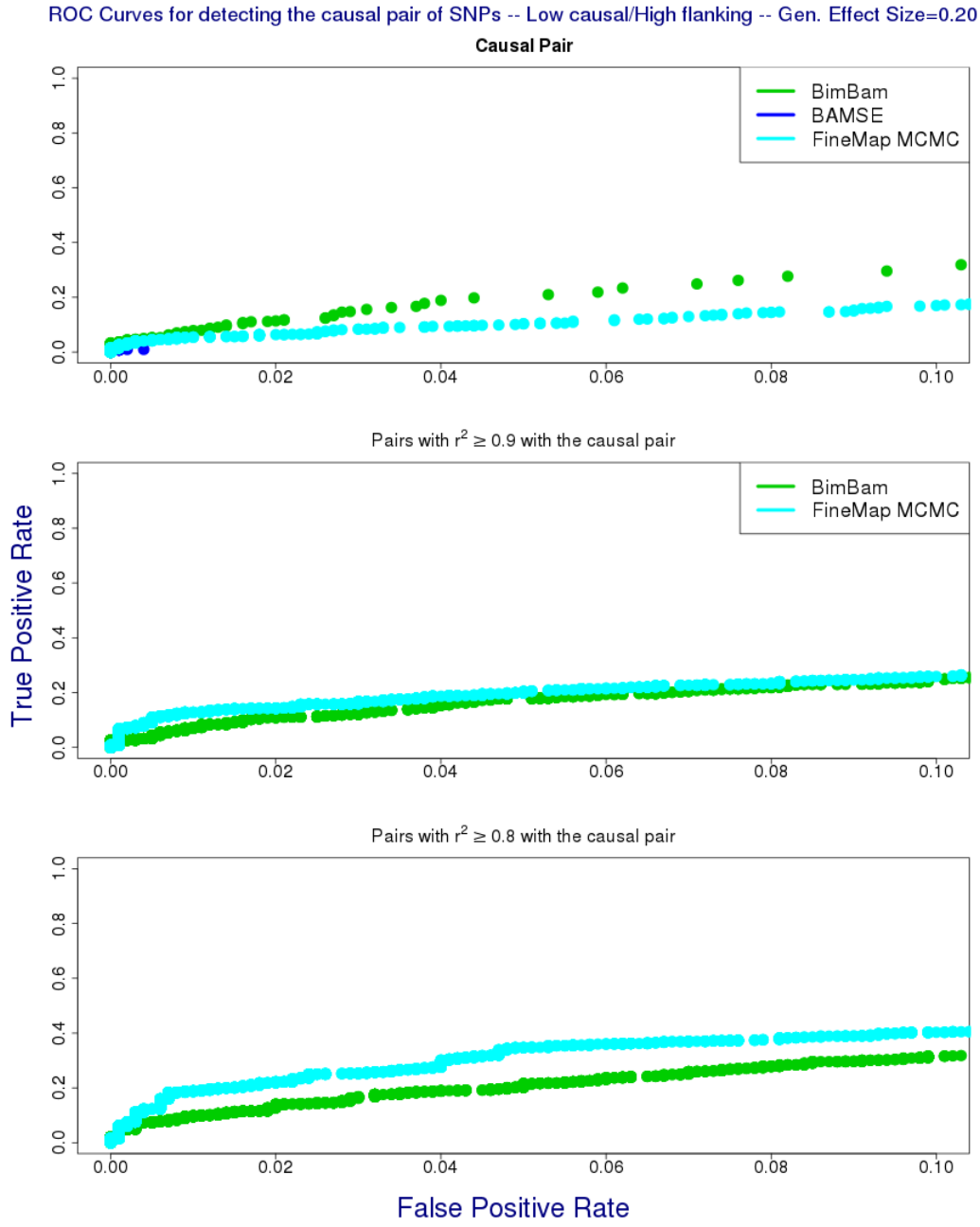

Figure E.10: ROC curves to detect the causal pair of variants (**BAMSE**, **BIMBAM** and **FineMap MCMC**) and to detect the pairs in high LD ($r^2 = 0.80$ and 0.90) with the causal combination of SNPs (**BIMBAM** and **FineMap MCMC**). The horizontal axes represent the false positive rates while the vertical axes represent the true positive rates. The different methods are coded using the following colours: **BIMBAM** (green), **BAMSE** (blue), **FineMap MCMC** (cyan) (LD: low causal/high flanking).

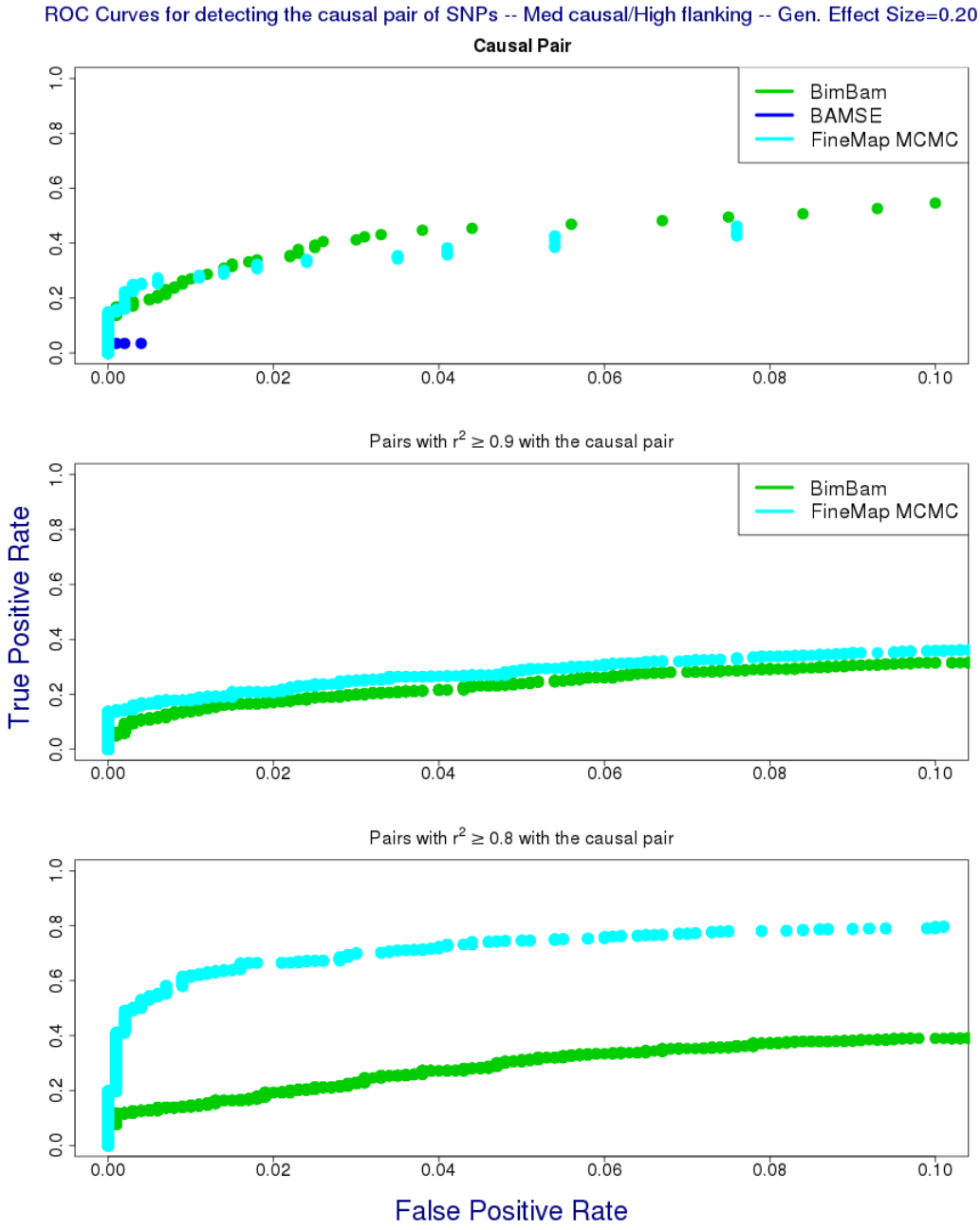

Figure E.11: ROC curves to detect the causal pair of variants (**BAMSE**, **BIMBAM** and **FineMap MCMC**) and to detect the pairs in high LD ($r^2 = 0.80$ and 0.90) with the causal combination of SNPs (**BIMBAM** and **FineMap MCMC**). The horizontal axes represent the false positive rates while the vertical axes represent the true positive rates. The different methods are coded using the following colours: **BIMBAM** (green), **BAMSE** (blue), **FineMap MCMC** (cyan) (LD: med causal/high flanking).

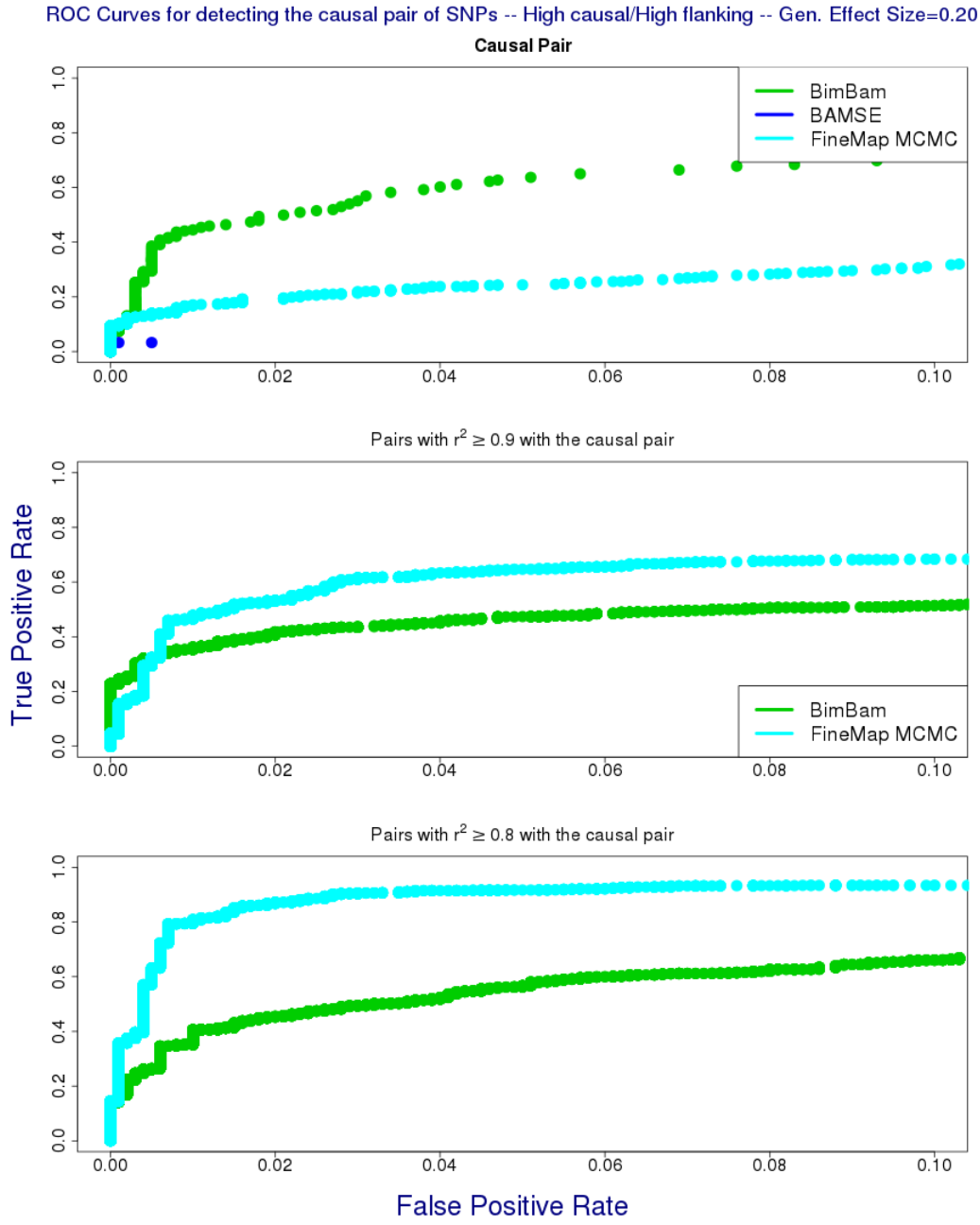

Figure E.12: ROC curves to detect the causal pair of variants (**BAMSE**, **BIMBAM** and **FineMap MCMC**) and to detect the pairs in high LD ($r^2 = 0.80$ and 0.90) with the causal combination of SNPs (**BIMBAM** and **FineMap MCMC**). The horizontal axes represent the false positive rates while the vertical axes represent the true positive rates. The different methods are coded using the following colours: **BIMBAM** (green), **BAMSE** (blue), **FineMap MCMC** (cyan) (LD: high causal/high flanking).

