THE UNIVERSITY OF CHICAGO

BAYESIAN ANALYSIS OF GENETIC ASSOCIATION DATA, ACCOUNTING FOR HETEROGENEITY

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF STATISTICS

BY
XIAOQUAN WEN

CHICAGO, ILLINOIS
AUGUST 2011

To Mingming

and

In Memoriam of My Mom, Zezhi

ABSTRACT

In this dissertation research, we tackle the statistical problem of analyzing potentially heterogeneous genetic association data. Most frequently, this type of the data arise from applications of genetic meta-analysis and study of gene-environment interactions. These two types of applications are both critical for understanding the effects of genetic variants on complex traits.

We propose a unified Bayesian framework to deal with potentially heterogeneous genetic association data. Within this framework, We address the problems of *whether* and *how* a particular genetic variant act on the phenotype of interest by Bayesian testing and model comparison approaches in a systematic way. We propose Bayesian models, derive easy-to-compute Bayes Factors for this purpose and discuss the general strategy for exploratory analysis in these settings.

Built on these results, we discuss a special type of application from genomics research: mapping eQTLs across tissue types. At a single gene level, this is a special application of gene-environment interactions we have discussed above. Nevertheless, the special feature of this setting is that there are information shared by many genes which are simultaneously measured in a single experiment. We propose a hierarchical mixture model to "pool" the information across genes and investigate the scope of the tissue specificity of eQTLs and potential biological "features" that are associated with tissue specificity.

Finally, to deal with missing data in meta-analyses settings, we propose a linear predictor approach that can efficiently "impute" allele frequencies based on observed summary-level data. The main statistical novelty is that we find a very natural shrinkage estimator of a high dimensional covariance matrix by incorporating knowledge from population genetic models.

ACKNOWLEDGEMENTS

I would like to express my most sincere gratitude to my adviser, Matthew Stephens, for his superb guidance over the years. His limitless support, encouragement and patience have been invaluable. It has been truly my great honor to work with such an extraordinary statistician and scientist.

I would also like to thank the faculty members in Department of Statistics. Dan Nicolae encouraged and helped me to apply the program and constantly gives me great advice. We have been collaborating since an even earlier time, from which I benefited a great deal. Mary Sara McPeek, Michael Stein and Peter McCullagh all generously spent their time discussing the problems I met during my dissertation research, their input has greatly improved this thesis.

My special thank goes to Jonathan Pritchard and Nancy Cox, who led me into this wonderful field of statistical genetics.

I am also grateful to all the past and current members of Prichard, Przeworski and Stephens groups: Yongtao Guan, John Marioni, Peter Carbonetto, Bryan Howie, Barbara Engelhardt, Kevin Bullaughey, John Novembre, Daniel Gaffney, Joe Pickrell, Pall Melsted, Jacob Degner, Roger Pique-Reg, John Zekos, Sebastian Zoellner, Graham Coop, Jordana Bell, Don Conrad, Sridhar Kudaravalli, Jean-Baptiste Veyrieras, Adi Alon and many many more. They have created a wonderfully vigorous and collaborative environment, where biologists, computer scientists, mathematicians, statisticians and physicists can freely exchange ideas and learn from each other. I will miss the group meetings, science bombs and Friday beers.

Finally and most importantly, I thank my wife Mingming Zhang, for her love, trust, support and sacrifice. Without her, this dissertation simply would not have been possible.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

Genetic association studies have become increasingly popular for understanding the genetic basis of complex human diseases. The fast advancement of experimental technology now enables simultaneously interrogating millions of genetic variants in human genome. However, statistically identifying genetic variants that are associated with complex phenotypes is still intrinsically difficult. Moreover, to understand *how* genetic variants impact complex phenotypes is a even more daunting challenge. In this dissertation, we aim to address some of these issues and provide statistically sound and computationally efficient solutions.

1.1 Background

We start by introducing some relevant terminologies in genetics.

A **SNP** (Single Nucleotide Polymorphism) is a single base pair mutation occurred in a DNA sequence. The possible nucleotides that can be presented at a SNP location are known as **alleles**. Almost all SNPs have at most two alleles in a populations. We will use 0 and 1 to denote the two alleles at each SNP with the labeling being essentially arbitrary. SNP is the most common type of the genetic variants: up to today, about 10 million SNPs in the human genome have been cataloged.

A **haplotype** is a combination of alleles at multiple SNPs residing on a single copy of a genome. Each haplotype can be represented by a string of binary (0/1) values. Each individual has two haplotypes, one inherited from each parent. Routine technologies read the two haplotypes simultaneously to produce a measurement of the individual's **genotype** at each SNP, which can be coded as 0,1 or 2 copies of the "1" allele. Thus the haplotypes themselves are not usually directly observed, although they can be inferred using statistical methods (Stephens *et al.* (2001)). Genotype data where the haplotypes are treated as unknown are referred to as "unphased", whereas if the haplotypes are measured or estimated they are referred to as "phased".

Genetic Association refers to the correlation between genetic polymorphism and some phenotype of interest, which is typically some measurable trait (e.g. quantitative

trait like height, weight, blood pressure) or characteristics (e.g. case/control status of a particular disease). In medicine, it has been known for a very long time that some diseases such as heart disease, cancer, and diabetes run in families, and family history, as a proxy for genetics, has been commonly used in medical practice for disease diagnosis and prevention. Understanding the genetic basis of complex disease not only helps us to reveal the biological pathways to diseases, but also leads to developing new drugs/therapies for better treatments.

In a **genetic association study**, both genotypes (typically SNPs) and phenotype information are measured for a set of collected samples and the aim is to identify genetic variants that are associated with the phenotype of interest. In this dissertation, we focus exclusively on studies that collect population samples from *unrelated* individuals.

One common approach to analyzing genetic association data is to examine one SNP at a time. For quantitative phenotypes, a single SNP genetic association can be modeled by the following simple linear regression model:

$$\boldsymbol{y} = \mu\mathbf{1} + \beta\boldsymbol{g} + \boldsymbol{e}\ , \quad \boldsymbol{e} \sim \mathrm{N}(\mathbf{0}, \sigma^{\mathbf{2}}\mathbf{I}), \tag{1.1}$$

where $\boldsymbol{y}$ and $\boldsymbol{g}$ are vectors of sample phenotypes and genotypes (coded as 0, 1 or 2 for each individual) respectively, μ is the intercept term and σ^2 is the residual error variance. The quantity of interest is the regression coefficient β, also known as genetic effect: the value of β is non-zero if and only if an association exists.

1.2 Heterogeneous Genetic Association Data

In genetic association analysis, it is often desired to analyze data from multiple potentially-heterogeneous subgroups. There are primarily two types of application/study that concern heterogeneous genetic association data:

1. To detect modest genetic association signals that are too weak to be detected in smaller individual studies, meta-analysis of multiple studies are often required. These studies are typically carried out by different investigators, at different

centers, which might be expected to exhibit heterogeneity of genetic effects.

2. Some genetic variants exhibit different effect sizes under different environmental exposures. This is known as gene-environment (G×E) interaction. Identifying G×E interactions has become increasingly important to understand and explain phenotypic variation.

In both cases, the genetic association data are structured in subgroups: in meta-analysis, the subgroups are formed by samples from different studies; while in study of G×E interactions, samples from different environmental exposures are naturally clustered into subgroups. The scientific questions we are interested in can be framed at two different levels: First, we aim to identify genetic variants that are associated with the phenotype of interest in *any* of the subgroups; then, we might be interested in investigating the heterogeneity of the genetic effects in details for identified genetic variants.

For meta-analysis, the motivation is to increase statistical power by accumulating more samples through multiple studies. The perception of heterogeneity, at least *a priori*, is that the genetic effects among different studies should be quite consistent (Munafò and Flint (2004)). Whereas in detection of G×E interactions, the goal is completely different: we are looking for genetic variants whose effects have large variation in different environmental conditions (Hunter (2005)).

1.3 The Bayesian Approach

Our choice of Bayesian approaches to analyze potentially heterogeneous genetic association data is primarily motivated by pragmatic rather than philosophical considerations. In particular, we take advantages of convenient model comparison procedures built into the Bayesian framework.

In studies of potentially heterogeneous genetic association data, multiple levels of scientific inquires can be phrased as model comparison problems. For a phenotype of interest and a target SNP, firstly, we are interested in knowing whether the SNP is associated with the phenotype. This is typically solved as a hypothesis testing

problem, which can also be framed as a special case of model comparison (comparing some non-null alternative model with the null model stating no association). Once we find evidence for association, we may be interested in investigating the levels of heterogeneity of genetic effects exhibited in different subgroups. This can be achieved by comparing a set of models explicitly describing various degrees of heterogeneity. Finally, if there are scientific hypotheses that explicitly explain the heterogeneity, we can construct models based on these scientific hypotheses and assess their relative plausibilities using observed data.

To give a simple example of the explanation step, suppose there are two possible biological processes that lead to the phenotype of interest – one involving the target SNP, the other does not, and the activity of the biological process depends on the environmental conditions. Now, given two different environmental conditions, with observed genetic association data, we may infer the activities of the two candidate processes in these two conditions. To do this, we compare all four possible genetic models that describe the association between the phenotype and the target SNP:

1. target SNP has no association with the phenotype in either condition.

2. target SNP is associated with the phenotype in condition 1 but has no association in condition 2.

3. target SNP is associated with the phenotype in condition 2 but has no association in condition 1.

4. target SNP has are associated with the phenotype in both conditions.

If the data are sufficiently informative, we expect the true model to obtain the strongest support.

The Bayesian device that compares evidence in the data for two competing models is the *Bayes Factor*. The Bayes Factor of model M_i to model M_j is defined as

$$\mathrm{BF}_{ij} = \frac{\Pr(\mathrm{Data}|M_i)}{\Pr(\mathrm{Data}|M_j)}. \tag{1.2}$$

It also follows from this definition that the Bayes Factor of model M_i to another model M_k can be computed as

$$\mathrm{BF}_{ik} = \frac{\mathrm{BF}_{ij}}{\mathrm{BF}_{kj}}. \tag{1.3}$$

In principle, BF_{ij} captures all the evidence provided by the data in favor of or against model M_i vs. model M_j. If prior probabilities for models $\eta_i = \Pr(M_i), i = 1, 2, \ldots$ are specified, then posterior probabilities of models can be easily computed via Bayes Factors, for example,

$$\Pr(M_i|\mathrm{Data}) = \frac{\eta_i \mathrm{BF}_{ij}}{\sum_k \eta_k \mathrm{BF}_{kj}}. \tag{1.4}$$

In comparison, in a Frequentist framework, if the candidate models are not nested, the procedure is usually not straightforward. It is sometimes possible to compare models by computing various model selection criteria. However, these quantities usually rely heavily on asymptotic assumption, which in practice may not be appropriate. Further, simple model selection criteria, e.g. AIC, typically are not well-defined for hierarchical models which happen to be our choice to model potentially heterogeneous genetic association data.

There are many other advantages of Bayesian methodology over classical Frequentist approaches in the context of genetic association studies. Stephens and Balding (2009) has a thorough discussion and systematic review on this perspective. The same arguments naturally apply in our context as well.

1.4 Outline of the Dissertation

In this chapter, we have briefly described the problems regarding analyzing potentially-heterogeneous genetic association data and set out to search for Bayesian solutions.

In chapter 2, we describe our theoretical results on Bayesian approaches for analyzing heterogeneous genetic association data. In particular, we develop some computationally-efficient approaches to computing Bayes factors in these settings. One of these approaches yields a Bayes Factor with a simple and intuitive analytic

form. Various interesting properties of the Bayes Factors will be discussed and demonstrated through real data examples.

In chapter 3, we apply our Bayesian methodology to mapping tissue-specific eQTLs. eQTL (expression quantitative trait loci) are genetic variants that are associated with gene expression levels. Investigating tissue-specific eQTL yields great insights into mechanisms of differential gene regulation (Montgomery and Dermitzakis (2011)). Study of tissue-specific eQTLs can be viewed as a special case of investigation of G×E interaction. We further build a hierarchical mixture model to investigate the scope of the tissue specificity of eQTLs and potential biological "features" that are connected with tissue specificity.

In chapter 4, we describe an imputation method for untyped genetic variants when only summary-level information (e.g. allele frequencies) are available. This is a important statistical method in dealing with missing genetic data when combining genetic association data across multiple studies. In particular, we consider the situation when only summary-level genetic data (e.g. allele frequencies) are made available. Such scenarios can arise due to policy (e.g. for protecting privacy) or experimental design (e.g. data generated by DNA pooling experiment). Our proposed method can accurately infer the frequencies of untyped genetic variants in these settings, and indeed substantially improve frequency estimates at typed variants in pooling experiments where observations are noisy. Our approach, which predicts each allele frequency using a linear combination of observed frequencies, is statistically straightforward, and related to a long history of the use of linear methods for estimating missing values (e.g. Kriging). The main statistical novelty is our approach to regularizing the covariance matrix estimates, and the resulting linear predictors, which is based on methods from population genetics. We find that, besides being both fast and flexible – allowing new problems to be tackled that cannot be handled by existing imputation approaches purpose-built for the genetic context – these linear methods are also very accurate. Indeed, imputation accuracy using this approach is similar to that obtained by state-of-the art imputation methods that use individual-level data, but at a fraction of the computational cost.

CHAPTER 2
BAYESIAN METHODS FOR ANALYZING HETEROGENEOUS GENETIC ASSOCIATION DATA

2.1 Introduction

In this chapter, we present Bayesian methods for analyzing genetic association data, allowing for potential heterogeneity among (pre-specified) subgroups. We are motivated by two distinct settings where heterogeneity may arise. The first setting is meta-analysis of multiple association studies of the same phenotype. These studies are usually carried out by different investigators, at different centers, therefore, heterogeneity of genetic effects are commonly expected (e.g. due to differences in the way phenotypes are measured, or due to systematic differences between individuals enrolled in each study). Such meta-analyses have become an increasingly popular and important statistical tool for detecting modest genetic associations that are too small to be detected in smaller individual studies (Teslovich *et al.* (2010), Zeggini *et al.* (2008)). The second setting is where genuine biological interactions may cause some genetic variants to exhibit different effects on individuals in different subgroups; for example, genetic effects can differ in males and females even at autosomal loci (Kong *et al.* (2008), Ober *et al.* (2008)). And in gene expression analyses that aim to detect genetic variants associated with gene expression levels, data are often available on individuals from different continental groups Stranger *et al.* (2007), Veyrieras *et al.* (2008), or on different tissue types Dimas *et al.* (2009), where heterogeneity of effects may be expected.

These two settings differ in the extent of the heterogeneity expected: for example, interactions could cause genetic variants to have effects in different directions in different subgroups, however this might be considered unlikely in the meta-analysis setting. They also differ in the extent to which heterogeneity may be of direct interest (e.g. in interactions) or largely a "nuisance" (e.g. in meta-analysis). However, the two settings also share an important element in common: the vast majority of genetic variants are unassociated with any given phenotype of interest, within *all* subgroups.

Consequently, it is of great interest to identify genetic variants that show association in *any* subgroups (i.e. rejecting the "global" null hypothesis of *no association within any subgroup*). Similar points are made in more detail in Lebrec *et al.* (2010), which develops and compares frequentist tests for this problem.

Our proposed methods handle heterogeneous genetic association data in these two different settings within a unified Bayesian framework. The key idea here is to take a *model comparison* approach rather than solely focus on *hypothesis testing*. More specifically, we construct a set of models (with the null model included) in which heterogeneity of the genetic effects is explicitly modeled. For identifying genetic association, we evaluate the evidence in the data for the null model versus the non-null models. In case of investigating the details of heterogeneity given association presented in some subgroup, we can further compare the support from the data for each alternative model. In Bayesian framework, the essential statistical device required to accomplish both tasks is the Bayes Factor.

In the rest of the chapter, we first discuss considerations and approaches for modeling heterogeneity, then proceed to show the computation of Bayes Factors for a set of proposed models. Finally, we use various data examples to illustrate our proposed Bayesian methods.

2.2 Models and Methods

We start with models for quantitative traits, then proceed to discuss Bayesian analysis procedures based on those proposed models. Later, we generalize these results to binary outcomes in case-control studies.

2.2.1 Notation and Assumptions

We assume (quantitative) phenotype data and genotype data are available on S predefined subgroups, and focus on assessing association between the phenotype and each genetic variant (SNP), one at a time. We assume that the data within subgroup s come from n_s randomly-sampled unrelated individuals. Let the n_s-vectors $\boldsymbol{y}_s$ and $\boldsymbol{g}_s$ denote, respectively, the corresponding phenotype data and the genotype data at

a single "target" SNP. We also let $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_S)$ and $\boldsymbol{G} = (\boldsymbol{g}_1, \ldots, \boldsymbol{g}_S)$ denote the complete set of phenotype and genotype data respectively.

2.2.2 Hierarchical Models for Quantitative Traits

In this section, we introduce a set of models for describing heterogeneous genetic effects for a target SNP across subgroups.

Within each subgroup, we model the association between phenotype and genotype using a standard linear model. Specifically, in subgroup s, we assume

$$\boldsymbol{y}_s = \mu_s \mathbf{1} + \beta_s \boldsymbol{g}_s + \boldsymbol{e}_s \ , \quad \boldsymbol{e}_s \sim \mathrm{N}(0, \sigma_s^2 I). \tag{2.1}$$

Here, we also assume residual errors are independent across subgroups.

The "global" null hypothesis of interest is that there is no genotype-phenotype association within any subgroup; that is, $\beta_s = 0$ for all s.

Under the alternative hypothesis we begin by assuming that the genetic effects among subgroups are *exchangeable*, and more specifically that they are normally distributed about some unknown mean. We consider two different definitions of genetic effects: the "standardized effects" $b_s := \beta_s/\sigma_s$, and the unstandardized effects, β_s, leading to the following models:

1. *Exchangeable Standardized Effects (ES model).* Under this model we assume that the standardized effects b_s are normally distributed among subgroups:

$$b_s|\sigma_s \sim \mathrm{N}(\bar{b}, \phi^2) \ \text{ or equivalently, } \ \beta_s|\sigma_s \sim \mathrm{N}(\sigma_s\bar{b}, \sigma_s^2\phi^2), \tag{2.2}$$

   so the hyper-parameters $\bar{b}$ and ϕ characterize, respectively, the mean and variance of effects among subgroups. We also assume a normal prior distribution for $\bar{b}$,

$$\bar{b} \sim \mathrm{N}(0, \omega^2). \tag{2.3}$$

   Finally, for the parameters σ_s and μ_s, which are common to both the null and

alternative hypotheses, we use the convenient conjugate priors

$$\mu_s\sigma_s|\sigma_s \sim \mathrm{N}(0, \sigma_s^2 v_s^2); \quad \sigma_s^{-2} \sim \Gamma(m_s/2, l_s/2). \tag{2.4}$$

When performing inference we consider the limits $v_s^2 \to \infty$ and $l_s, m_s \to 0$, in which posteriors for both σ_s and μ_s are well-defined.