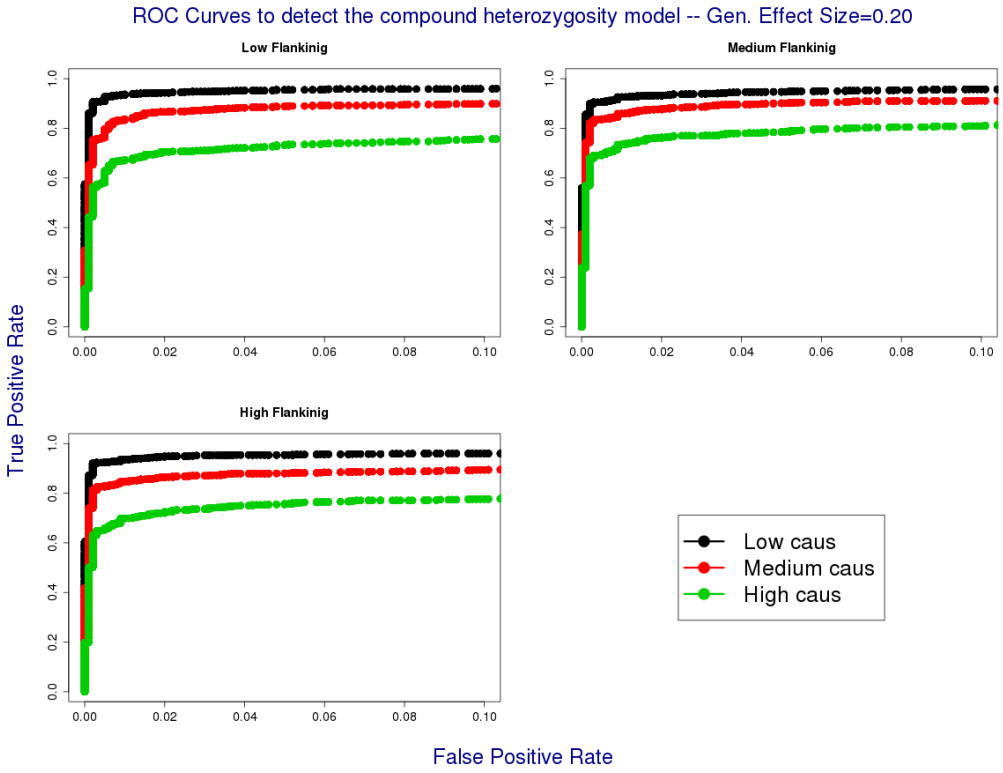

Figure E.13: ROC curves to detect the compound heterozygosity model for the data under the examined structures of LD. The horizontal axes represent the false positive rates while the vertical axes represent the true positive rates. The three pairs of axes describe the LD structures between the causal SNPs and the flanking sites. The black ROC curves show the results for low LD between the causal SNPs, the red ROC curve is for medium LD between the causal SNPs and the green ROC curve is for high LD between the causal SNPs.

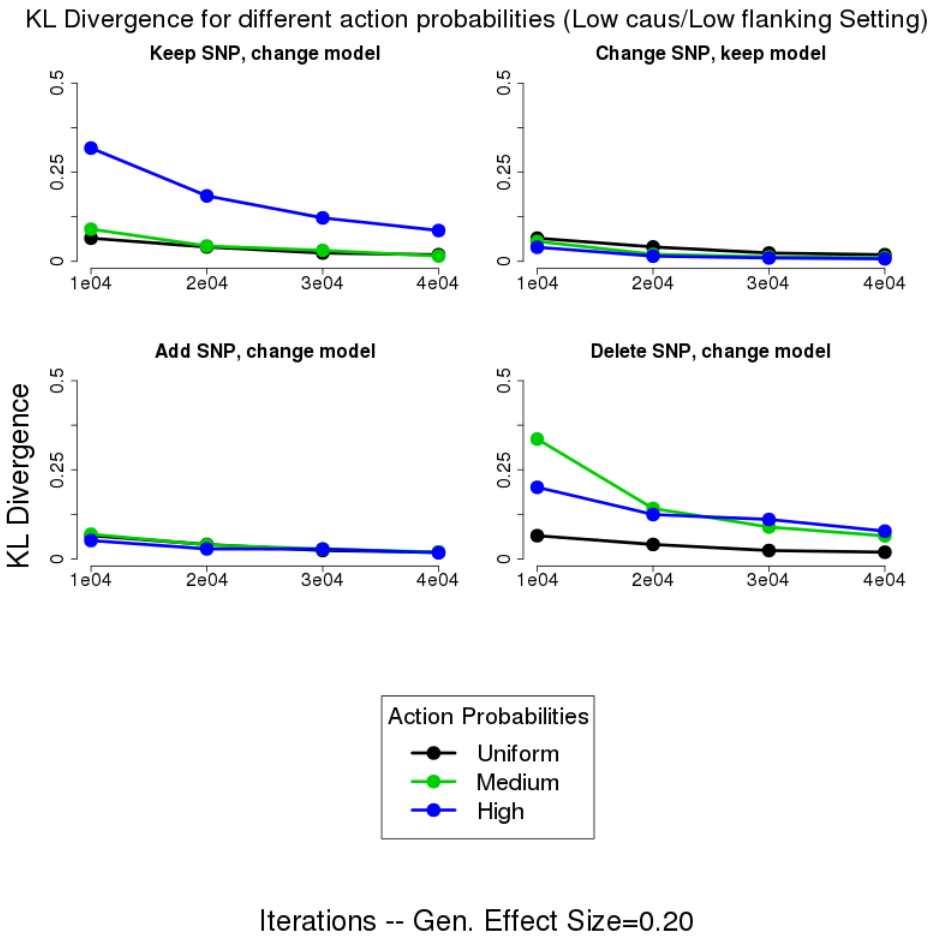

Figure E.14: Kullback-Leibler divergence between the true and estimated marginal posterior distribution of each SNP for a single dataset for different number of iterations. The horizontal axis describes the number of iterations while the vertical axis shows the KL divergence. The black line represents the uniform action probabilities scenario, the green line represents the medium action probabilities scenario and the blue line denotes the high action probabilities scenario (LD: low causal/low flanking).

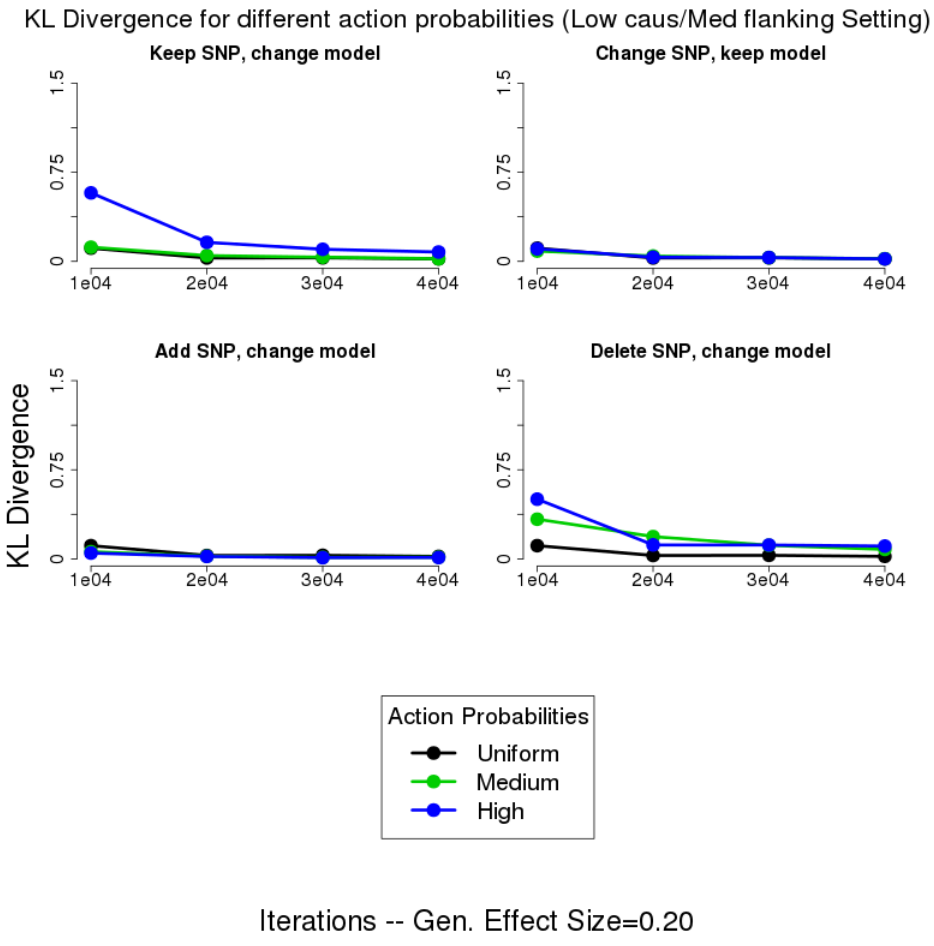

Figure E.15: Kullback-Leibler divergence between the true and estimated marginal posterior distribution of each SNP for a single dataset for different number of iterations. The horizontal axis describes the number of iterations while the vertical axis shows the KL divergence. The black line represents the uniform action probabilities scenario, the green line represents the medium action probabilities scenario and the blue line denotes the high action probabilities scenario (LD: low causal/med flanking).

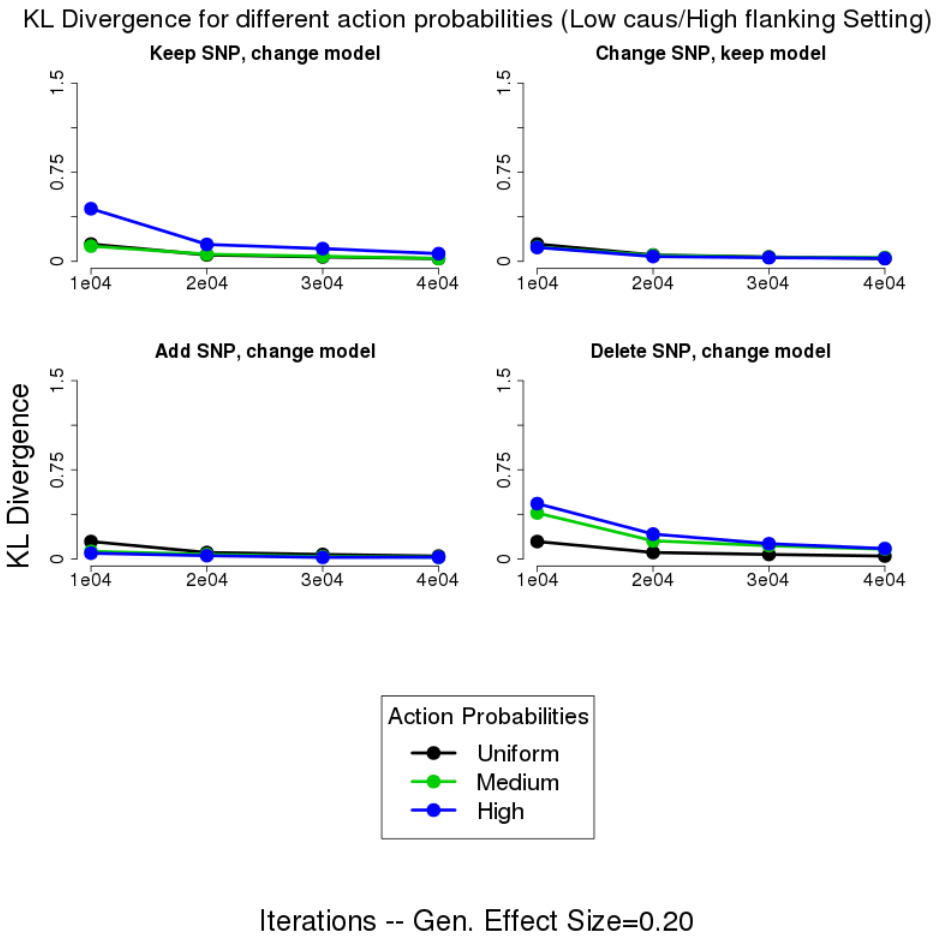

Figure E.16: Kullback-Leibler divergence between the true and estimated marginal posterior distribution of each SNP for a single dataset for different number of iterations. The horizontal axis describes the number of iterations while the vertical axis shows the KL divergence. The black line represents the uniform action probabilities scenario, the green line represents the medium action probabilities scenario and the blue line denotes the high action probabilities scenario (LD: low causal/high flanking).

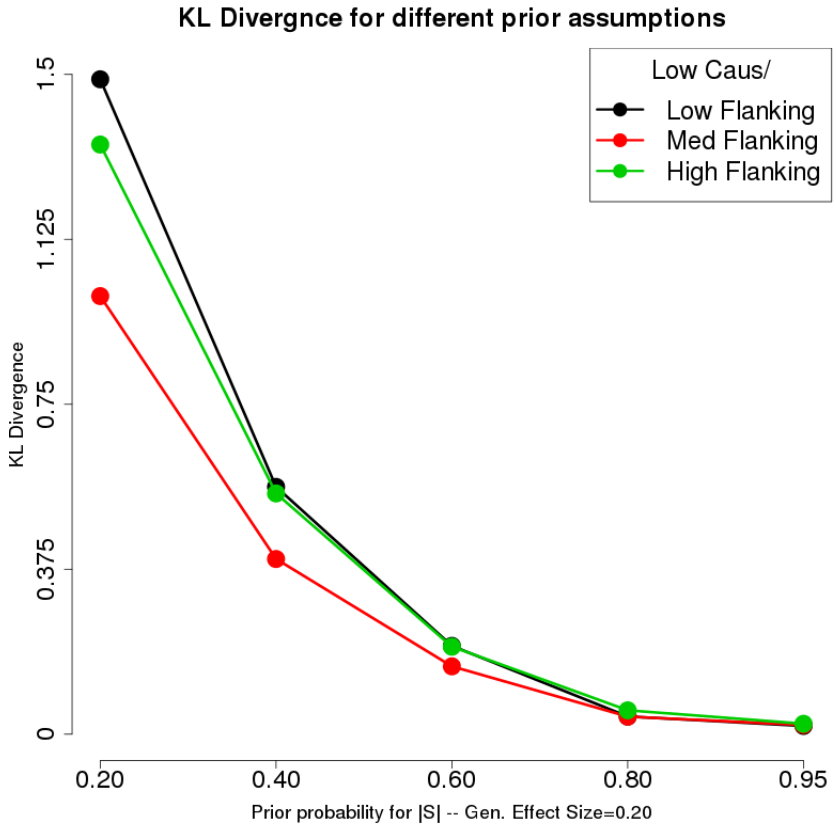

Figure E.17: Kullback-Leibler divergence between the true and estimated marginal posterior distribution of each SNP for a single dataset for different prior probabilities for $|S|$. The horizontal axis presents the different values of p and the vertical axis shows the KL divergence. The LD between the causal SNPs is low while the flanking LD is low (black line), medium (red line) and high (green line) (LD: low causal/low,med,high flanking).

Appendix F

Probability to Detect the True Model using the Pritchard Approach

Table F.1 presents the probability of each model of our meta-analysis approach for the different values of the parameters examined when the marginal likelihood is calculated under Pritchard's method [67]. A general remark is that when all cohorts are associated, for any level of heterogeneity, the true model exhibits the largest probability and this agrees with the analogous observations for the probability estimates when Chib's method is used. On the other hand, we observe dissimilarities in the probability estimates of the general random effects model and in some cases for the fixed effects effects model which allows for cohorts not to be associated when these are the true underlying models respectively. More specifically, for all values of β, as heterogeneity increases we observe increased probability for the general random effects model as the number of associated cohorts decreases. In most cases, though, these estimates are not the largest and the random effects, all cohorts associated, model outperforms the rest. Quite surprisingly, when the data is simulated under the fixed effects model with decreasing number of associated cohorts, the random effects model assuming all cohorts associated outperforms the true model.

		$\beta = 0.143$				$\beta = 0.101$				$\beta = 0.071$				$\beta = 0.045$			
		All		Subset		All		Subset		All		Subset		All		Subset	
	Ass.	Random	Fixed	Random	Fixed	Random	Fixed	Random	Fixed	Random	Fixed	Random	Fixed	Random	Fixed	Random	Fixed
$\lambda = 1000$	10	0.583	0.220	0.002	0.195	0.526	0.253	0.000	0.221	0.560	0.280	0.001	0.159	0.656	0.244	0.007	0.093
	8	0.304	0.007	0.001	0.688	0.495	0.064	0.002	0.439	0.695	0.124	0.006	0.175	0.658	0.224	0.027	0.091
	6	0.208	0.003	0.006	0.783	0.529	0.041	0.008	0.422	0.699	0.118	0.010	0.173	0.664	0.190	0.039	0.107
	4	0.256	0.004	0.024	0.716	0.627	0.043	0.021	0.309	0.728	0.108	0.024	0.140	0.653	0.147	0.063	0.137
	2	0.473	0.015	0.050	0.462	0.677	0.070	0.043	0.210	0.620	0.137	0.095	0.148	0.527	0.167	0.125	0.181
$\lambda = 100$	10	0.712	0.002	0.036	0.250	0.642	0.000	0.021	0.337	0.681	0.002	0.027	0.290	0.761	0.003	0.017	0.219
	8	0.565	0.002	0.050	0.383	0.561	0.001	0.050	0.388	0.581	0.004	0.036	0.379	0.735	0	0.026	0.239
	6	0.406	0.003	0.072	0.519	0.447	0.006	0.056	0.491	0.555	0.007	0.049	0.389	0.636	0.013	0.063	0.288
	4	0.291	0.004	0.139	0.566	0.404	0.017	0.108	0.471	0.572	0.015	0.074	0.339	0.595	0.031	0.077	0.297
	2	0.341	0.033	0.175	0.451	0.477	0.050	0.147	0.326	0.481	0.086	0.139	0.294	0.554	0.084	0.126	0.236
$\lambda = 15$	10	0.984	0	0.013	0.003	0.986	0	0.009	0.005	0.984	0	0.012	0.004	0.984	0	0.011	0.005
	8	0.944	0	0.044	0.012	0.943	0	0.039	0.018	0.968	0	0.021	0.011	0.967	0	0.023	0.010
	6	0.841	0	0.116	0.043	0.848	0	0.110	0.042	0.861	0	0.084	0.055	0.871	0	0.079	0.050
	4	0.256	0.004	0.024	0.716	0.642	0	0.227	0.131	0.649	0.003	0.210	0.138	0.667	0.003	0.202	0.128
	2	0.362	0.011	0.360	0.267	0.378	0.014	0.338	0.270	0.391	0.011	0.309	0.289	0.395	0.030	0.313	0.262
$\lambda = 5$	10	0.965	0	0.035	0	0.917	0	0.028	0.055	0.98	0	0.02	0	0.969	0	0.031	0
	8	0.927	0	0.073	0	0.921	0	0.079	0.000	0.944	0	0.056	0	0.927	0	0.073	0
	6	0.855	0	0.142	0.003	0.848	0	0.149	0.003	0.870	0	0.129	0.001	0.862	0	0.136	0.002
	4	0.703	0	0.265	0.032	0.690	0.001	0.284	0.025	0.686	0	0.281	0.033	0.699	0	0.272	0.029
	2	0.368	0.005	0.458	0.169	0.359	0.005	0.481	0.155	0.356	0.003	0.458	0.183	0.359	0.009	0.462	0.170
$\lambda^{-1} = 0$	10	0.355	0.434	0	0.211	0.332	0.522	0	0.146	0.384	0.518	0	0.098	0.446	0.510	0.003	0.041
	8	0.222	0.015	0.001	0.762	0.484	0.107	0	0.409	0.606	0.264	0	0.130	0.531	0.397	0.014	0.058
	6	0.182	0.002	0	0.816	0.583	0.039	0.001	0.377	0.724	0.173	0.003	0.100	0.611	0.299	0.036	0.054
	4	0.242	0.001	0.008	0.749	0.692	0.033	0.011	0.264	0.766	0.129	0.026	0.079	0.571	0.229	0.071	0.129
	2	0.506	0.009	0.040	0.445	0.742	0.081	0.028	0.149	0.636	0.152	0.083	0.129	0.472	0.194	0.142	0.192

Table F.1: Probability to detect the correct model under which the data are simulated from for different values of β, λ and different number of associated cohorts using Pritchard's method to obtain the marginal likelihood (The correct model is highlighted in red while the model which exhibits the highest probability is highlighted in blue).

Further, we examine the stability of this marginal likelihood approach by performing two independent runs of the data simulated under $\beta = 0.143$ and $\lambda = 1000, 15$ respectively (Table F.2). In both circumstances, we observe very similar results between the two runs allowing us to conclude that this approach does not face any stability issues.

		Run 1				Run 2			
		All		Subset		All		Subset	
	Ass.	Random	Fixed	Random	Fixed	Random	Fixed	Random	Fixed
$\lambda = 1000$	10	0.583	0.220	0.002	0.195	0.578	0.211	0.001	0.210
	8	0.304	0.007	0.001	0.688	0.303	0.010	0.003	0.684
	6	0.208	0.003	0.006	0.783	0.233	0.002	0.004	0.761
	4	0.256	0.004	0.024	0.716	0.254	0.002	0.027	0.717
	2	0.473	0.015	0.050	0.462	0.466	0.022	0.043	0.469
$\lambda = 15$	10	0.984	0	0.013	0.003	0.986	0	0.013	0.001
	8	0.944	0	0.044	0.012	0.946	0	0.043	0.011
	6	0.841	0	0.116	0.043	0.840	0	0.116	0.044
	4	0.666	0	0.207	0.127	0.662	0	0.215	0.123
	2	0.362	0.011	0.360	0.267	0.367	0.013	0.335	0.285

Table F.2: Comparison of two runs calculating the probability for the case where $\beta = 0.143$ and $\lambda = 1000, 15$ and different number of associated cohorts using Pritchard's method to obtain the marginal likelihood (The correct model is highlighted in red while the model which exhibits the highest probability is highlighted in blue).

Bibliography

[1] The HapMap Project. `hapmap.ncbi.nlm.nih.gov/`, 2005. [Online; accessed: 05-April-2013].

[2] GWA Studies. `http://www.genome.gov/gwastudies/`, 2013. [Online; accessed: 29-March-2013].

[3] ALBRECHTSEN ANDERS, CASTELLA SOFIE, ANDERSEN GITTE, HANSEN TORBEN, PEDERSEN OLUF, AND NIELSEN,RASMUS. A Bayesian Multilocus Association Method: Allowing for Higher-Order Interaction in Association Studies. *Genetics 176*, 2 (2007), 1197–1208.

[4] BAGOS PANTELIS. Meta-Analysis of Haplotype-Association Studies: Comparison of Methods and Empirical Evaluation of the Literature. *BMC Genetics 12*, 1 (2011), 8.

[5] BAGOS PANTELIS G., AND LIAKOPOULOS THEODORE D. A Multipoint Method for Meta-Analysis of Genetic Association Studies. *Genetic Epidemiology 34*, 7 (2010), 702–715.

[6] BALDING DAVID J. A Tutorial on Statistical Methods for Population Association Studies. *Nature Review Genetics 7*, 10 (October 2006), 781–791.

[7] BANSAL VIKAS, LIBIGER ONDREJ, TORKAMANI ALI, AND SCHORK NICHOLAS J. Statistical Analysis Strategies for Association Studies Involving Rare Variants. *Nat Reviews Genetics 11*, 11 (November 2010), 773–785.

[8] BEAUMONT MARK A., AND RANNALA BRUCE. The Bayesian Revolution in Genetics. *Nature Reviews Genetics 5*, 4 (April 2004), 251–261.

[9] BELLENGUEZ CELINE, BEVAN STEVE, GSCHWENDTNER ANDREAS, SPENCER CHRIS C. A., BURGESS ANNETTE I., PIRINEN MATTI, JACKSON CAROLINE A., TRAYLOR MATTHEW, STRANGE AMY, SU ZHAN, BAND GAVIN, SYME PAUL D., MALIK RAINER, PERA JOANNA, NORRVING BO, LEMMENS ROBIN, FREEMAN COLIN, SCHANZ RENATA, JAMES TOM, POOLE DEBORAH, MURPHY LEE, SEGAL HELEN, CORTELLINI LYNELLE, CHENG YU-CHING, WOO DANIEL, NALLS MICHAEL A., MULLER-MYHSOK BERTRAM, MEISINGER CHRISTA, SEEDORF UDO, ROSS-ADAMS HELEN, BOONEN STEVEN, WLOCH-KOPEC DOROTA, VALANT VALERIE, SLARK JULIA, FURIE KAREN, DELAVARAN HOSSEIN, LANGFORD CORDELIA, DELOUKAS PANOS, EDKINS SARAH, HUNT SARAH, GRAY EMMA, DRONOV SERGE, PELTONEN LEENA, GRETARSDOTTIR SOLVEIG, THORLEIFSSON GUDMAR, THORSTEINSDOTTIR UNNUR, STEFANSSON KARI, BONCORAGLIO GIORGIO B., PARATI EUGENIO A., ATTIA JOHN, HOLLIDAY ELIZABETH, LEVI CHRIS, FRANZOSI MARIA-GRAZIA, GOEL ANUJ, HELGADOTTIR ANNA, BLACKWELL JENEFER M., BRAMON ELVIRA, BROWN MATTHEW A., CASAS JUAN P., CORVIN AIDEN, DUNCANSON AUDREY, JANKOWSKI JANUSZ, MATHEW CHRISTOPHER G., PALMER COLIN N. A., PLOMIN ROBERT, RAUTANEN ANNA, SAWCER STEPHEN J., TREMBATH RICHARD C., VISWANATHAN ANANTH C., WOOD NICHOLAS W., WORRALL BRADFORD B., KITTNER STEVEN J., MITCHELL BRAXTON D., KISSELA BRETT, MESCHIA JAMES F., THIJS VINCENT, LINDGREN ARNE, MACLEOD MARY JOAN, SLOWIK AGNIESZKA, WALTERS MATTHEW, ROSAND JONATHAN, SHARMA PANKAJ, FARRALL MARTIN, SUDLOW CATHIE L. M., ROTHWELL PETER M., DICHGANS MARTIN, DONNELLY PETER, MARKUS HUGH S., ET AL. Genome-Wide Association Study Identifies a Variant in HDAC9 Associated with

Large Vessel Ischemic Stroke. *Nature Genetics 44*, 3 (March 2012), 328–333.

[10] BISHOP, CHRISTOPHER M. *Pattern Recognition and Machine Learning*. Springer, 2006.

[11] BROOKS STEPHEN P., AND GELMAN ANDREW. General Methods for Monitoring Convergence of Iterative Simulations. *Journal of Computational and Graphical Statistics 7*, 4 (1998), 434–455.

[12] CHIB SIDDHARTHA. Marginal Likelihood from the Gibbs Output. *Journal of the American Statistical Association 90*, 432 (December 1995), 1313–1321.

[13] CHIB SIDDHARTHA, AND GREENBERG EDWARD. Understanding the Metropolis-Hastings Algorithm. *The American Statistician 49*, 4 (1995), 327–335.

[14] CLARK ANDREW G. The role of haplotypes in candidate gene studies. *Genetic Epidemiology 27*, 4 (2004), 321–333.