2. *Exchangeable Effects (EE model).* Under this model we assume that the unstandardized effects β_s are normally distributed:

$$\beta_s \sim \mathrm{N}(\bar{\beta}, \psi^2), \tag{2.5}$$

where $\bar{\beta}$ and ψ play similar roles to $\bar{b}$ and ϕ in the ES model. We also assume a normal prior for $\bar{\beta}$,

$$\bar{\beta} \sim \mathrm{N}(0, w^2), \tag{2.6}$$

and priors for (μ_s, σ_s):

$$\mu_s \sim \mathrm{N}(0, u_s^2); \quad \sigma_s^{-2} \sim \Gamma(m_s/2, l_s/2). \tag{2.7}$$

At subgroup level, this prior specification is very similar with the commonly used semi-conjugate priors in Bayesian linear regression.

Again, we use the limits $u_s^2 \to \infty$ and $l_s, m_s \to 0$, as discussed further below.

In both the ES and EE models the alternative hypothesis involves two key hyperparameters, one (ω in the ES model and w in the EE model) that controls the prior expected size of the average effect, and another (ϕ in the ES model and ψ in the EE model) that controls the prior expected degree of heterogeneity among subgroups. A complimentary view is that $\omega^2 + \phi^2$ (respectively, $w^2 + \psi^2$) controls the expected (marginal) effect size in each study and ϕ/ω (respectively, ψ/w) controls the degree of heterogeneity.

Of the two models, the ES model has the advantage that it results in analyses (e.g. Bayes Factors) that are invariant to the phenotype measurement scale used

within each subgroup. This not only makes it more robust to users accidentally specifying phenotype measurements in different subgroups on different scales (possibly a non-trivial issue in complex analyses involving collaboration among many research groups), but also means that it can be applied when measurement scales may be difficult to harmonize across subgroups, for example due to the use of different measurement technologies. For these reasons we prefer the ES model for general use. However, in some cases the EE model may be easier to apply. For example, if one has access only to published point estimates and standard errors for the effect size β_s in each study, then this suffices to approximate the Bayes Factor under the EE model, but not under the ES model. Note that the ES and EE models will produce similar results to each other if the residual error variances are similar in all subgroups.

A Curved Exponential Family Normal Prior

In some genetic applications, heterogeneity of effect sizes among subgroups are generally expected, but only with certain degree. For example, in meta-analysis, we may expect genuine genetic association to possess the property that effect sizes across studies predominantly show the same sign (Owen (2009)). To reflect this type of the prior belief in "constrained" heterogeneity, we introduce a novel curved exponential family normal (CEFN) prior: under ES model, we replace (2.2) with

$$b_s \sim \mathrm{N}(\bar{b}, k^2\bar{b}^2). \tag{2.8}$$

In this formulation, the prior mean and variance are functionally related. As a consequence, the distance from prior mean ($\bar{b}$) to origin 0 is measured as $\frac{1}{k}$ units of prior standard deviation. This relationship can be translated into the following probability statement,

$$\Pr\left(\mathrm{b}_s \text{ has a different sign from } \bar{b}\right) = \Phi(-\frac{1}{|k|}), \tag{2.9}$$

where Φ is the cumulative probability function of standard normal distribution. For example, when $k = 1/2$, sampling from this prior distribution, the probability of obtaining a value of b_s having an opposite sign to $\bar{b}$ is approximately 2.3%. As the

value of k decreases, the restriction becomes more stringent. When $k = 0$, the prior indicates all b_s are exactly same as $\bar{b}$, i.e. there is no heterogeneity of effects across subgroups.

For the EE model, a similar curved exponential family prior can be applied as

$$\beta_s \sim \mathrm{N}(\bar{\beta}, k^2\bar{\beta}^2). \tag{2.10}$$

2.2.3 Use of Proposed Models

This section concerns the general statistical strategy on choice of the models proposed in previous sections for analyzing potential-heterogeneous genetic data. We argue that the prior expectations of heterogeneity of genetic effects are typically context-dependent, and we should construct appropriate context-dependent models to reflect our beliefs.

In meta-analysis, for a genuine genetic association, genetic effects in different participating studies are expected to be similar, but not necessarily identical. At the same time, we also tend to believe the maximum degree of heterogeneity in effects should be constrained, for example, most of the effects in different subgroups are expected to be in same direction. Thus, we can construct a set of ES models with CEFN priors of small k values (or regular ES model with only "small" ϕ/ω values) and compare with the global null.

If subgroup labels can be regarded as proxies of environmental conditions, our first interest in a genome scan is still to reject the global null hypothesis of no association in any subgroups. Once we confirm there is some association existing in some subgroup, we then are interested in investigating the nature of the heterogeneity of genetic effects in different subgroups: are the effects fairly consistent or are they very different? If data give strong support to the latter case, we have likely found an instance of G×E interaction.

One possible implementation of this strategy is to construct a series of alternative ES models that span a range of expected marginal effect sizes ($\phi^2 + \omega^2$) and various degrees of heterogeneities (ϕ/ω). To compare with the global null model, we obtain the evidence from the data by averaging over all considered alternative models. To

further identify potential G×E interactions once the null is rejected, we compare the evidence for models with high heterogeneity versus low heterogeneity: in an instance of true G×E interaction, the genetic effects are expected to exhibit a large degree of heterogeneity in different environment conditions. Conversely, we would not claim to have identified a G×E interaction if the effects are "broadly consistent".

The above strategies should work well as a general guideline for exploratory data analysis. In case of G×E interaction, with more available information, it is possible to construct more explicit models that *explain* the heterogeneity. Consider a simple, hypothetical example of two mutually exclusive genetic pathways leading to the phenotype of interest. Suppose a target SNP only actively involves in pathway A (therefore, is associated with the phenotype) but not in pathway B, and the activity of the pathways depends on different environmental conditions. Given the genetic association data, if the correspondence between environmental conditions and pathway activities are *known*, it is very natural to separately model the genetic effects of the target SNP in different environmental conditions according to the activities of the pathways: for the group of environmental conditions where pathway A is active, we can model the genetic effects within this group using a meta-analysis approach (e.g. the ES model with low degrees of heterogeneities); whereas for the group of conditions where pathway B is active, the genetic effects there should be independently described by the null model. Essentially, we cluster the genetic effects (into a zero cluster and a non-zero cluster) according to the activities of the pathways, thus explicitly explain the variation of genetic effects between clusters. In practice, the relationship between environment condition and pathway activity is typically unknown (i.e. the cluster membership of each genetic effect under certain environmental condition is latent) and may be of great interest. In principle, we can enumerate all possible environmental condition-pathway correspondence relationships and evaluate them by computing the support from observed data. We will show such an example in cross-population eQTL mapping later in this chapter.

2.2.4 Bayes Factors for Testing the Global Null Hypothesis

We now derive Bayes Factors for testing the "global" null hypothesis that the phenotype is not associated with the target SNP in any of the subgroups, versus the alternative hypotheses (ES and EE) outlined above.

Starting with the ES model, recall that this model is indexed by two parameters, ϕ and ω. Within this model, the global null hypothesis, which is most naturally written as $\beta_s \equiv 0$ for all s, can also be written as

$$H_0 : \phi = \omega = 0. \tag{2.11}$$

To compare the support in the data for this null hypothesis against a particular alternative ES model specified by parameters (ϕ, ω), we use the Bayes Factor:

$$\mathrm{BF}^{\mathrm{ES}}(\phi, \omega) = \frac{P(\boldsymbol{Y}|\boldsymbol{G}, \phi, \omega)}{P(\boldsymbol{Y}|\boldsymbol{G}, H_0)}. \tag{2.12}$$

Note that this Bayes factor depends on the values of the prior hyper-parameters v_s, l_s and m_s; however, because these hyper-parameters are common to both the null and alternative hypotheses, the value of the Bayes Factor is not especially sensitive to the values chosen. As noted above we take the limits

$$v_s^2 \to \infty,\ l_s \to 0,\ m_s \to 0,\ \forall s. \tag{2.13}$$

Each value of ω, ϕ corresponds to a particular alternative model, with ω controlling the typical average effect size, and ϕ controlling the degree of heterogeneity among subgroups (or in a re-parameterization, $\omega^2 + \phi^2$ controls the expected marginal effect size in each subgroup and ϕ/ω controls the degree of heterogeneity). As discussed in section 2.2.3, there may be reasonable uncertainty about appropriate values for ϕ and ω due to the unknown mechanism that causes heterogeneity. To allow for this, we specify a prior distribution on a set of plausible values $\{(\phi^{(i)}, \omega^{(i)}) : i = 1, \ldots, M\}$. We give a specific choice of such prior in the applications below. If π_i denotes the prior weight on $(\phi^{(i)}, \omega^{(i)})$ then the resulting Bayes Factor against H_0 is the weighted

average of the individual BFs:

$$\mathrm{BF}^{\mathrm{ES}}_{\mathrm{all}} = \sum_{i=1}^{M} \pi_i \mathrm{BF}^{\mathrm{ES}}(\phi^{(i)}, \omega^{(i)}). \tag{2.14}$$

Calculating the Bayes Factors

Calculating $\mathrm{BF}^{\mathrm{ES}}(\phi, \omega)$ and $\mathrm{BF}^{\mathrm{EE}}(\psi, w)$ boils down to evaluating a complicated multi-dimensional integral. In appendix A, we show two different approximations to these integrals, both based on applying Laplace's method and both having error terms that decay inversely with the average sample size across subgroups. The first of these, which effectively follows methods from Butler and Wood (2002) for computing confluent hyper-geometric functions, is very accurate, even for small sample sizes. Indeed, for the special case of a single subgroup ($S = 1$), the approximation becomes *exact*, and for small numbers of subgroups we have checked numerically (appendix D) that it provides almost identical results to an alternative approach based on adaptive quadrature (which is practical only for small S). However, it requires a numerical optimization step and has a somewhat complex form, which although not a practical barrier to its use does hinder intuitive interpretation. In what follows we simply use $\widehat{\mathrm{BF}}^{\mathrm{ES}}$ to denote this approximation.