[15] COOKSON WILLIAM, LIANG LIMING, ABECASIS GONCALO R., MOFFATT MIRIAM, AND LATHROP MARK. Mapping Complex Disease Traits With Global Gene Expression. *Nature Reviews Genetics 10*, 3 (March 2009), 184–194.

[16] CORDELL HEATHER J. Detecting Gene-Gene Interactions That Underlie Human Diseases. *Nature Review Genetics 10*, 6 (June 2009), 392–404.

[17] DE BAKKER PAUL I.W., FERREIRA MANUEL A.R., JIA XIAOMING, NEALE BENJAMIN M., RAYCHAUDHURI SOUMYA, AND VOIGHT BENJAMIN F. Practical Aspects of Imputation-Driven Meta-Analysis of Genome-Wide Association Studies. *Human Molecular Genetics 17*, R2 (2008), R122–R128.

[18] DELANEAU, OLIVIER, MARCHINI, JONATHAN, AND ZAGURY, JEAN-FRANCOIS. A Linear Complexity Phasing Method for Thousands of Genomes. *Nature Methods 9*, 2 (February 2012), 179–181.

[19] DELOUKAS PANOS, KANONI STAVROULA, WILLENBORG CHRISTINA, FARRALL MARTIN, ASSIMES THEMISTOCLES L., THOMPSON JOHN R., INGELSSON ERIK, SALEHEEN DANISH, ERDMANN JEANETTE, GOLDSTEIN BENJAMIN A., STIRRUPS KATHLEEN, KONIG INKE R., CAZIER JEAN-BAPTISTE, JOHANSSON ASA, HALL ALISTAIR S., LEE JONG-YOUNG, WILLER CRISTEN J, CHAMBERS JOHN C., ESKO TONU, FOLKERSEN LASSE, GOEL ANUJ, GRUNDBERG ELIN, HAVULINNA AKI S., HO WEANG K., HOPEWELL JEMMA C., ERIKSSON NICLAS, KLEBER MARCUS E., KRISTIANSSON KATI, LUNDMARK PER, LYYTIKAINEN LEO-PEKKA, RAFELT SUZANNE, SHUNGIN DMITRY, STRAWBRIDGE RONA J., THORLEIFSSON GUDMAR, TIKKANEN EMMI, VAN ZUYDAM NATALIE, VOIGHT BENJAMIN F., WAITE LINDSAY L., ZHANG WEIHUA, ZIEGLER ANDREAS, ABSHER DEVIN, ALTSHULER DAVID, BALMFORTH ANTHONY J., BARROSO INES, BRAUND PETER S., BURGDORF CHRISTOF, CLAUDI-BOEHM SIMONE, COX DAVID, DIMITRIOU MARIA, DO RON, DONEY ALEX S. F., MOKHTARI NOUREDDINE EL, ERIKSSON PER, FISCHER KRISTA, FONTANILLAS PIERRE, FRANCO-CERECEDA ANDERS, GIGANTE BRUNA, GROOP LEIF, GUSTAFSSON STEFAN, HAGER JORG, HALLMANS GORAN, HAN BOK-GHEE, HUNT SARAH E., KANG HYUN M., ILLIG THOMAS, KESSLER THORSTEN, KNOWLES JOSHUA W., KOLOVOU GENOVEFA, KUUSISTO JOHANNA, LANGENBERG CLAUDIA, LANGFORD CORDELIA, LEANDER KARIN, LOKKI MARJA-LIISA, LUNDMARK ANDERS, MCCARTHY MARK I., MEISINGER CHRISTA, MELANDER OLLE, MIHAILOV EVELIN, MAOUCHE SERAYA, MORRIS ANDREW D., MULLER-NURASYID MARTINA, NIKUS KJELL, PEDEN JOHN F., RAYNER N WILLIAM, RASHEED ASIF, ROSINGER SILKE, RUBIN DIANA, RUMPF MORITZ P., SCHAFER ARNE, SIVANANTHAN MOHAN, SONG CI, STEWART ALEXANDRE F. R., TAN SIAN-TSUNG, THORGEIRSSON GUDMUNDUR, SCHOOT C ELLEN VAN DER, WAGNER PETER J., WELLS GEORGE A, WILD PHILIPP S., YANG TSUN-PO, AMOUYEL PHILIPPE, ARVEILER DOMINIQUE, BASART HAN-

NEKE, BOEHNKE MICHAEL, BOERWINKLE ERIC, BRAMBILLA PAOLO, CAMBIEN FRANCOIS, CUPPLES ADRIENNE L., ULF, D. F., DEHGHAN ABBAS, DIEMERT PATRICK, EPSTEIN STEPHEN E., EVANS ALUN, FERRARIO MARCO M., FERRIERES JEAN, GAUGUIER DOMINIQUE, GO ALAN S, GOODALL ALISON H., GUDNASON VILLI, HAZEN STANLEY L., HOLM HILMA, IRIBARREN CARLOS, JANG YANGSOO, KAHONEN MIKA, KEE FRANK, KIM HYO-SOO, KLOPP NORMAN, KOENIG WOLFGANG, KRATZER WOLFGANG, KUULASMAA KARI, LAAKSO MARKKU, LAAKSONEN REIJO, LEE JI-YOUNG, LIND LARS, OUWEHAND WILLEM H., PARISH SARAH, PARK JEONG E., PEDERSEN NANCY L., PETERS ANNETTE, QUERTERMOUS THOMAS, RADER DANIEL J., SALOMAA VEIKKO, SCHADT ERIC, SHAH SVATI H., SINISALO JUHA, STARK KLAUS, STEFANSSON KARI, TREGOUET DAVID-ALEXANDRE, VIRTAMO JARMO, WALLENTIN LARS, WAREHAM NICHOLAS, ZIMMERMANN MARTINA E., NIEMINEN MARKKU S., HENGSTENBERG CHRISTIAN, SANDHU MANJINDER S., PASTINEN TOMI, SYVANEN ANN-CHRISTINE, HOVINGH G KEES, DEDOUSSIS GEORGE, FRANKS PAUL W., LEHTIMAKI TERHO, METSPALU ANDRES, ZALLOUA PIERRE A., SIEGBAHN AGNETA, SCHREIBER STEFAN, RIPATTI SAMULI, BLANKENBERG STEFAN S., PEROLA MARKUS, CLARKE ROBERT, BOEHM BERNHARD O., O'DONNELL CHRISTOPHER, REILLY MUREDACH P., MARZ WINFRIED, COLLINS RORY, KATHIRESAN SEKAR, HAMSTEN ANDERS, KOONER JASPAL S., THORSTEINSDOTTIR UNNUR, DANESH JOHN, PALMER COLIN N. A., ROBERTS ROBERT, WATKINS HUGH, SCHUNKERT HERIBERT, SAMANI NILESH J., ET AL. Large-Scale Association Analysis Identifies New Risk Loci for Coronary Artery Disease. *Nature Genetics 45*, 1 (January 2013), 25–33.

[20] DENISON D.G.T., HOLMES C.C., MALLICK B.K., AND SMITH A.F.M. *Bayesian Methods for Nonlinear Classification and Regression*. Wiley Series in Probability and Statistics, 2002.

[21] EASTON DOUGLAS F., POOLEY KAREN A., DUNNING ALISON M., PHAROAH PAUL D. P., THOMPSON DEBORAH, BALLINGER DENNIS G., STRUEWING JEFFERY P., MORRISON JONATHAN, FIELD HELEN, LUBEN ROBERT, WAREHAM NICHOLAS, AHMED SHAHANA, HEALEY CATHERINE S., BOWMAN RICHARD, MEYER KERSTIN B., HAIMAN CHRISTOPHER A., KOLONEL LAURENCE K., HENDERSON BRIAN E., LE MARCHAND LOIC, BRENNAN PAUL, SANGRAJRANG SULEEPORN, GABORIEAU VALERIE, ODEFREY FABRICE, SHEN CHEN-YANG, WU PEI-EI, WANG HUI-CHUN, ECCLES DIANA, EVANS D. GARETH, PETO JULIAN, FLETCHER OLIVIA, JOHNSON NICHOLA, SEAL SHEILA, STRATTON MICHAEL R., RAHMAN NAZNEEN, CHENEVIX-TRENCH GEORGIA, BOJESEN STIG E., NORDESTGAARD BORGE G., AXELSSON CHRISTEN K., GARCIA-CLOSAS MONTSERRAT, BRINTON LOUISE, CHANOCK STEPHEN, LISSOWSKA JOLANTA, PEPLONSKA BEATA, NEVANLINNA HELI, FAGERHOLM RAINER, EEROLA HANNALEENA, KANG DAEHEE, YOO KEUN-YOUNG, NOH DONG-YOUNG, AHN SEI-HYUN, HUNTER DAVID J., HANKINSON SUSAN E., COX DAVID G., HALL PER, WEDREN SARA, LIU JIANJUN, LOW YEN-LING, BOGDANOVA NATALIA, SCHURMANN PETER, DORK THILO, TOLLENAAR ROB A. E. M., JACOBI CATHARINA E., DEVILEE PETER, KLIJN JAN G. M., SIGURDSON ALICE J., DOODY MICHELE M., ALEXANDER BRUCE H., ZHANG JINGHUI, COX ANGELA, BROCK IAN W., MACPHERSON GORDON, REED MALCOLM W. R., COUCH FERGUS J., GOODE ELLEN L., OLSON JANET E., MEIJERS-HEIJBOER HANNE, ANS, V. D. O., UITTERLINDEN ANDRE, RIVADENEIRA FERNANDO, MILNE ROGER L., RIBAS GLORIA, GONZALEZ-NEIRA ANNA, BENITEZ JAVIER, HOPPER JOHN L., MCCREDIE MARGARET, SOUTHEY MELISSA, GILES GRAHAM G., SCHROEN CHRIS, JUSTENHOVEN CHRISTINA, BRAUCH HILTRUD, HAMANN UTE, KO YON-DSCHUN, SPURDLE AMANDA B., BEESLEY JONATHAN, CHEN XIAOQING, MANNERMAA ARTO, KOSMA VELI-MATTI, KATAJA VESA, HARTIKAINEN JAANA, DAY NICHOLAS E., COX DAVID R., PON-

DER BRUCE A. J., ET AL. Genome-Wide Association Study Identifies Novel Breast Cancer Susceptibility Loci. *Nature 447*, 7148 (June 2007), 1087–1093.

[22] EELES ROSALIND A., KOTE-JARAI ZSOFIA, GILES GRAHAM G., OLAMA ALI AMIN AL, GUY MICHELLE, JUGURNAUTH SARAH K, MULHOLLAND SHANI, LEONGAMORNLERT DANIEL A., EDWARDS STEPHEN M., MORRISON JONATHAN, FIELD HELEN I., SOUTHEY MELISSA C., SEVERI GIANLUCA, DONOVAN JENNY L., HAMDY FREDDIE C., DEARNALEY DAVID P., MUIR KENNETH R., SMITH CHARMAINE, BAGNATO MELISA, ARDERN-JONES AUDREY T., HALL AMANDA L., O'BRIEN LYNNE T, GEHR-SWAIN BEATRICE N., WILKINSON ROSEMARY A., COX ANGIE, LEWIS SARAH, BROWN PAUL M., JHAVAR SAMEER G., TYMRAKIEWICZ MALGORZATA, LOPHATANANON ARTITAYA, BRYANT SARAH L., HORWICH ALAN, HUDDART ROBERT A., KHOO VINCENT S., PARKER CHRISTOPHER C., WOODHOUSE CHRISTOPHER J., THOMPSON ALAN, CHRISTMAS TIM, OGDEN CHRIS, FISHER CYRIL, JAMIESON CHARLES, COOPER COLIN S., ENGLISH DALLAS R., HOPPER JOHN L., NEAL DAVID E., EASTON DOUGLAS F., ET AL. Multiple Newly Identified Loci Associated With Prostate Cancer Susceptibility. *Nature Genetics 40*, 3 (March 2008), 316–321.

[23] FALCONER D.S., AND MACKAY TRUDY F.C. *Introduction to Quantitative Genetics*, fourth ed. Longman Group Ltd, 1996.

[24] FERREIRA TERESA, AND MARCHINI JONATHAN. Modeling Interactions With Known Risk Loci-a Bayesian Model Averaging Approach. *Annals of Human Genetics 75*, 1 (2011), 1–9.

[25] FRANKE ANDRE, MCGOVERN DERMOT P. B., BARRETT JEFFREY C., WANG KAI, RADFORD-SMITH GRAHAM L., AHMAD TARIQ, LEES CHARLIE W, BALSCHUN TOBIAS, LEE JAMES, ROBERTS REBECCA, ANDERSON CARL A., BIS JOSHUA C., BUMPSTEAD SUZANNE, ELLINGHAUS DAVID, FESTEN

ELEONORA M., GEORGES MICHEL, GREEN TODD, HARITUNIANS TALIN, JOSTINS LUKE, LATIANO ANNA, MATHEW CHRISTOPHER G., MONTGOMERY GRANT W., PRESCOTT NATALIE J., RAYCHAUDHURI SOUMYA, ROTTER JEROME I., SCHUMM PHILIP, SHARMA YASHODA, SIMMS LISA A., TAYLOR KENT D., WHITEMAN DAVID, WIJMENGA CISCA, BALDASSANO ROBERT N., BARCLAY MURRAY, BAYLESS THEODORE M., BRAND STEPHAN, BUNING CARSTEN, COHEN ALBERT, COLOMBEL JEAN-FREDERICK, COTTONE MARIO, STRONATI LAURA, DENSON TED, DE VOS MARTINE, D'INCA RENATA, DUBINSKY MARLA, EDWARDS CATHRYN, FLORIN TIM, FRANCHIMONT DENIS, GEARRY RICHARD, GLAS JURGEN, VAN GOSSUM ANDRE, GUTHERY STEPHEN L., HALFVARSON JONAS, VERSPAGET HEIN W., HUGOT JEAN-PIERRE, KARBAN AMIR, LAUKENS DEBBY, LAWRANCE IAN, LEMANN MARC, LEVINE ARIE, LIBIOULLE CECILE, LOUIS EDOUARD, MOWAT CRAIG, NEWMAN WILLIAM, PANES JULIAN, PHILLIPS ANNE, PROCTOR DEBORAH D., REGUEIRO MIGUEL, RUSSELL RICHARD, RUTGEERTS PAUL, SANDERSON JEREMY, SANS MIQUEL, SEIBOLD FRANK, STEINHART A. HILLARY, STOKKERS PIETER C. F., TORKVIST LEIF, KULLAK-UBLICK GERD, WILSON DAVID, WALTERS THOMAS, TARGAN STEPHAN R., BRANT STEVEN R., RIOUX JOHN D., D'AMATO MAURO, WEERSMA RINSE K., KUGATHASAN SUBRA, GRIFFITHS ANNE M., MANSFIELD JOHN C., VERMEIRE SEVERINE, DUERR RICHARD H., SILVERBERG MARK S., SATSANGI JACK, SCHREIBER STEFAN, CHO JUDY H., ANNESE VITO, HAKONARSON HAKON, DALY MARK J., PARKES MILES, ET AL. Genome-Wide Meta-Analysis Increases to 71 the Number of Confirmed Crohn's Disease Susceptibility Loci. *Nature Genetics 42*, 12 (December 2010), 1118–1125.

[26] FRIDLEY BROOKE L. Bayesian Variable and Model Selection Methods for Genetic Association Studies. *Genetic Epidemiology 33*, 1 (2009), 27–37.

[27] GABRIELSEN, MAIKEN E., ROMUNDSTAD, PAL, LANGHAMMER ARNULF,

KROKAN, HANS E., AND SKORPEN, FRANK. Association Between a 15q25 Gene Variant, Nicotine-Related Habits, Lung Cancer and COPD Among 56 307 Individuals From the HUNT Study in Norway. *Eur Jurnal of Human Genetics* (February 2013).

[28] GAMARRA SOLEDAD, GARCIA-EFFRON GUILLERMO, MONTESERIN CARMEN, ET AL. Beta-Thalassaemia Major in a Spanish Patient due to a Compound Heterozygosity for CD39 C →T/−28 A →C. *Advances in Hematology* (2009).

[29] GELMAN ANDREW, CARLIN JOHN B., STERN HAL S., AND RUBIN DONALD B. *Bayesian Data Analysis*. Chapman & Hall, 2004.

[30] GELMAN ANDREW, AND RUBIN DONALD B. Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science 7*, 4 (1992), 457–472.

[31] GIBSON GREG. Rare and Common Variants: Twenty Arguments. *Nature Review Genetics 13*, 2 (February 2012), 135–145.

[32] GUAN YONGTAO, AND STEPHENS MATTHEW. Bayesian Variable Selection Regression for Genome-Wide Association Studies and Other Large-Scale Problems. *Annals of Applied Statistics 5*, 3 (2011), 1780–1815.

[33] HAIMAN CHRISTOPHER A., STRAM DANIEL O., PIKE MALCOLM C., KOLONEL LAURENCE N., BURTT NOEL P., ALTSHULER DAVID, HIRSCHHORN JOEL, AND HENDERSON BRIAN E. A Comprehensive Haplotype Analysis of CYP19 and Breast Cancer Risk: the Multiethnic Cohort. *Human Molecular Genetics 12*, 20 (2003), 2679–2692.

[34] HAN BUHM, AND ESKIN ELEAZAR. Random-Effects Model Aimed at Discovering Associations in Meta-Analysis of Genome-wide Association Studies. *The American Journal of Human Genetics 88*, 5 (2011), 586–598.

[35] HAN BUHM, AND ESKIN ELEAZAR. Interpreting Meta-Analyses of Genome-Wide Association Studies. *PLoS Genetics 8*, 3 (March 2012), e1002555.

[36] HIGGINS JULIAN P. T., AND THOMPSON SIMON G. Quantifying Heterogeneity in a Meta-Analysis. *Statistics in Medicine 21*, 11 (2002), 1539–1558.

[37] HOWIE BRYAN, FUCHSBERGER CHRISTIAN, STEPHENS MATTHEW, MARCHINI JONATHAN, AND ABECASIS GONCALO R. Fast and Accurate Genotype Imputation in Genome-Wide Association Studies Through Pre-Phasing. *Nature Genetics 44*, 8 (August 2012), 955–959.

[38] KAVVOURA FOTINI K., AND IOANNIDIS JOHN P. A. Methods for Meta-Analysis in Genetic Association Studies:a Review of their Potential and Pitfalls. *Human Genetics 123* (2008), 1–14.

[39] KETTUNEN JOHANNES, TUKIAINEN TARU, SARIN ANTTI-PEKKA, ORTEGA-ALONSO ALFREDO, TIKKANEN EMMI, LYYTIKAINEN LEO-PEKKA, KANGAS ANTTI J., SOININEN PASI, WURTZ PETER, SILANDER KAISA, DICK DANIELLE M., ROSE RICHARD J., SAVOLAINEN MARKKU J., VIIKARI JORMA, KAHONEN MIKA, LEHTIMAKI TERHO, PIETILAINEN KIRSI H., INOUYE MICHAEL, MCCARTHY MARK I., JULA ANTTI, ERIKSSON JOHAN, RAITAKARI OLLI T., SALOMAA VEIKKO, KAPRIO JAAKKO, JARVELIN MARJO-RIITTA, PELTONEN LEENA, PEROLA MARKUS, FREIMER NELSON B., ALA-KORPELA MIKA, PALOTIE AARNO, AND RIPATTI SAMULI. Genome-Wide Association Study Identifies Multiple Loci Influencing Human Serum Metabolite Levels. *Nature Genetics 44*, 3 (March 2012), 269–276.

[40] LEBREC JEREMIE J., STIJNEN THEO, AND VAN HOUWELINGEN HANS C. Dealing with Heterogeneity between Cohorts in Genomewide SNP Association Studies. *Statistical Applications in Genetics and Molecular Biology 9*, 1 (January 2010), 1544–6115.

[41] LEE, JONG-MIN, GILLIS1,TAMMY, SRINIDHI MYSORE,JAYALAKSHMI, ET AL. Common SNP-Based Haplotype Analysis of the 4p16.3 Huntington Disease Gene Region. *The American Journal of Human Genetics 90*, 3 (2012), 434–444.

[42] LEPPERT MARK F., WEISS ROBERT B., BAKER, TIMOTHY B., CANNON DALE S., VON NIEDERHAUSERN, ANDREW, DUNN, DIANE M., MATSUNAMI, NORI, SINGH NANDA A., BAIRD, LISA, COON, HILARY, MCMAHON, WILLIAM M., PIPER, MEGAN E., FIORE, MICHAEL C., SCHOLAND, MARY BETH, CONNETT, JOHN E., KANNER, RICHARD E., GAHRING, LORISE C., HOIDAL, JOHN R., ROGERS, SCOTT W., ET AL. A Candidate Gene Approach Identifies the CHRNA5-A3-B4 Region as a Risk Factor for Age-Dependent Nicotine Addiction. *PLoS Genetics 4*, 7 (July 2008), e1000125.

[43] LIU, JASON Z, AND MARCHINI JONATHAN. META. `http://mathgen.stats.ox.ac.uk/genetics\_software/meta/meta.html`, 2010. [Online; accessed: 13-February-2013].

[44] LIU JASON Z., TOZZI FEDERICA, WATERWORTH DAWN M., PILLAI SREEKUMAR G., MUGLIA PIERANDREA, MIDDLETON LEFKOS, BERRETTINI WADE, KNOUFF CHRISTOPHER W., YUAN XIN, WAEBER GERARD, VOLLENWEIDER PETER, PREISIG MARTIN, WAREHAM NICHOLAS J., ZHAO JING HUA, LOOS RUTH J F, BARROSO INES, KHAW KAY-TEE, GRUNDY SCOTT, BARTER PHILIP, MAHLEY ROBERT, KESANIEMI ANTERO, MCPHERSON RUTH, VINCENT JOHN B., STRAUSS JOHN, KENNEDY JAMES L., FARMER ANNE, MCGUFFIN PETER, DAY RICHARD, MATTHEWS KEITH, BAKKE PER, GULSVIK AMUND, LUCAE SUSANNE, ISING MARCUS, BRUECKL TANJA, HORSTMANN SONJA, WICHMANN H-ERICH, RAWAL RAJESH, DAHMEN NORBERT, LAMINA CLAUDIA, POLASEK OZREN, ZGAGA LINA, HUFFMAN JENNIFER, CAMPBELL SUSAN, KOONER JASPAL, CHAMBERS JOHN C., BURNETT MARY SUSAN, DEVANEY JOSEPH M., PICHARD AUGUSTO D., KENT KENNETH M., SATLER LOWELL, LINDSAY JOSEPH M., WAKSMAN RON, EPSTEIN STEPHEN, WILSON JAMES F., WILD SARAH H., CAMPBELL HARRY, VITART VERONIQUE, REILLY MUREDACH P., LI MINGYAO, QU LIMING, WILENSKY ROBERT, MATTHAI

WILLIAM, HAKONARSON HAKON H., RADER DANIEL J., FRANKE ANDRE, WITTIG MICHAEL, SCHAFER ARNE, UDA MANUELA, TERRACCIANO ANTONIO, XIAO XIANGJUN, BUSONERO FABIO, SCHEET PAUL, SCHLESSINGER DAVID, CLAIR DAVID ST, RUJESCU DAN, ABECASIS GONCALO R., GRABE HANS JORGEN, TEUMER ALEXANDER, VOLZKE HENRY, PETERSMANN ASTRID, JOHN ULRICH, RUDAN IGOR, HAYWARD CAROLINE, WRIGHT ALAN F., KOLCIC IVANA, WRIGHT BENJAMIN J., THOMPSON JOHN R., BALMFORTH ANTHONY J., HALL ALISTAIR S., SAMANI NILESH J., ANDERSON CARL A., AHMAD TARIQ, MATHEW CHRISTOPHER G., PARKES MILES, SATSANGI JACK, CAULFIELD MARK, MUNROE PATRICIA B., FARRALL MARTIN, DOMINICZAK ANNA, WORTHINGTON JANE, THOMSON WENDY, EYRE STEVE, BARTON ANNE, MOOSER VINCENT, FRANCKS CLYDE, MARCHINI JONATHAN, ET AL. Meta-Analysis and Imputation Refines the Association of 15q25 With Smoking Quantity. *Nature Genetics 42*, 5 (May 2010), 436–440.

[45] LIU JIMMY Z., ALMARRI MOHAMED A., GAFFNEY DANIEL J., MELLS GEORGE F., JOSTINS LUKE, CORDELL HEATHER J., DUCKER SAMANTHA J., DAY DARREN B., HENEGHAN MICHAEL A., NEUBERGER JAMES M., DONALDSON PETER T., BATHGATE ANDREW J., BURROUGHS ANDREW, DAVIES MERVYN H., JONES DAVID E., ALEXANDER GRAEME J., BARRETT JEFFREY C., SANDFORD RICHARD N., ANDERSON CARL A., ET AL. Dense fine-mapping study identifies new susceptibility loci for primary biliary cirrhosis. *Nature Genetics 44*, 10 (October 2012), 1137–1141. 10.1038/ng.2395.

[46] MAGI REEDIK, LINDGREN CECILIA M., AND MORRIS ANDREW P. Meta-Analysis of Sex-Specific Genome-Wide Association Studies. *Genetic Epidemiology 34*, 8 (2010), 846–853.