The second approximation is less accurate for small samples sizes, but converges asymptotically (with average sample size) to the correct answer. For the special case of $S = 1$ it yields an analogue of the approximate Bayes Factors from Wakefield (2009) and Johnson (2008), and in what follows we use $\mathrm{ABF}^{\mathrm{ES}}$ to denote this approximation under the ES model. The nice feature of $\mathrm{ABF}^{\mathrm{ES}}$ is that it has an intuitive analytic form with close connections to standard Frequentist test statistics for meta-analysis. Proposition 1 below gives this analytic form in detail.

Before stating proposition 1, we introduce some notation.

- *Association Testing in a Single Subgroup*

  First, let us consider analyzing a single subgroup, s. Let $\hat{\beta}_s$ and $\hat{\sigma}_s$ denote the least square estimates of β_s and σ_s from the linear regression model (2.1)

using only data from s. Then an estimate for the standardized effect b_s and its standard error $\delta_s := \text{se}(\hat{b}_s)$ can be obtained from

$$\hat{b}_s = \hat{\beta}_s / \hat{\sigma}_s, \tag{2.15}$$

$$\delta_s^2 = \frac{1}{\boldsymbol{g}_s' \boldsymbol{g}_s - n_s \bar{g}_s^2}. \tag{2.16}$$

The usual Frequentist statistic T_s for testing $b_s = 0$ can be represented as

$$T_s^2 = \frac{\hat{b}_s^2}{\text{se}(\hat{b}_s)^2} = \frac{\hat{\beta}_s^2}{\hat{\sigma}_s^2 \delta_s^2}. \tag{2.17}$$

(Note that T_s is also equal to $\hat{\beta}_s / \text{se}(\hat{\beta}_s)$, which is the t-statistic for testing $\beta_s = 0$ in Frequentist framework).

Both Wakefield (2009) and Johnson (2008) derive an approximate Bayes Factor for testing $b_s \sim \text{N}(0, \phi^2)$ vs. $b_s = 0$, which has the form

$$\text{ABF}^{\text{ES}}_{\text{single}}(T_s, \delta_s; \phi) = \sqrt{\frac{\delta_s^2}{\delta_s^2 + \phi^2}} \exp\left(\frac{T_s^2}{2} \frac{\phi^2}{\delta_s^2 + \phi^2} \right). \tag{2.18}$$

- *Testing Average Effect in a Random-effect Meta-analysis Model*

  Now consider the standard Frequentist test of $\bar{b} = 0$ in a random-effect meta-analysis of all subgroups, with $b_s \sim \text{N}(\bar{b}, \phi^2)$. If ϕ is considered known, the standard estimate for $\bar{b}$ and its standard error $\zeta := \text{se}(\hat{\bar{b}})$ are

$$\hat{\bar{b}} = \frac{\sum_s (\delta_s^2 + \phi^2)^{-1} \hat{b}_s}{\sum_s (\delta_s^2 + \phi^2)^{-1}}, \tag{2.19}$$

  and

$$\zeta^2 = \frac{1}{\sum_s (\delta_s^2 + \phi^2)^{-1}}. \tag{2.20}$$

  (Note that in this context, each $\hat{b}_s$ is regarded as an estimate of $\bar{b}$.)

The usual Frequentist statistic $\mathcal{T}_{\text{es}}^2$ for testing $\bar{b} = 0$ is

$$\mathcal{T}_{\text{es}}^2 = \frac{\hat{\bar{b}}^2}{\text{se}(\hat{\bar{b}})^2}. \tag{2.21}$$

We can "translate" this test statistic into the Bayes Factor by applying Johnson's recipe (Johnson (2005, 2008)) and the resulting approximate Bayes Factor for testing $\bar{b} \sim \text{N}(0, \omega^2)$ vs. $\bar{b} = 0$ is given by

$$\text{ABF}_{\text{single}}^{\text{ES}}(\mathcal{T}_{\text{es}}^2, \zeta; \omega) = \sqrt{\frac{\zeta^2}{\zeta^2 + \omega^2}} \exp\left(\frac{\mathcal{T}_{\text{es}}^2}{2} \frac{\omega^2}{\zeta^2 + \omega^2}\right). \tag{2.22}$$

We are now able to describe the analytic form of the overall approximate Bayes Factor $\text{ABF}^{\text{ES}}(\phi, \omega)$, as a simple product of the ABFs (2.18) and (2.22).

PROPOSITION 1. *Under the ES model,*

$$\text{ABF}^{\text{ES}}(\phi, \omega) = \text{ABF}_{\text{single}}^{\text{ES}}(\mathcal{T}_{\text{es}}{}^2, \zeta; \omega) \cdot \prod_s \text{ABF}_{\text{single}}^{\text{ES}}(T_s^2, \delta_s; \phi). \tag{2.23}$$

Furthermore, $\text{ABF}^{\text{ES}}(\phi, \omega)$ *converges to the true Bayes Factor as* $n_s \to \infty$ *for all subgroups* s*.*

Proof. appendix A.1. □

Proposition 1 breaks down the overall evidence for association into parts that are due to the evidence in each individual subgroup (the second term) and a part that reflects the consistency of the effects across subgroups (the first term). In particular, if all subgroups show effects in the same direction, then the first term will tend to be large ($\gg 1$) and provide a "boost" in the evidence for association compared with the situations when the effects across subgroups are in different directions.

A similar result holds for the EE model and is given in appendix A.2. The detailed computation of Bayes Factors involving CEFN priors is shown in appendix A.3. In appendix D, we show the evaluation of numerical accuracy of various Bayes Factor approximations described above.

Small Sample Size Corrections for the Approximate Bayes Factor

The accuracy of ABF^{ES} relies on the sample sizes in subgroups: when sample sizes are small in some subgroups, the approximation may become inaccurate. In particular, we consider the behavior of the approximate Bayes Factor when the null hypothesis is true. A valid Bayes Factor has the property that

$$\text{E}(\text{BF}|H_0) = 1, \tag{2.24}$$

where the expectation is taken with respect to the data distribution under the null model. Unfortunately, when sample sizes are small, (2.24) can be violated (as the expected value is strictly greater than 1) for some ABF^{ES} (appendix C), which leads to inaccurate approximation results (appendix D). Therefore, when applying (A.38) and (2.23) special care must be taken in small sample situations.

We now propose a simple correction procedure for small sample sizes, which ensures the resulting approximation satisfies property (2.24). Specifically, we modify (2.23) into the following form

$$\text{A}^*\text{BF}^{\text{ES}}(\phi, \omega) = \text{ABF}^{\text{ES}}_{\text{single}}(q({\mathcal{T}_{\text{es}}}^2), \zeta; \omega) \cdot \prod_s \text{ABF}^{\text{ES}}_{\text{single}}(q_s(T_s^2), \delta_s; \phi). \tag{2.25}$$

where the function q_s denote a one-to-one quantile transformation from a t-distribution with $n_s - 2$ degree of freedom to a standard normal distribution, and the function q is defined as

$$q(\mathcal{T}_{\text{es}})^2 = \frac{\hat{\tilde{b}}^2_{\text{cor}}}{\zeta^2}, \tag{2.26}$$

where

$$\hat{\tilde{b}}_{\text{cor}} = \frac{\sum_s (\delta_s^2 + \phi^2)^{-1} \delta_s q_s(T_s)}{\sum_s (\delta_s^2 + \phi^2)^{-1}}. \tag{2.27}$$

Note, the quantile transformation functions q_s and q converge to the identity mappings as $n_s \to \infty$ and the asymptotic property of (2.23) is preserved. Other details on approximation (2.25) are discussed in appendix C.

The correction is practically very effective: $\text{A}^*\text{BF}^{\text{ES}}$ yields satisfying accuracy when subgroup-level sample sizes decrease to 40 to 50. We demonstrate this correction

with the real data example in appendix D.

2.2.5 Properties of Bayes Factors

In this section, we discuss some interesting and important properties of the Bayes Factors described above.

Data Reduction in Computing Bayes Factors

All four Bayes Factors ($\mathrm{BF}^{\mathrm{ES}}, \mathrm{BF}^{\mathrm{EE}}, \mathrm{ABF}^{\mathrm{ES}}$ and $\mathrm{ABF}^{\mathrm{EE}}$) depend on the observed data in each subgroup only through a set of summary statistics, i.e., a 6-tuple (n_s, $\mathbf{1}'\boldsymbol{y}_s$, $\mathbf{1}'\boldsymbol{g}_s$, $\boldsymbol{y}_s'\boldsymbol{y}_s$, $\boldsymbol{g}_s'\boldsymbol{g}_s$, $\boldsymbol{y}_s'\boldsymbol{g}_s$). Therefore, to perform the proposed Bayesian method in a meta-analysis context, there is no need to collect full data set from each individual study. If the approximate Bayes Factors are sufficient, the summary statistics from each study are reduced to only $(\hat{b}_s, \mathrm{se}(\hat{b}_s))$ for the ES model and $(\hat{\beta}_s, \mathrm{se}(\hat{\beta}_s))$ for the EE model.

Induced Single Study Bayes Factors

For the ES model, in the special case of one subgroup ($S = 1$), both the actual Bayes Factor and our approximations to it (2.23) reduce to results from previous work. More specifically, $\widehat{\mathrm{BF}}^{\mathrm{ES}}$ becomes exact in this case, as the Bayes factor derived by Servin and Stephens (2007), whereas the approximation is the same as the ABF in Wakefield (2009) (see also Johnson (2005) and Johnson (2008)).

Non-informative Subgroup Data

If most of sample genotypes of the target SNP concentrate in only one of the three genotype categories in subgroup s, i.e. the sample variance of the genotype data tends to 0 (in our notation, $\delta_s^2 \to \infty$), the data from subgroup s contain little information on the correlation of the phenotype and the target SNP. Such a scenario could arise in cross-population genetic studies where allele frequencies of SNPs have large variation

in different populations. It is not uncommon to observe SNPs with modest minor allele frequencies in one population is monomorphic in another population.

It can be shown from the approximate Bayes Factor (2.23) or the expression (A.13) of the exact Bayes Factors (appendix A) that for both EE and ES models,

$$\lim_{\delta_s^2 \to \infty} \mathrm{BF}_{\text{include s}} = \mathrm{BF}_{\text{exclude s}}, \tag{2.28}$$

which indicates whether including data from subgroup s has little effect on the resulting Bayes Factor. This property shows that the proposed Bayesian procedure correctly characterizes the non-informativeness of the data. Although this property seems very intuitive and we might expect every statistical procedure to possess it, for some methods (e.g. Fisher's combined probability test), (2.28) does not hold.

Extreme Models and Corresponding Bayes Factors

The proposed hierarchical models are very flexible, covering a wide range of possible heterogeneity by setting different values for (ϕ, ω). The following two special cases correspond to the extremes of no heterogeneity and maximum heterogeneity:

1. *Fixed Effect Model.* A fixed effect model assumes genetic effects are homogeneous across subgroups. We obtain this extreme case by setting $\phi = 0$, which consequently forces all unobserved b_s identical to $\bar{b}$.

   The resulting approximate Bayes factor (2.23) simplifies to

$$\mathrm{ABF}_{\mathrm{fix}}^{\mathrm{ES}}(\omega) := \mathrm{ABF}^{\mathrm{ES}}(\phi = 0, \omega) = \sqrt{\frac{\zeta^2}{\zeta^2 + \omega^2}} \cdot \exp\left(\frac{\mathcal{T}_{\mathrm{es}}^2}{2} \frac{\omega^2}{\zeta^2 + \omega^2}\right), \tag{2.29}$$

where

$$\zeta = \frac{1}{\sum_s \delta_s^{-2}}, \tag{2.30}$$

$$\hat{\bar{b}} = \frac{\sum_s \delta_s^{-2} \frac{\hat{\beta}_s}{\hat{\sigma}_s}}{\sum_s \delta_s^{-2}}, \tag{2.31}$$

$$\mathcal{T}_{\text{es}}^2 = \frac{\hat{\bar{b}}^2}{\zeta^2}. \tag{2.32}$$

Note, a Frequentist test under the same fixed effect employs the same test statistic $\mathcal{T}_{\text{es}}$.

2. *Maximum Heterogeneity Model.* Consider a class of ES models with prior expected marginal effect sizes $(\phi^2 + \omega^2)$ set as constant. Then the model with $\omega = 0$ represents the maximum possible heterogeneity among all models in the class. Under this setting, the average effect $\bar{b}$ is strictly 0 and conditional on ϕ, b_s from different subgroups are independent.

   It can be shown from (A.13) and (A.28) that for both the EE and the ES models, the exact Bayes Factor under this particular setting BF_{maxH} is the product of the individual Bayes Factors, i.e.

$$\text{BF}_{\text{maxH}} = \prod_s \text{BF}_{\text{single},\, s}, \tag{2.33}$$

   where $\text{BF}_{\text{single},\, s}$ is the exact Bayes Factor calculated using data only from subgroup s. For the approximate Bayes Factor, this relationship also holds, i.e.

$$\text{ABF}_{\text{maxH}}^{\text{ES}}(\phi) := \text{ABF}^{\text{ES}}(\phi, \omega = 0) = \prod_s \sqrt{\frac{\delta_s^2}{\delta_s^2 + \phi^2}} \exp\left(\frac{T_s^2}{2} \frac{\phi^2}{\delta_s^2 + \phi^2}\right). \tag{2.34}$$

In the simplest case, we can use these two extreme models as proxies for low and high degrees of heterogeneities. Comparing BF_{maxH} and BF_{fix} provides us with a sense of the support from data for strong versus weak heterogeneity. As a demonstration, we simulate two sets of data for two different SNPs, both consisting of 5

subgroups. Figure 2.1 shows estimated effect sizes ($\hat{\beta}_s$) and their corresponding 95% confidence intervals estimated from the simulated data. SNP 1 is simulated to have consistent effects across subgroups. SNP 2 mimics an example of G×E interaction, where the direction of the effects are different in subgroups (this phenomena is known as "qualitative interaction"). For both SNPs, we compute $\text{ABF}^{\text{EE}}{}_{\text{fix}}$ and $\text{ABF}^{\text{EE}}{}_{\text{maxH}}$ with $w^2 + \psi^2 = 0.01$ for both SNPs. The resulting approximate Bayes Factors are shown in Table 2.1.

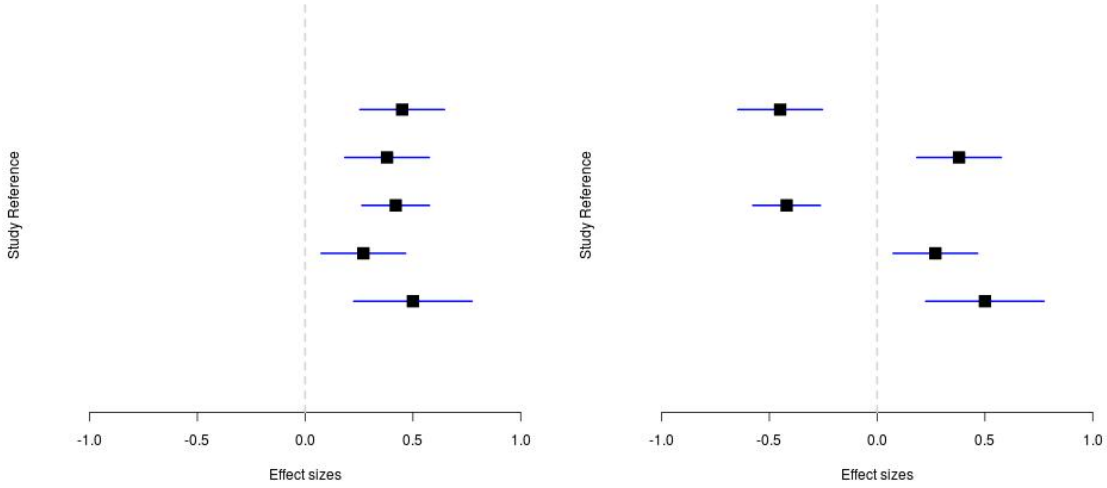

Figure 2.1: Forest plots of simulated data sets of two SNPs. SNP 1 and SNP 2 is simulated to mimic situations in meta-analysis and G×E interaction study respectively. For each subgroup data, estimated $\hat{\beta}_s$ and its 95% confidence interval is plotted.

Name	$\log_{10}(\text{ABF}^{\text{EE}}_{\text{fix}})$	$\log_{10}(\text{ABF}^{\text{EE}}_{\text{maxH}})$	$\log_{10}(\text{ABF}^{\text{EE}}_{\text{all}})$
SNP 1	14.07	8.40	13.77
SNP 2	−0.25	8.40	8.07

Table 2.1: Approximate Bayes Factors of the extreme models for simulated data set. $\text{ABF}^{\text{EE}}_{\text{all}}$ is computed by averaging the two extreme Bayes Factors.

First of all, in both cases, evidence against the null is very strong. This can be seen by computing $\text{ABF}^{\text{EE}}_{\text{all}}$ using two extreme models. In the case of SNP 1,

there is very strong evidence in favor of the fixed effect model suggesting effect sizes across subgroups are quite consistent. While for SNP 2, the Bayes Factor indicates the support for maximum heterogeneity model is overwhelming. Although these results are expected, the magnitude of the evidence favoring one model than the other summarized by Bayes Factors ($10^{5.67}$ in SNP 1, and $10^{8.15}$ in SNP 2) is striking.