[47] MANNING ALISA K., HIVERT MARIE-FRANCE, SCOTT ROBERT A, GRIMSBY JONNA L., BOUATIA-NAJI NABILA, CHEN HAN, RYBIN DENIS, LIU CHING-TI,

BIELAK LAWRENCE F., PROKOPENKO INGA, AMIN NAJAF, BARNES DANIEL, CADBY GEMMA, HOTTENGA JOUKE-JAN, INGELSSON ERIK, JACKSON ANNE U., JOHNSON TOBY, KANONI STAVROULA, LADENVALL CLAES, LAGOU VASILIKI, LAHTI JARI, LECOEUR CECILE, LIU YONGMEI, MARTINEZ-LARRAD MARIA TERESA, MONTASSER MAY E., NAVARRO PAU, PERRY JOHN R. B., RASMUSSEN-TORVIK LAURA J., SALO PERTTU, SATTAR NAVEED, SHUNGIN DMITRY, STRAWBRIDGE RONA J., TANAKA TOSHIKO, VAN DUIJN CORNELIA, M., AN PING, MARIZA, D. A., ANDREWS JEANETTE S., ASPELUND THOR, ATALAY MUSTAFA, AULCHENKO YURII, BALKAU BEVERLEY, BANDINELLI STEFANIA, BECKMANN JACQUES S., BEILBY JOHN P., BELLIS CLAIRE, BERGMAN RICHARD N., BLANGERO JOHN, BOBAN MLADEN, BOEHNKE MICHAEL, BOERWINKLE ERIC, BONNYCASTLE LORI L., BOOMSMA DORRET I., BORECKI INGRID B., BOTTCHER YVONNE, BOUCHARD CLAUDE, BRUNNER ERIC, BUDIMIR DANIJELA, CAMPBELL HARRY, CARLSON OLGA, CHINES PETER S., CLARKE ROBERT, COLLINS FRANCIS S., CORBATON-ANCHUELO ARTURO, COUPER DAVID, ULF, D. F., DEDOUSSIS GEORGE V., DELOUKAS PANOS, DIMITRIOU MARIA, EGAN JOSEPHINE M., EIRIKSDOTTIR GUDNY, ERDOS MICHAEL R., ERIKSSON JOHAN G., EURY ELODIE, FERRUCCI LUIGI, FORD IAN, FOROUHI NITA G., FOX CAROLINE S., FRANZOSI MARIA GRAZIA, FRANKS PAUL W., FRAYLING TIMOTHY M., FROGUEL PHILIPPE, GALAN PILAR, ECO, D. G., GIGANTE BRUNA, GLAZER NICOLE L., GOEL ANUJ, GROOP LEIF, GUDNASON VILMUNDUR, HALLMANS GORAN, HAMSTEN ANDERS, HANSSON OLA, HARRIS TAMARA B., HAYWARD CAROLINE, HEATH SIMON, HERCBERG SERGE, HICKS ANDREW A., HINGORANI AROON, HOFMAN ALBERT, HUI JENNIE, HUNG JOSEPH, JARVELIN MARJO-RIITTA, JHUN MIN A., JOHNSON PAUL C. D., JUKEMA J. WOUTER, JULA ANTTI, KAO W. H., KAPRIO JAAKKO, KARDIA SHARON L. R., KEINANEN-KIUKAANNIEMI SIRKKA, KIVIMAKI MIKA, KOLCIC IVANA, KOVACS PETER, KUMARI MEENA, KUUSISTO JO-

HANNA, KYVIK KIRSTEN OHM, LAAKSO MARKKU, LAKKA TIMO, LANNFELT LARS, LATHROP G. MARK, LAUNER LENORE J., LEANDER KARIN, LI GUO, LIND LARS, LINDSTROM JAANA, LOBBENS STEPHANE, LOOS RUTH J F, LUAN JIAN'AN, LYSSENKO VALERIYA, MAGI REEDIK, MAGNUSSON PATRIK K. E., MARMOT MICHAEL, MENETON PIERRE, MOHLKE KAREN L., MOOSER VINCENT, MORKEN MARIO A., MILJKOVIC IVA, NARISU NARISU, O'CONNELL JEFF, ONG KEN K., OOSTRA BEN A., PALMER LYLE J., PALOTIE AARNO, PANKOW JAMES S., PEDEN JOHN F., PEDERSEN NANCY L., PEHLIC MARINA, PELTONEN LEENA, PENNINX BRENDA, PERICIC MARIJANA, PEROLA MARKUS, PERUSSE LOUIS, PEYSER PATRICIA A., POLASEK OZREN, PRAMSTALLER PETER P., PROVINCE MICHAEL A., RAIKKONEN KATRI, RAURAMAA RAINER, REHNBERG EMIL, RICE KEN, ROTTER JEROME I., RUDAN IGOR, RUOKONEN AIMO, SAARISTO TIMO, SABATER-LLEAL MARIA, SALOMAA VEIKKO, SAVAGE DAVID B., SAXENA RICHA, SCHWARZ PETER, SEEDORF UDO, SENNBLAD BENGT, SERRANO-RIOS MANUEL, SHULDINER ALAN R., SIJBRANDS ERIC J. G., SISCOVICK DAVID S., SMIT JOHANNES H., SMALL KERRIN S., SMITH NICHOLAS L., SMITH ALBERT VERNON, STANCAKOVA ALENA, STIRRUPS KATHLEEN, STUMVOLL MICHAEL, SUN YAN V., SWIFT AMY J., TONJES ANKE, TUOMILEHTO JAAKKO, TROMPET STELLA, UITTERLINDEN ANDRE G., UUSITUPA MATTI, VIKSTROM MAX, VITART VERONIQUE, VOHL MARIE-CLAUDE, VOIGHT BENJAMIN F., VOLLENWEIDER PETER, WAEBER GERARD, WATERWORTH DAWN M., WATKINS HUGH, WHEELER ELEANOR, WIDEN ELISABETH, WILD SARAH H., WILLEMS SARA M., WILLEMSEN GONNEKE, WILSON JAMES F., WITTEMAN JACQUELINE C. M., WRIGHT ALAN F., YAGHOOTKAR HANIEH, ZELENIKA DIANA, ZEMUNIK TATIJANA, ZGAGA LINA, WAREHAM NICHOLAS J., MCCARTHY MARK I., BARROSO INES, WATANABE RICHARD M., FLOREZ JOSE C., DUPUIS JOSEE, MEIGS JAMES B., LANGENBERG CLAUDIA, ET AL. A Genome-Wide Approach Accounting for Body Mass Index Identifies Genetic Vari-

ants Influencing Fasting Glycemic Traits and Insulin Resistance. *Nature Genetics 44*, 6 (June 2012), 659–669.

[48] MANOLIO TERI A., BAILEY-WILSON JOAN E., AND COLLINS FRANCIS S. Genes, Environment and the Value of Prospective Cohort Studies. *Nature Reviews Genetics 7*, 10 (October 2006), 812–820.

[49] MANOLIO TERI A., COLLINS FRANCIS S., COX NANCY J., GOLDSTEIN DAVID B., HINDORFF LUCIA A., HUNTER DAVID J., MCCARTHY MARK I., RAMOS ERIN M., CARDON LON R., CHAKRAVARTI ARAVINDA, CHO JUDY H., GUTTMACHER ALAN E., KONG AUGUSTINE, KRUGLYAK LEONID, MARDIS ELAINE, ROTIMI CHARLES N., SLATKIN MONTGOMERY, VALLE DAVID, WHITTEMORE ALICE S., BOEHNKE MICHAEL, CLARK ANDREW G., EICHLER EVAN E., GIBSON GREG, HAINES JONATHAN L., MACKAY TRUDY F. C., MCCARROLL STEVEN A., AND VISSCHER PETER M. Finding the Missing Heritability of Complex Diseases. *Nature 461*, 7265 (October 2009), 747–753.

[50] MARCHINI JONATHAN. SNPTESTv2. `https://mathgen.stats.ox.ac.uk/genetics\_software/snptest/snptest.html`, 2012. [Online; accessed: 04-March-2013].

[51] MARCHINI JONATHAN, DONNELLY PETER, AND CARDON LON R. Genome-Wide Strategies for Detecting Multiple Loci That Influence Complex Diseases. *Nature Genetics 37*, 4 (April 2005), 413–417.

[52] MARCHINI JONATHAN, AND HOWIE BRYAN. Genotype Imputation for Genome-Wide Association Studies. *Nature Reviews Genetics 11*, 7 (June 2010), 499–511.

[53] MARKUS, P. Genome-Wide Association Approaches for Identifying Loci for Human Height Genes. *Best Practice & Research Clinical Endocrinology & Metabolism 25*, 1 (2011), 19–23.

[54] MCCARTHY MARK I., ABECASIS GONCALO R., CARDON LON R., GOLDSTEIN

DAVID B., LITTLE JULIAN, IOANNIDIS JOHN P. A., AND HIRSCHHORN JOEL N. Genome-Wide Association Studies for Complex Traits: Consensus, Uncertainty and Challenges. *Nature Review Genetics 9*, 5 (May 2008), 356–369.

[55] MOON, SUNGDAE, KIMA, JU-HEE, HANA, JE-HO, ET AL. Novel Compound Heterozygous Mutations in the Fructose-1,6-bisphosphatase Gene Cause Hypoglycemia and Lactic Acidosis. *Metabolism 60*, 1 (2011), 107–113.

[56] MORRIS ANDREW P. Direct Analysis of Unphased SNP Genotype Data in Population-Based Association Studies via Bayesian Partition Modelling of Haplotypes. *Genetic Epidemiology 29*, 2 (2005), 91–107.

[57] MORRIS ANDREW P. A Flexible Bayesian Framework for Modeling Haplotype Association with Disease, Allowing for Dominance Effects of the Underlying Causative Variants. *The American Journal of Human Genetics 79*, 4 (2006), 679–694.

[58] MORRIS ANDREW P. Transethnic Meta-Analysis of Genomewide Association Studies. *Genetic Epidemiology 35*, 8 (2011), 809–822.

[59] MORRIS ANDREW P., VOIGHT BENJAMIN F., TESLOVICH TANYA M., ET AL. Large-Scale Association Analysis Provides Insights into the Genetic Architecture and Pathophysiology of Type 2 Diabetes. *Nature Genetics 44*, 9 (September 2012), 981–990.

[60] NAKAOKA HIROFUMI, AND INOUE ITURO. Meta-Analysis of Genetic Association Studies: Methodologies, Between-Study Heterogeneity and Winner's Curse. *Journal of Human Genetics 54*, 11 (Oct. 2009), 615–623.

[61] NAYKOO NIYAZ A., HAMEEDD,IQRA, AASIF,MIR, ET AL. Single Nucleotide Polymorphisms, Haplotype Association and Tumour Expression of the Vascular Endothelial Growth Factor (VEGF) Gene with Lung Carcinoma. *Gene* (2013).

[62] NEUPANE BINOD, LOEB MARK, ANAND SONIA S., AND BEYENE JOSEPH. Meta-Analysis of Genetic Association Studies Under Heterogeneity. *European Jurnal of*

Human Genetics 20, 11 (November 2012), 1174–1181.

[63] NEWCOMBE PAUL J., VERZILLI CLAUDIO, CASAS JUAN P., HINGORANI AROON D., SMEETH,LIAM, AND WHITTAKER JOHN C. Multilocus Bayesian Meta-Analysis of Gene-Disease Associations. *The American Journal of Human Genetics 84*, 5 (2009), 567–580.

[64] NJAJOU OMER T., HOUWING-DUISTERMAAT JEANINE J., OSBORNE RICHARD H., VAESSEN NORBERT, VERGEER JEANETTE, HEERINGA JAN, POLS HUIBERT A.P., HOFMAN ALBERT, VAN DUIJN CORNELIA M, ET AL. A Population-Based Study of the Effect of the HFE C282Y and H63D Mutations on Iron Metabolism. *European Journal of Human Genetics 11*, 3 (2003), 225–231.

[65] OKADA YUKINORI, SIM XUELING, GO MIN JIN, WU JER-YUARN, GU DONGFENG, TAKEUCHI FUMIHIKO, TAKAHASHI ATSUSHI, MAEDA SHIRO, TSUNODA TATSUHIKO, CHEN PENG, LIM SU-CHI, WONG TIEN-YIN, LIU JIAN-JUN, YOUNG TERRI L., AUNG TIN, SEIELSTAD MARK, TEO YIK-YING, KIM YOUNG JIN, LEE JONG-YOUNG, HAN BOK-GHEE, KANG DAEHEE, CHEN CHIEN-HSIUN, TSAI FUU-JEN, CHANG LI-CHING, FANN S-J CATHY, MEI HAO, RAO DABEERU C., HIXSON JAMES E., CHEN SHUFENG, KATSUYA TOMOHIRO, ISONO MASATO, OGIHARA TOSHIO, CHAMBERS JOHN C., ZHANG WEIHUA, KOONER JASPAL S., ALBRECHT EVA, YAMAMOTO KAZUHIKO, KUBO MICHI-AKI, NAKAMURA YUSUKE, KAMATANI NAOYUKI, KATO NORIHIRO, HE JIANG, CHEN YUAN-TSONG, CHO YOON SHIN, TAI E-SHYONG, TANAKA TOSHIHIRO, ET AL. Meta-Analysis Identifies Multiple Loci Associated With Kidney Function-Related Traits in East Asian Populations. *Nature Genetics 44*, 8 (August 2012), 904–909.