2.2.6 Model for Case-Control Data

In situations when phenotypes are case/control status, we replace the linear model (2.1) for each subgroup by a logistic regression model: for individual i in subgroup s, the phenotype-genotype association is modeled by

$$\log \frac{\Pr(y_{si}=1|g_{si})}{\Pr(y_{si}=0|g_{si})} = \mu_s + \beta_s g_{si}. \tag{2.35}$$

Furthermore, we use the same form for the priors on μ_s, β_s and $\bar{\beta}$ as described in the EE model for quantitative traits.

Computing Bayes Factors for this model is challenging because the marginal likelihood is analytically intractable. To ease the computation, we approximate the subgroup-level log-likelihood function $l(\beta_s, \mu_s)$ given by (2.35) using an asymptotic expansion around its maximum likelihood estimates. The calculation then becomes straightforward and we show it in appendix B.

The resulting approximate Bayes Factor has the same form as in (2.23) and (A.38). Let $\hat{\beta}_s$ denote the MLE of β_s using the data from subgroup s only. For the alternative model specified by parameters (ψ, w),

$$\begin{aligned} \mathrm{ABF}^{\mathrm{CC}}(\psi, w) =& \sqrt{\frac{\xi^2}{\xi^2+w^2}} \exp\left(\frac{\mathcal{Z}_{\mathrm{cc}}^2}{2}\frac{w^2}{\xi^2+w^2}\right) \\ & \cdot \prod_s \left(\sqrt{\frac{\gamma_s^2}{\gamma_s^2+\psi^2}} \exp\left(\frac{Z_s^2}{2}\frac{\psi^2}{\gamma_s^2+\psi^2}\right)\right), \end{aligned} \tag{2.36}$$

where

$$\gamma_s^2 := \mathrm{se}(\hat{\beta}_s)^2, \tag{2.37}$$

$$Z_s^2 = \frac{\hat{\beta}_s^2}{\gamma_s^2}, \tag{2.38}$$

$$\hat{\bar{\beta}} = \frac{\sum_s (\gamma_s^2 + \psi^2)^{-1} \hat{\beta}_s}{\sum_s (\gamma_s^2 + \psi^2)^{-1}}, \tag{2.39}$$

$$\xi^2 := \mathrm{se}(\hat{\bar{\beta}})^2 = \frac{1}{\sum_s (\gamma_s^2 + \psi^2)^{-1}}, \tag{2.40}$$

$$\mathcal{Z}_{\mathrm{cc}}^2 = \frac{\hat{\bar{\beta}}^2}{\zeta^2}. \tag{2.41}$$

2.3 Data Application

2.3.1 Global Lipids Study

The global lipids study (Teslovich *et al.* (2010)) is a large scale meta-analysis of genome-wide genetic association studies of blood lipids phenotypes. In this study, more than 100,000 individuals of European ancestry were amassed through 46 separate studies (grouped into 25 studies in their final analysis). For each individual, quantitative phenotypes of total cholesterol (TC), low-density lipoprotein cholesterol (LDL-C), high-density lipoprotein cholesterol (HDL-C) and triglycerides (TG) were measured. The whole genome screening of genetic variants were performed and missing genotypes were imputed: in total, about 2.7 million common SNPs were included in the final association analysis. In each individual study, all four phenotypes were independently quantile normal transformed; single SNP association testings were performed for all SNPs and all phenotypes using the linear model (2.1) and the estimated effect sizes and their standard errors were reported. In the meta-analysis stage, they collected those summary-level information from each individual study and performed a version of the Frequentist fixed effect testing procedure, known as Stouffer's method (Willer *et al.* (2010)). In the end, they reported 95 significantly associated loci (the fixed effect p-value $< 10^{-8}$), with 59 showing genome-wide significant association with lipid traits for the first time.

We use this dataset to study the behaviors of proposed Bayesian methods under the settings of genetic meta-analysis. Because the data from each individual study are available only in forms of the summary statistics (the estimated effects and their standard errors), we choose to apply the EE models and compute the corresponding approximate Bayes Factors. For imputed genotypes, we follow Guan and Stephens (2008) to substitute imputed mean genotypes in linear model (2.1). To formalize their arguments, we also provide a justification in appendix E. We compare the results obtained from the fixed effect model, the EE model with CEFN priors and the maximum heterogeneity model using the LDL-C phenotype.

For all three different types of models, we assume a discrete uniform prior on the overall genetic effect size (i.e., $w^2 + \psi^2$ in the regular EE models; $(k^2+1)w^2$ in the EE model with CEFN priors) on $\mathrm{E}(\beta_s^2) = 0.1^2, 0.2^2, 0.4^2, 0.6^2$ and 0.8^2. For the fixed effect model, ψ is set to 0; for the maximum heterogeneity model, we set $w = 0$; for CEFN prior, we specify $k = 0.326$ which reflects a prior belief that the effect in a particular study having an opposite sign to the average effect is only 1 in 1,000.

We select the top 10,000 associated SNPs based on the fixed effect p-values reported in Teslovich *et al.* (2010) and compare their approximate Bayes Factors from the three different models. The comparisons between resulting approximate Bayes Factors are shown in Figure 2.2.

We note the resulting fixed effect Bayes Factors ($\mathrm{ABF}_{\mathrm{fix}}^{\mathrm{EE}}$) yield a very consistent ranking of the associated SNPs with the ranking based on the reported Frequentist fixed effect p-values. This is expected: both the approximate Bayes Factors and p-values of the fixed effect model essentially utilize the same test statistics as we show in (2.29).

As discussed in section 2.2.3, the fixed effect model may be considered too restrictive and certain level of heterogeneity in genetic effects across studies is generally expected in this meta-analysis context. We find the EE model with CEFN prior fits this scenario quite well. In general, the very top Bayes Factors from this model are quite consistent with the fixed effect model: for the top 1,500 LDL-C associated SNPs reported (values of $\log_{10}(\mathrm{ABF}_{\mathrm{fix}}^{\mathrm{EE}})$ range from 6.75 to 174.20), the rank correlation between $\mathrm{ABF}_{\mathrm{fix}}^{\mathrm{EE}}$ and $\mathrm{ABF}_{\mathrm{cefn}}^{\mathrm{EE}}$ reaches 0.99, which indicates the strongest association

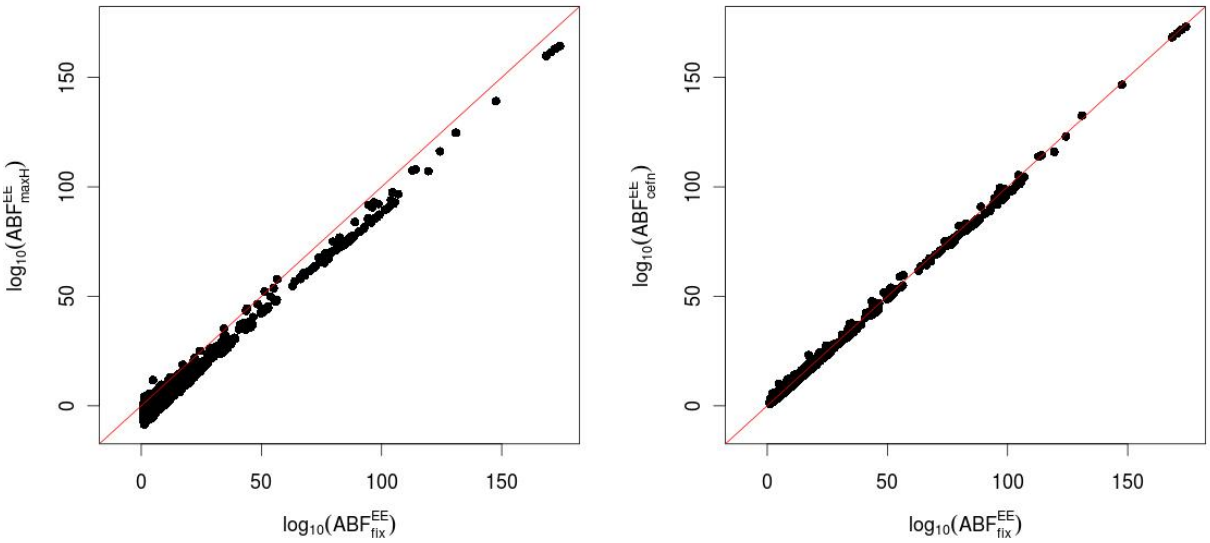

Figure 2.2: Comparison of approximate Bayes Factors using the fixed effect model, the maximum heterogeneity model and the EE model with CEFN prior in the meta-analysis of LDL-C phenotype. The top 10,000 associated SNPs based on the fixed effect p-values are plotted: on the left panel, $\log_{10}(\text{ABF}^{\text{EE}}_{\text{maxH}})$ vs. $\log_{10}(\text{ABF}^{\text{EE}}_{\text{fix}})$ is shown; on the right panel, the plot is shown for $\log_{10}(\text{ABF}^{\text{EE}}_{\text{CEFN}})$ vs. $\log_{10}(\text{ABF}^{\text{EE}}_{\text{fix}})$

signals in this dataset exhibit very low level of heterogeneity across studies. However, the agreement between the two Bayes Factors becomes weaker for SNPs with intermediate ranks: for SNPs ranked 5,000 to 8,000 in the original report, with values of $\log_{10}(\text{ABF}^{\text{EE}}_{\text{fix}})$ ranging from 1.39 to 2.45, the rank correlation between $\text{ABF}^{\text{EE}}_{\text{fix}}$ and $\text{ABF}^{\text{EE}}_{\text{cefn}}$ decreases to 0.66.

The maximum heterogeneity model conceptually is not ideal for identifying consistent association signals from meta-analysis. As shown in Figure 2.2, $\text{ABF}^{\text{EE}}_{\text{maxH}}$ constantly under-evaluate the evidence of association by ignoring the consistency of the directions of the signals across studies. Sometimes, this under-evaluation can be severe. We show such an example with SNP rs512535. This SNP is located in the promoter region of APOB gene which is a known to be associated with LDL-C phenotype. The estimated effects of this SNP from each individual study are mostly modest (shown in Figure 2.3), but the directions of the effects are quite consistent.

Using the maximum heterogeneity model, we obtain $\text{ABF}^{\text{EE}}_{\text{maxH}} = 0.73$. In comparison, $\text{ABF}^{\text{EE}}_{\text{fix}} = 10^{7.85}$ and $\text{ABF}^{\text{EE}}_{\text{cefn}} = 10^{6.84}$ and the reported Frequentist fixed effect test p-value is 1.132×10^{-9}.

For all three models that we have considered, with such a large sample size in the meta-analysis, it seems none of them (even the maximum heterogeneity model) misses extremely strong association signals as we show in Figure 2.2, the main differences among the models mainly are reflected in the rankings of modest to relatively strong signals. Overall, the maximum heterogeneity model lacks power by ignoring the consistency of the effect direction; the fixed effect model makes a too strong assumption of no heterogeneity and may also lose power when the true effects indeed have some levels of heterogeneity; the EE model with CEFN prior seems the most appropriate in this context by allowing but restricting possible heterogeneities of effect sizes.

In addition, we find that comparing Bayes Factors from the fixed effect model and the maximum heterogeneity model is practically useful for quality control purpose. Particularly, we look for SNPs whose $\text{ABF}^{\text{EE}}_{\text{maxH}} \gg \text{ABF}^{\text{EE}}_{\text{fix}}$. These are typically the results that association signals are driven by a relatively small number of studies (in extreme cases, the signal can be driven by a single study), but the genetic effects lack of consistency across all studies. In the global lipid study, we encounter such an example in SNP rs11984900 with phenotype HDL-C, in which $\text{ABF}^{\text{EE}}_{\text{maxH}} = 10^{16.79}$ and $\text{ABF}^{\text{EE}}_{\text{fix}} = 0.11$. It turns out this SNP shows a strikingly significant association only in one of the participating studies with p-value $= 7.24 \times 10^{-35}$, but in the remaining 24 studies the minimum p-value only reaches 0.086. In meta-analysis context, this phenomenon is closely related to "winner's curse", i.e. an significant association in a particular study fails to be replicated by subsequent studies. Most likely, the initial finding of significant association is attributed to some artifacts, e.g. experimental errors, population stratifications, in particular studies. In practice, the researchers should follow up on these identified SNPs to ensure the quality of the meta-analysis.

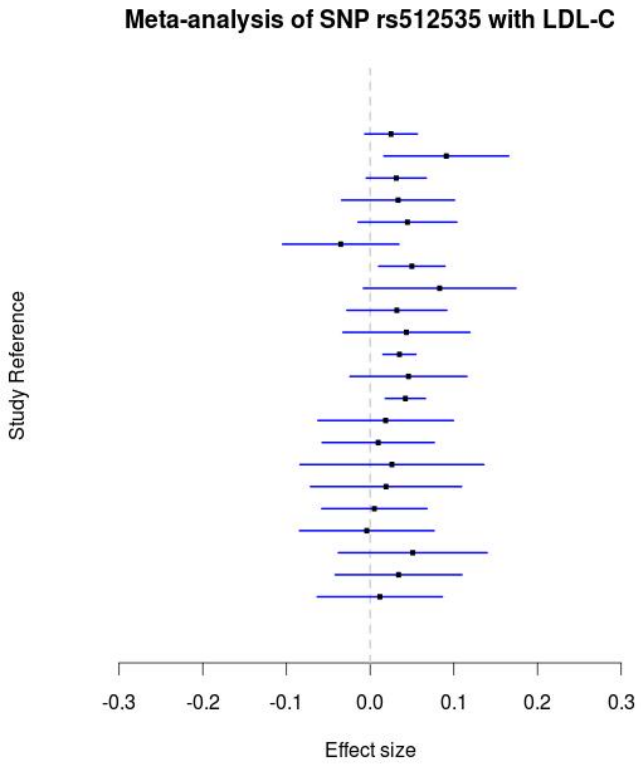

Figure 2.3: The genetic effect of SNP rs512535 with LDL-C estimated from individual studies. The point estimates and their corresponding 95% confidence intervals are shown in the forest plot. Most effects are modest but the direction of the effects are consistently positive. For this SNP, $\mathrm{ABF}^{\mathrm{EE}}_{\mathrm{maxH}} = 0.73$, $\mathrm{ABF}^{\mathrm{EE}}_{\mathrm{fix}} = 10^{7.85}$ and $\mathrm{ABF}^{\mathrm{EE}}_{\mathrm{cefn}} = 10^{6.84}$.

2.3.2 deCODE Recombination Study

The deCODE recombination study (Kong *et al.* (2008)) is designed to find genetic variants that explain genome-wide recombination rate variation. The study genotyped 1,887 males and 1,702 females from the Icelandic population and performed a genome-wide scan searching for association signals of SNP genotypes and the phenotype of recombination rate estimates.

Prior to this study, it was already commonly known that male and female recombination maps are quite different at genome-wide scales; the researchers therefore analyzed the data separately for males and females. They estimated the genetic effect sizes on the recombination phenotype assuming an additive model (2.1). For recom-

bination rate in males, they found three highly correlated SNPs in a small region on chromosome 4p16.3 show strong association signals. Interestingly, when compared with results in females, each of these three SNPs still shows strong association; however, the effect size points to opposite direction (i.e. the allele associated with low recombination rate in males is associated with high recombination rate in females).

We perform the Bayesian analysis on the three reported SNPs. We obtain the summary-level statistics of genome-wide scan result from Table 1 of Kong *et al.* (2008). In particular, we use their point estimates of effect sizes $\hat{\beta}_{\text{male}}$ and $\hat{\beta}_{\text{female}}$ and compute $\text{se}(\hat{\beta}_{\text{male}})$ and $\text{se}(\hat{\beta}_{\text{male}})$ from corresponding reported p-values.

We apply EE model by treating males and females as two subgroups and consider 4 levels of expected marginal overall effect sizes with $\sqrt{\psi^2 + w^2} = 5, 10, 20, 40$ (in scale of centi-Morgan) and 5 levels of heterogeneity levels with $\psi^2/w^2 = 0, 0.5, 1, 2, \infty$. In total, we obtain a grid of 4×5 different (ψ, w) combinations and we treat every grid value as *a priori* equally likely when computing $\text{ABF}^{\text{EE}}_{\text{all}}$.

The resulting Bayes Factors are shown in Table 2.2. The overall evidence against the global null is overwhelmingly strong, and there is little doubt that we should reject the global null. However, if we only concentrate on the fixed effect models or models allowing small degrees of heterogeneities, the evidence is *much* weaker.

	Male	Female	Bayes Factors		
SNP	Effect (p-value)	Effect (p-value)	$\text{ABF}^{\text{EE}}_{\text{fix}}$	$\text{ABF}^{\text{EE}}_{\text{maxH}}$	$\text{ABF}^{\text{EE}}_{\text{all}}$
rs3796619	-67.9 (1.1×10^{-14})	67.6 (7.9×10^{-6})	$10^{3.07}$	$10^{14.44}$	$10^{13.91}$
rs1670533	-66.1 (1.8×10^{-11})	92.8 (4.1×10^{-8})	$10^{1.10}$	$10^{13.16}$	$10^{12.58}$
rs2045065	-66.2 (1.6×10^{-11})	92.2 (6.0×10^{-8})	$10^{1.18}$	$10^{13.07}$	$10^{12.49}$

Table 2.2: Bayesian meta-analysis result of genetic association of remobination rate. The SNPs and their estimated effect sizes and p-values are directly taken from Kong *et al.* (2008) Table 1. We compute approximate Bayes Factor assuming EE model using only those reported summary statistics.

Although the p-values of the three SNPs indicate the genetic effects in males and females are separately both significant, it does not directly assess the magnitude of the subgroup interaction. In comparison, within our proposed Bayesian framework, by

comparing the fixed effect model and the maximum heterogeneity model, the quantity $\mathrm{ABF}^{\mathrm{EE}}_{\mathrm{maxH}}/\mathrm{ABF}^{\mathrm{EE}}_{\mathrm{fix}}$ provides a direct measure of the strength of genetic interaction in males and females.