[66] PORCU ELEONORA, MEDICI MARCO, PISTIS GIORGIO, VOLPATO CLAUDIA B., WILSON SCOTT G., CAPPOLA ANNE R., BOS STEFFAN D., DEELEN JORIS, MARTIN, D. H., FREATHY RACHEL M., LAHTI JARI, LIU CHUNYU, LOPEZ

LORNA M., NOLTE ILJA M., O'CONNELL JEFFREY R., TANAKA TOSHIKO, TROMPET STELLA, ARNOLD ALICE, BANDINELLI STEFANIA, BEEKMAN MARIAN, BÖHRINGER STEFAN, BROWN SUZANNE J., BUCKLEY BRENDAN M., CAMASCHELLA CLARA, M., D. C. A. J., DAVIES GAIL, H., D. V. M. C., FORD IAN, FORSEN TOM, FRAYLING TIMOTHY M., FUGAZZOLA LAURA, GÖGELE MARTIN, HATTERSLEY ANDREW T., HERMUS AD R., HOFMAN ALBERT, HOUWING-DUISTERMAAT JEANINE J., JENSEN RICHARD A., KAJANTIE EERO, KLOPPENBURG MARGREET, LIM EE M., MASCIULLO CORRADO, MARIOTTI STEFANO, MINELLI COSETTA, MITCHELL BRAXTON D., NAGARAJA RAMAIAH, NETEA-MAIER ROMANA T., PALOTIE AARNO, PERSANI LUCA, PIRAS MARIA G., PSATY BRUCE M., RÄIKKÖNEN KATRI, RICHARDS J. BRENT, RIVADENEIRA FERNANDO, SALA CINZIA, SABRA MONA M., SATTAR NAVEED, SHIELDS BEVERLEY M., SORANZO NICOLE, STARR JOHN M., STOTT DAVID J., SWEEP FRED C. G. J., USALA GIANLUCA, M., V. D. K. M., DIANA, V. H., ALIES, V. M., H.VERMEULEN SITA, VISSER W. EDWARD, WALSH JOHN P., WESTENDORP RUDI G. J., WIDEN ELISABETH, ZHAI GUANGJU, CUCCA FRANCESCO, DEARY IAN J., ERIKSSON JOHAN G., FERRUCCI LUIGI, FOX CAROLINE S., JUKEMA J. WOUTER, KIEMENEY LAMBERTUS A., PRAMSTALLER PETER P., SCHLESSINGER DAVID, SHULDINER ALAN R., SLAGBOOM ELINE P., UITTERLINDEN ANDRÉ G., VAIDYA BIJAY, VISSER THEO J., WOLFFENBUTTEL BRUCE H. R., MEULENBELT INGRID, ROTTER JEROME I., SPECTOR TIM D., HICKS ANDREW A., TONIOLO DANIELA, SANNA SERENA, PEETERS ROBIN P., NAITZA SILVIA, ET AL. A Meta-Analysis of Thyroid-Related Traits Reveals Novel Loci and Gender-Specific Differences in the Regulation of Thyroid Function. *PLoS Genetics 9*, 2 (February 2013), e1003266.

[67] PRITCHARD JONATHAN K., STEPHENS MATTHEW, AND DONNELLY PETER. Inference of Population Structure Using Multilocus Genotype Data. *Genetics 155*, 2

(June 2000), 945–959.

[68] PURCELL SHAUN, NEALE BENJAMIN, TODD-BROWN KATHE, THOMAS LORI, FERREIRA MANUEL A.R., BENDER DAVID, MALLER JULIAN, SKLAR PAMELA, DE BAKKER PAUL I.W., DALY MARK J., AND SCHAM PAK C. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics 81*, 3 (2007), 559–575.

[69] SACCONE NANCY L., CULVERHOUSE ROBERT C., SCHWANTES-AN TAE-HWI, CANNON DALE S., CHEN XIANGNING, CICHON SVEN, GIEGLING INA, HAN SHIZHONG, HAN YOUNGHUN, KESKITALO-VUOKKO KAISU, KONG XIANGYANG, LANDI MARIA TERESA, MA JENNIE Z., SHORT SUSAN E., STEPHENS SARAH H., STEVENS VICTORIA L., SUN LINGWEI, WANG YUFEI, WENZLAFF ANGELA S., AGGEN STEVEN H., BRESLAU NAOMI, BRODERICK PETER, CHATTERJEE NILANJAN, CHEN JINGCHUN, HEATH ANDREW C., HELIÖVAARA MARKKU, HOFT NICOLE R., HUNTER DAVID J., JENSEN MAJKEN K., MARTIN NICHOLAS G., MONTGOMERY GRANT W., NIU TIANHUA, PAYNE THOMAS J., PELTONEN LEENA, PERGADIA MICHELE L., RICE JOHN P., SHERVA RICHARD, SPITZ MARGARET R., SUN JUZHONG, WANG JEN C., WEISS ROBERT B., WHEELER WILLIAM, WITT STEPHANIE H., YANG BAO-ZHU, CAPORASO NEIL E., EHRINGER MARISSA A., EISEN TIM, GAPSTUR SUSAN M., GELERNTER JOEL, HOULSTON RICHARD, KAPRIO JAAKKO, KENDLER KENNETH S., KRAFT PETER, LEPPERT MARK F., LI MING D., MADDEN PAMELA A. F., NÖTHEN MARKUS M., PILLAI SREEKUMAR, RIETSCHEL MARCELLA, RUJESCU DAN, SCHWARTZ ANN, AMOS CHRISTOPHER I., BIERUT LAURA J., ET AL. Multiple Independent Loci at Chromosome 15q25.1 Affect Smoking Quantity: a Meta-Analysis and Comparison with Lung Cancer and COPD. *PLoS Genetics 6*, 8 (August 2010), e1001053.

[70] SACCONE NANCY L., WANG, JEN C., BRESLAU, NAOMI, JOHNSON, ERIC O.,

HATSUKAMI, DOROTHY, SACCONE, SCOTT F., GRUCZA, RICHARD A., SUN, LINGWEI, DUAN, WEIMIN, BUDDE, JOHN, CULVERHOUSE ROBERT C., FOX, LOUIS, HINRICHS, ANTHONY L., STEINBACH JOSEPH HENRY, WU, MENG, RICE, JOHN P., GOATE, ALISON M., BIERUT LAURA J., ET AL. The CHRNA5-CHRNA3-CHRNB4 Nicotinic Receptor Subunit Gene Cluster Affects Risk for Nicotine Dependence in African-Americans and in European-Americans. *Cancer Research 69*, 17 (2009), 6848–6856.

[71] SAMUELS MARK, ANDLYNCH, MARY E., TE MORSCHE, RENE H.M., AND DRENTH, JOOST P.H. Compound Heterozygosity in Sodium Channel Nav1.7 in a Family With Hereditary Erythermalgia. *Molecular Pain 4*, 1 (2008), 21.

[72] SCHAID DANIEL J. Evaluating Associations of Haplotypes With Traits. *Genetic Epidemiology 27*, 4 (2004), 348–364.

[73] SCHEET PAUL, AND STEPHENS MATTHEW. A Fast and Flexible Statistical Model for Large-Scale Population Genotype Data: Applications to Inferring Missing Genotypes and Haplotypic Phase. *The American Journal of Human Genetics 78*, 4 (2006), 629–644.

[74] SCOTT LAURA J., MUGLIA PIERANDREA, KONG XIANGYANG Q., GUAN, WEIHUA, UPMANYU RUCHI, ET AL. Genome-Wide Association and Meta-Analysis of Bipolar Disorder in Individuals of European Ancestry. *Proceedings of the National Academy of Sciences 106*, 18 (2009), 7501–7506.

[75] SERVIN BERTRAND, AND STEPHENS MATTHEW. Imputation-Based Analysis of Association Studies: Candidate Regions and Quantitative Traits. *PLoS Genetics 3*, 7 (July 2007), e114.

[76] SHEA JESSICA, AGARWALA VINEETA, PHILIPPAKIS ANTHONY A., MAGUIRE JARED, BANKS ERIC, DEPRISTO MARK, THOMSON BRIAN, GUIDUCCI CANDACE, ONOFRIO ROBERT C., KATHIRESAN SEKAR, GABRIEL STACEY, BURTT

NOEL P., DALY MARK J., GROOP LEIF, ALTSHULER DAVID, ET AL. Comparing Strategies to Fine-Map the Association of Common SNPs at Chromosome 9p21 With Type 2 Diabetes and Myocardial Infarction. *Nature Genetics 43*, 8 (August 2011), 801–805.

[77] SHIMIZU HIROSHI, TAKIZAWA,YASUKO, PULKKINEN,LEENA, ET AL. Epidermolysis Bullosa Simplex Associated With Muscular Dystrophy: Phenotype-Genotype Correlations and Review of the Literature. *Journal of the American Academy of Dermatology 41*, 6 (1999), 950–956.

[78] SLADEK ROBERT, ROCHELEAU GHISLAIN, RUNG JOHAN, DINA CHRISTIAN, SHEN LISHUANG, SERRE DAVID, BOUTIN PHILIPPE, VINCENT DANIEL, BELISLE ALEXANDRE, HADJADJ SAMY, BALKAU BEVERLEY, HEUDE BARBARA, CHARPENTIER GUILLAUME, HUDSON THOMAS J., MONTPETIT ALEXANDRE, PSHEZHETSKY ALEXEY V., PRENTKI MARC, POSNER BARRY I., BALDING DAVID J., MEYRE DAVID, POLYCHRONAKOS CONSTANTIN, FROGUEL PHILIPPE, ET AL. A Genome-Wide Association Study Identifies Novel Risk Loci for Type 2 Diabetes. *Nature 445*, 7130 (February 2007), 881–885.

[79] SORENSEN DANIEL, AND GIANOLA DANIEL. *Likelihood, Bayesian, and MCMC Methods in Quantitative Genetics*. Springer, 2002.

[80] SPEED, DOUG, AND TAVARÉ SIMON. Sparse Partitioning: Nonlinear Regression With Binary or Tertiary Predictors, With Application to Association Studies. *Annals of Applied Statistics 5*, 2A (2011), 873–893.

[81] SPELIOTES ELIZABETH K, WILLER CRISTEN J, BERNDT SONJA I., MONDA KERI L., THORLEIFSSON GUDMAR, JACKSON ANNE U., ALLEN HANA LANGO, LINDGREN CECILIA M., LUAN JIAN'AN, MAGI REEDIK, RANDALL JOSHUA C., VEDANTAM SAILAJA, WINKLER THOMAS W., QI LU, WORKALEMAHU TSEGASELASSIE, HEID IRIS M., STEINTHORSDOTTIR VALGERDUR, STRINGHAM HEATHER M., WEEDON MICHAEL N., WHEELER ELEANOR, WOOD AN-

DREW R., FERREIRA TERESA, ET AL. Association Analyses of 249,796 Individuals Reveal 18 New Loci Associated With Body Mass Index. *Nature Genetics 42*, 11 (November 2010), 937–948.

[82] SPENCER CHRIS C. A., SU ZHAN, DONNELLY PETER, AND MARCHINI JONATHAN. Designing Genome-Wide Association Studies: Sample Size, Power, Imputation, and the Choice of Genotyping Chip. *PLoS Genetics 5*, 5 (May 2009), e1000477.

[83] STAHL ELI A., RAYCHAUDHURI SOUMYA, REMMERS ELAINE F., XIE GANG, EYRE STEPHEN, THOMSON BRIAN P., LI YONGHONG, KURREEMAN FINA A. S., ZHERNAKOVA ALEXANDRA, HINKS ANNE, GUIDUCCI CANDACE, CHEN ROBERT, ALFREDSSON LARS, AMOS CHRISTOPHER I., ARDLIE KRISTIN G., BARTON ANNE, BOWES JOHN, BROUWER ELISABETH, BURTT NOEL P., CATANESE JOSEPH J., COBLYN JONATHAN, COENEN MARIEKE J. H., COSTENBADER KAREN H., CRISWELL LINDSEY A., CRUSIUS J. BART A., CUI JING, W., D. B. P. I., DE JAGER PHILIP L., DING BO, EMERY PAUL, FLYNN EDWARD, HARRISON PILLE, HOCKING LYNNE J., HUIZINGA TOM W. J., KASTNER DANIEL L., KE XIAYI, LEE ANNETTE T., LIU XIANGDONG, MARTIN PAUL, MORGAN ANN W., PADYUKOV LEONID, POSTHUMUS MARCEL D., RADSTAKE TIMOTHY R. D. J., REID DAVID M., SEIELSTAD MARK, SELDIN MICHAEL F., SHADICK NANCY A., STEER SOPHIA, TAK PAUL P., THOMSON WENDY, VAN DER HELM-VAN MIL ANNETTE, H. M., VAN DER HORST-BRUINSMA IRENE, E., ELLEN, V. D. S. C., VAN RIEL PIET, L. C. M., WEINBLATT MICHAEL E., WILSON ANTHONY G., WOLBINK GERT JAN, WORDSWORTH B. PAUL, WIJMENGA CISCA, KARLSON ELIZABETH W., TOES RENE E. M., NIEK, D. V., BEGOVICH ANN B., WORTHINGTON JANE, SIMINOVITCH KATHERINE A., GREGERSEN PETER K., KLARESKOG LARS, PLENGE ROBERT M., ET AL. Genome-Wide Association Study Meta-Analysis Identifies Seven New Rheumatoid Arthritis Risk Loci.

Nature Genetics 42, 6 (June 2010), 508–514.

[84] STAHL ELI A., RAYCHAUDHURI SOUMYA, REMMERS ELAINE F., XIE GANG, EYRE STEPHEN, THOMSON BRIAN P., LI YONGHONG, KURREEMAN FINA A. S., ZHERNAKOVA ALEXANDRA, HINKS ANNE, GUIDUCCI CANDACE, CHEN ROBERT, ALFREDSSON LARS, AMOS CHRISTOPHER I., ARDLIE KRISTIN G., BARTON ANNE, BOWES JOHN, BROUWER ELISABETH, BURTT NOEL P., CATANESE JOSEPH J., COBLYN JONATHAN, COENEN MARIEKE J. H., COSTENBADER KAREN H., CRISWELL LINDSEY A., CRUSIUS J. BART A., CUI JING, W., D. B. P. I., DE JAGER PHILIP L., DING BO, EMERY PAUL, FLYNN EDWARD, HARRISON PILLE, HOCKING LYNNE J., HUIZINGA TOM W. J., KASTNER DANIEL L., KE XIAYI, LEE ANNETTE T., LIU XIANGDONG, MARTIN PAUL, MORGAN ANN W., PADYUKOV LEONID, POSTHUMUS MARCEL D., RADSTAKE TIMOTHY R. D. J., REID DAVID M., SEIELSTAD MARK, SELDIN MICHAEL F., SHADICK NANCY A., STEER SOPHIA, TAK PAUL P., THOMSON WENDY, VAN DER HELM-VAN MIL ANNETTE, H. M., VAN DER HORST-BRUINSMA IRENE, E., ELLEN, V. D. S. C., VAN RIEL PIET, L. C. M., WEINBLATT MICHAEL E., WILSON ANTHONY G., WOLBINK GERT JAN, WORDSWORTH B. PAUL, WIJMENGA CISCA, KARLSON ELIZABETH W., TOES RENE E. M., NIEK, D. V., BEGOVICH ANN B., WORTHINGTON JANE, SIMINOVITCH KATHERINE A., GREGERSEN PETER K., KLARESKOG LARS, PLENGE ROBERT M., ET AL. Genome-Wide Association Study Meta-Analysis Identifies Seven New Rheumatoid Arthritis Risk Loci. *Nature Genetics 42*, 6 (June 2010), 508–514.

[85] STEPHENS MATTHEW, AND BALDING DAVID J. Bayesian Statistical Methods for Genetic Association Studies. *Nature Reviews Genetics 10*, 10 (October 2009), 681–690.

[86] STEPHENS MATTHEW, SMITH NICHOLAS J., AND DONNELLY PETER. A New Statistical Method for Haplotype Reconstruction from Population Data. *The Ameri-*

can Journal of Human Genetics 68, 4 (2001), 978–989.

[87] STRANGER BARBARA E., MONTGOMERY STEPHEN B., DIMAS ANTIGONE S., PARTS LEOPOLD, STEGLE OLIVER, INGLE CATHERINE E., SEKOWSKA MAGDA, SMITH GEORGE DAVEY, EVANS DAVID, GUTIERREZ-ARCELUS MARIA, PRICE ALKES, RAJ TOWFIQUE, NISBETT JAMES, NICA ALEXANDRA C., BEAZLEY CLAUDE, DURBIN RICHARD, DELOUKAS PANOS, AND DERMITZAKIS EMMANOUIL T. Patterns of Cis Regulatory Variation in Diverse Human Populations. *PLoS Genetics 8*, 4 (April 2012), e1002639.

[88] SU ZHAN, CARDIN NIALL, THE WELLCOME TRUST CASE CONTROL CONSORTIUM, DONNELLY PETER, AND MARCHINI JONATHAN. A Bayesian Method for Detecting and Characterizing Allelic Heterogeneity and Boosting Signals in Genome-Wide Association Studies. *Statistical Science 24*, 4 (2009), 430–450.

[89] SU ZHAN, AND MARCHINI JONATHAN. HAPGEN. `https://mathgen.stats.ox.ac.uk/genetics\_software/hapgen/hapgen2.html`, 2011. [Online; accessed: 22-February-2013].

[90] TABOR HOLLY K., RISCH NEIL J., AND MYERS RICHARD M. Candidate-Gene Approaches for Studying Complex Genetic Traits: Practical Considerations. *Nature Review Genetics 3*, 5 (May 2002), 391–397.

[91] TAN, JIN AI, KOK, JUAN LOONG, TAN, KIM LIAN, ET AL. Thalassemia Intermedia in HbH-CS Disease With Compound Heterozygosity for Beta-Thalassemia: Challenges in Hemoglobin Analysis and Clinical Diagnosis. *Genes Genet. Syst. 84* (2009), 67–71.

[92] TEO YIK-YING, SMALL KERRIN S., AND KWIATKOWSKI DOMINIC P. Methodological Challenges of Genome-Wide Association Analysis in Africa. *Nature Reviews Genetics 11*, 2 (February 2010), 149–160.

[93] TEWHEY RYAN, BANSAL VIKAS, TORKAMANI ALI, TOPOL ERIC J., AND

SCHORK NICHOLAS J. The Importance of Phase Information for Human Genomics. *Nature Review Genetics 12*, 3 (March 2011), 215–223.

[94] THE 1000 GENOMES PROJECT CONSORTIUM. A Map of Human Genome Variation from Population-scale Sequencing. *Nature 467*, 7319 (October 2010), 1061–1073.

[95] THE 1000 GENOMES PROJECT CONSORTIUM. An Integrated Map of Genetic Variation from 1,092 Human Genomes. *Nature 491*, 7422 (November 2012), 56–65.

[96] THE INTERNATIONAL HAPMAP3 CONSORTIUM. Integrating Common and Rare Genetic Variation in Diverse Human Populations. *Nature 467*, 7311 (September 2010), 52–58.

[97] THE WELLCOME TRUST CASE CONTROL CONSORTIUM. Genome-Wide Association Study of 14,000 Cases of Seven Common Diseases and 3,000 Shared Controls. *Nature 447*, 7145 (June 2007), 661–678.

[98] THE WELLCOME TRUST CASE CONTROL CONSORTIUM. Bayesian Refinement of Association Signals for 14 Loci in 3 Common Diseases. *Nature Genetics 44*, 12 (December 2012), 1294–1301.

[99] THOMAS DUNCAN. Gene–Environment-Wide Association Studies: Emerging Approaches. *Nature Review Genetics 11*, 4 (March 2010), 259–272.

[100] TIERNEY LUKE. Markov Chains for Exploring Posterior Distributions. *The Annals of Statistics* (1994).

[101] TWEE-HEE ONG RICK, WANG XU, LIU XUANYAO, AND TEO YIK-YING. Efficiency of Trans-Ethnic Genome-Wide Meta-Analysis and Fine-Mapping. *European Journal of Human Genetics 20*, 12 (December 2012), 1300–1307.

[102] WALDRON E.R.B., WHITTAKER J.C., AND BALDING DAVID J. Fine Mapping of Disease Genes via Haplotype Clustering. *Genetic Epidemiology 30*, 2 (2006), 170–179.

[103] WANG JUN, L., WANG, YUN, GUO LIHENG, CAI SU-PING, DAI MEIZHI, YANG QIAONA, YU WENHAN, YAN NAIHONG, ZHOU XIAOMIN, FU JIN, GUO XINWU, AND AND, H. Exome Sequencing Identifies Compound Heterozygous Mutations in CYP4V2 in a Pedigree with Retinitis Pigmentosa. *PLoS ONE 7*, 5 (May 2012), e33673.

[104] WANG, YUFEI, BRODERICK, PETER, MATAKIDOU, ATHENA, EISEN, TIMOTHY, AND HOULSTON, RICHARD S. Chromosome 15q25 (CHRNA3-CHRNA5) Variation Impacts Indirectly on Lung Cancer Risk. *PLoS ONE 6*, 4 (April 2011), e19085.

[105] WEI WEN-HUA, HEMANI GIB, GYENESEI ATTILA, VITART VERONIQUE, NAVARRO PAU, HAYWARD CAROLINE, CABRERA CLAUDIA P., HUFFMAN JENNIFER E., KNOTT SARA A., HICKS ANDREW A., RUDAN IGOR, PRAMSTALLER PETER P., WILD SARAH H., WILSON JAMES F., CAMPBELL HARRY, HASTIE NICHOLAS D., WRIGHT ALAN F., AND HALEY CHRIS S. Genome-Wide Analysis of Epistasis in Body Mass Index Using Multiple Human Populations. *European Jurnal Human Genetics 20*, 8 (August 2012), 857–862.

[106] WELCH KATHERINE O., MARIN RUTH S., PANDYA ARTI, AND ARNOS KATHLEEN S. Compound Heterozygosity for Dominant and Recessive GJB2 Mutations: Effect on Phenotype and Review of the Literature. *American Journal of Medical Genetics Part A 143A*, 14 (2007), 1567–1573.

[107] WILLER CRISTEN J., LI YUN, AND ABECASIS GONCALO R. METAL: Fast and Efficient Meta-Analysis of Genomewide Association Scans. *Bioinformatics 26*, 17 (2010), 2190–2191.

[108] WU YING, WAITE LINDSAY L., JACKSON ANNE U., SHEU WAYNE H-H., BUYSKE STEVEN, ABSHER DEVIN, ARNETT DONNA K., BOERWINKLE ERIC, BONNYCASTLE LORI L., CARTY CARA L., CHENG IONA, COCHRAN BARBARA, CROTEAU-CHONKA DAMIEN C., DUMITRESCU LOGAN, EATON CHARLES B., FRANCESCHINI NORA, GUO XIUQING, HENDERSON BRIAN E., HINDORFF LU-

CIA A., KIM ERIC, KINNUNEN LEENA, KOMULAINEN PIRJO, LEE WEN-JANE, LE MARCHAND LOIC, LIN YI, LINDSTRÖM JAANA, LINGAAS-HOLMEN ODDGEIR, MITCHELL SABRINA L., NARISU NARISU, ROBINSON JENNIFER G., SCHUMACHER FRED, STANČÁKOVÁ ALENA, SUNDVALL JOUKO, SUNG YUN-JU, SWIFT AMY J., WANG WEN-CHANG, WILKENS LYNNE, WILSGAARD TOM, YOUNG ALICIA M., ADAIR LINDA S., BALLANTYNE CHRISTIE M., BŮŽKOVÁ PETRA, CHAKRAVARTI ARAVINDA, COLLINS FRANCIS S., DUGGAN DAVID, FERANIL ALAN B., HO LOW-TONE, HUNG YI-JEN, HUNT STEVEN C., HVEEM KRISTIAN, JUANG JYH-MING J., KESÄNIEMI ANTERO Y., KUUSISTO JOHANNA, LAAKSO MARKKU, LAKKA TIMO A., LEE I-TE, LEPPERT MARK F., MATISE TARA C., MOILANEN LEENA, NJØLSTAD INGER, PETERS ULRIKE, QUERTERMOUS THOMAS, RAURAMAA RAINER, ROTTER JEROME I., SARAMIES JOUKO, TUOMILEHTO JAAKKO, UUSITUPA MATTI, WANG TZUNG-DAU, BOEHNKE MICHAEL, HAIMAN CHRISTOPHER A., CHEN YII-DER I., KOOPERBERG CHARLES, ASSIMES THEMISTOCLES L., CRAWFORD DANA C., HSIUNG CHAO A., NORTH KARI E., MOHLKE KAREN L., ET AL. Trans-Ethnic Fine-Mapping of Lipid Loci Identifies Population-Specific Signals and Allelic Heterogeneity That Increases the Trait Variance Explained. *PLoS Genetics 9*, 3 (March 2013), e1003379.

[109] ZEGGINI ELEFTHERIA, SCOTT LAURA J., SAXENA RICHA, VOIGHT BENJAMIN F., MARCHINI JONATHAN L., HU TIANLE, DE BAKKER PAUL I.W., ABECASIS GONCALO R., ALMGREN PETER, ANDERSEN GITTE, ARDLIE KRISTIN, BOSTROM KRISTINA BENGTSSON, BERGMAN RICHARD N., BONNYCASTLE LORI L., BORCH-JOHNSEN KNUT, BURTT NOEL P., CHEN HONG, CHINES PETER S., DALY MARK J., DEODHAR PARIMAL, DING CHIA-JEN, DONEY ALEX S. F., DUREN WILLIAM L., ELLIOTT KATHERINE S., ERDOS MICHAEL R., FRAYLING TIMOTHY M., FREATHY RACHEL M., GIANNINY LAUREN,

GRALLERT HARALD, GRARUP NIELS, GROVES CHRISTOPHER J., GUIDUCCI CANDACE, HANSEN TORBEN, HERDER CHRISTIAN, HITMAN GRAHAM A., HUGHES THOMAS E., ISOMAA BO, JACKSON ANNE U., JORGENSEN TORBEN, KONG AUGUSTINE, KUBALANZA KARI, KURUVILLA FINNY G., KUUSISTO JOHANNA, LANGENBERG CLAUDIA, LANGO HANA, LAURITZEN TORSTEN, LI YUN, LINDGREN CECILIA M., LYSSENKO VALERIYA, MARVELLE AMANDA F., MEISINGER CHRISTA, MIDTHJELL KRISTIAN, MOHLKE KAREN L., MORKEN MARIO A., MORRIS ANDREW D., NARISU NARISU, NILSSON PETER, OWEN KATHARINE R., PALMER COLIN N.A., PAYNE FELICITY, PERRY JOHN R. B., PETTERSEN ELIN, PLATOU CARL, PROKOPENKO INGA, QI LU, QIN LI, RAYNER NIGEL W., REES MATTHEW, ROIX JEFFREY J., SANDBAEK ANELLI, SHIELDS BEVERLEY, SJOGREN MARKETA, STEINTHORSDOTTIR VALGERDUR, STRINGHAM HEATHER M., SWIFT AMY J., THORLEIFSSON GUDMAR, THORSTEINSDOTTIR UNNUR, TIMPSON NICHOLAS J., TUOMI TIINAMAIJA, TUOMILEHTO JAAKKO, WALKER MARK, WATANABE RICHARD M., WEEDON MICHAEL N., WILLER CRISTEN J, ILLIG THOMAS, HVEEM KRISTIAN, HU FRANK B., LAAKSO MARKKU, STEFANSSON KARI, PEDERSEN OLUF, WAREHAM NICHOLAS J., BARROSO INES, HATTERSLEY ANDREW T., COLLINS FRANCIS S., GROOP LEIF, MCCARTHY MARK I., BOEHNKE MICHAEL, ALTSHULER DAVID, ET AL. Meta-Analysis of Genome-Wide Association Data and Large-Scale Replication Identifies Additional Susceptibility Loci for Type 2 Diabetes. *Nature Genetics 40*, 5 (May 2008), 638–645.