We can further quantify the interaction by explicitly considering the direction of the effect size (with respect to the pre-defined allele) in the model. To do so, we modify the EE model with CEFN prior in the following way: for average effect $\bar{\beta}$ in positive direction, we use the prior

$$\begin{aligned} \beta_s &\sim \mathrm{N}(\bar{\beta}, k^2\bar{\beta}^2), \\ \bar{\beta} &\sim \mathrm{HN}(0, w^2), \end{aligned} \tag{2.42}$$

where HN stands for half-normal distribution. Similarly, for negative $\bar{\beta}$, we use

$$\begin{aligned} \beta_s &\sim \mathrm{N}(\bar{\beta}, k^2\bar{\beta}^2), \\ -\bar{\beta} &\sim \mathrm{HN}(0, w^2). \end{aligned} \tag{2.43}$$

This modified model enables us to specify the directions of genetic effects in the alternative models. For each subgroup, there are three prior possibilities for the underlying average genetic effect of any target SNP: no effect, positive effect or negative effect. For the male and the female subgroups in the deCODE data, we enumerate all 3^2 possible configurations and evaluate the support from the data by computing the corresponding Bayes Factors. We apply the above modified model to compute the Bayes Factors of all 9 configurations for SNP rs3796619. In particular, we set $k = 0.326$ and keep the grid of overall marginal prior genetic effects the same as in the previous exploratory analysis. In addition to the Bayes Factors, we also assume a prior weighting for the 9 configurations as $10^6 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1$, i.e. the null model is assigned most of the prior mass, while each non-null alternative is considered equally likely *a priori*.

The resulting approximate Bayes Factors and posterior probabilities for all configurations are shown in Table 2.3. The strongest support from data based on the model is given to the configuration that asserts the genetic effect is negative in males and positive in females, and this particular configuration has the posterior probability

0.977.

Configuration (male:female)	$\log_{10}(\text{ABF}^{\text{EE}}_{\text{cefn}*})$	Posterior Probability
0 : 0	0.000	3.4×10^{-9}
0 : +	3.067	3.9×10^{-12}
0 : −	0.446	9.4×10^{-15}
+ : 0	8.537	1.2×10^{-6}
+ : +	12.357	0.007
+ : −	8.983	3.2×10^{-6}
− : 0	11.394	8.4×10^{-4}
− : +	**14.461**	**0.977**
− : −	12.635	0.015

Table 2.3: Quantify the subgroup interaction for SNP rs3796619. The log 10 of approximate Bayes Factors based on modified EE model with modified CEFN priors ($\log_{10}(\text{ABF}^{\text{EE}}_{\text{cefn}*})$) for all possible configurations (the left column) are shown. The strongest Bayes Factors is obtained from the model that asserts the genetic effect is negative in males and positive in females (highlighted).

2.3.3 Population eQTL Study

In this section, we apply the proposed Bayesian methods in mapping expression quantitative trait loci (eQTL) using multi-population data. An eQTL is a genetic variant (here we only focus on SNPs) that is associated with gene expression phenotype. The dataset we analyzed consists of gene expression measurements from lymphoblastoid cell lines of 141 unrelated individuals from Hapmap project The International HapMap Consortium (2005)). These individuals were sampled from three major population groups (41 Europeans (CEU), 59 Asians (ASN) and 41 Africans (YRI)) and were fully sequenced in the pilot project of the 1000 Genomes project (Durbin *et al.* (2010)). The gene expression levels, measured using the Illumina Sentrix Human-6 Expression BeadChip, came from Stranger *et al.* (2007). We focus on the 8,427 distinct autosomal genes that were confirmed to be expressed in the same African samples by an independent RNA-seq experiment (Pickrell *et al.* (2010)). The SNP genotype data were obtained from final release (March, 2010) of the pilot SNP calls

from 1000 genome project (Note, there is no additional allele frequency filtering applied to the SNPs). In total, 14.4 million SNPs are considered in our analysis. In addition to the original normalization procedure for gene expression measurement described in Stranger *et al.* (2007), we perform quantile normal transformations for each selected gene within each population group separately to adjust for population effects.

Our first goal is to identify eQTLs in the *cis* regions of selected genes: for a given gene, we examine the associations between the expression levels and genotypes of SNPs within the region enclosed by 500kb upstream of the transcription start site and 500 kb downstream of the transcription end site. We group the samples by their population of origin to form three subgroups and apply the proposed Bayesian methods.

We use the ES model with a grid of (ϕ, ω) values. Specifically, we consider five levels of $\sqrt{\phi^2 + \omega^2}$ values: 0.1, 0.2, 0.4, 0.8, 1.6, and seven degrees of heterogeneities characterized by ϕ^2/ω^2 values: $0, 1/4, 1/2, 1, 2, 4, \infty$. Further, we assign these 35 grid values equal prior weight. For each gene, we compute $\widehat{\mathrm{BF}}^{\mathrm{ES}}_{\mathrm{meta}}$ for each *cis* SNP and report the highest ranked SNP. The histogram of $\log_{10}(\widehat{\mathrm{BF}}^{\mathrm{ES}}_{\mathrm{meta}})$ values of these top SNPs are shown in Figure 2.4. Among the top SNPs, there are about 14% having Bayes Factors greater than 10^5 and about 18% having Bayes Factors greater than 10^4, suggesting a significant portion of genes having *cis*-eQTLs with strong effects.

We then proceed to examine the heterogeneities of potential eQTL effects in populations. Specifically, we compare the maximum heterogeneity model and the fixed effect model for each of the top ranked SNPs selected from the previous step. The results are shown in Figure 2.5. Overall, for most SNPs considered, there is no strong evidence for large degree of heterogeneity in eQTL effects. Nevertheless, there exist a few cases where the data suggest the effects of eQTLs are indeed strongly heterogeneous in different populations. We show four of such examples in Table 2.4 and Figure 2.6.

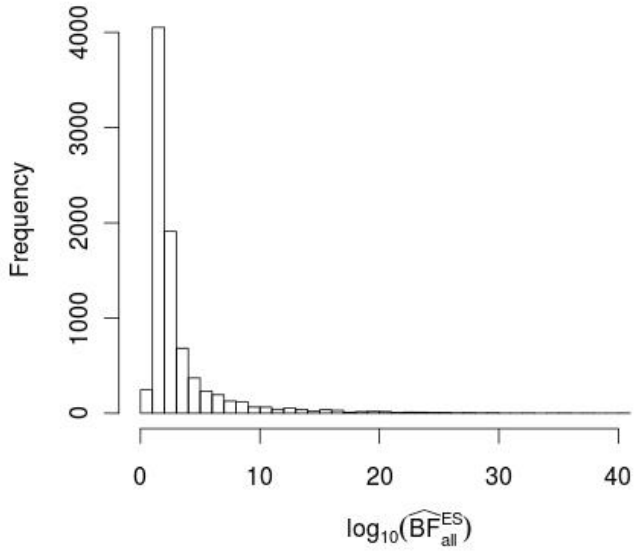

Figure 2.4: Histogram of $\log_{10}(\widehat{\mathrm{BF}}^{\mathrm{ES}}_{\mathrm{meta}})$ values of top ranked *cis*-SNPs for each of the 8,427 genes examined.

SNP	Gene	$\log_{10}(\widehat{\mathrm{BF}}^{\mathrm{ES}}_{\mathrm{fix}})$	$\log_{10}(\widehat{\mathrm{BF}}^{\mathrm{ES}}_{\mathrm{maxH}})$	$\log_{10}(\widehat{\mathrm{BF}}^{\mathrm{ES}}_{\mathrm{all}})$
rs9595893	RP11-298P3.4	4.54	19.25	19.02
rs380359	PLA2G4C	4.56	11.63	11.53
rs3180068	PAQR8	6.97	12.03	11.95
rs11070253	BUB1B	0.33	7.08	6.81

Table 2.4: Examples of eQTL SNPs showing strong heterogeneity of genetic effects in different populations.

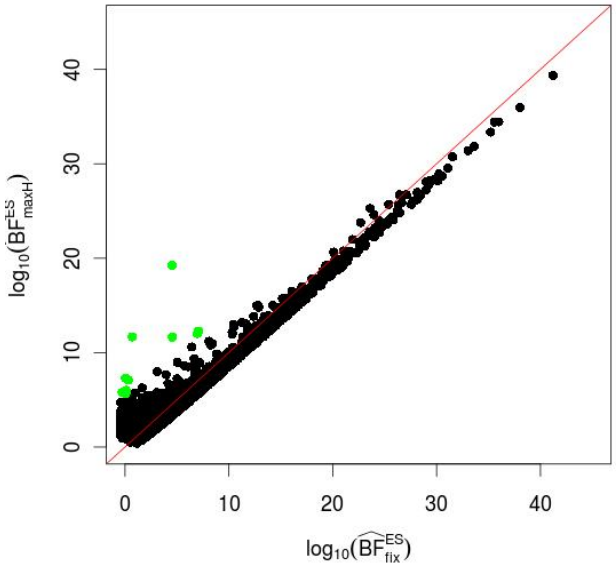

Figure 2.5: Comparisons of maximum heterogeneity models and fixed effect models for all top ranked SNPs. 10 out of all 8,427 SNPs with $\widehat{\mathrm{BF}}^{\mathrm{ES}}_{\mathrm{maxH}}/\widehat{\mathrm{BF}}^{\mathrm{ES}}_{\mathrm{fix}} > 10^5$ and $\widehat{\mathrm{BF}}^{\mathrm{ES}}_{\mathrm{meta}} > 10^5$ are highlighted in green. The effects of these SNPs appear highly heterogeneous in different populations

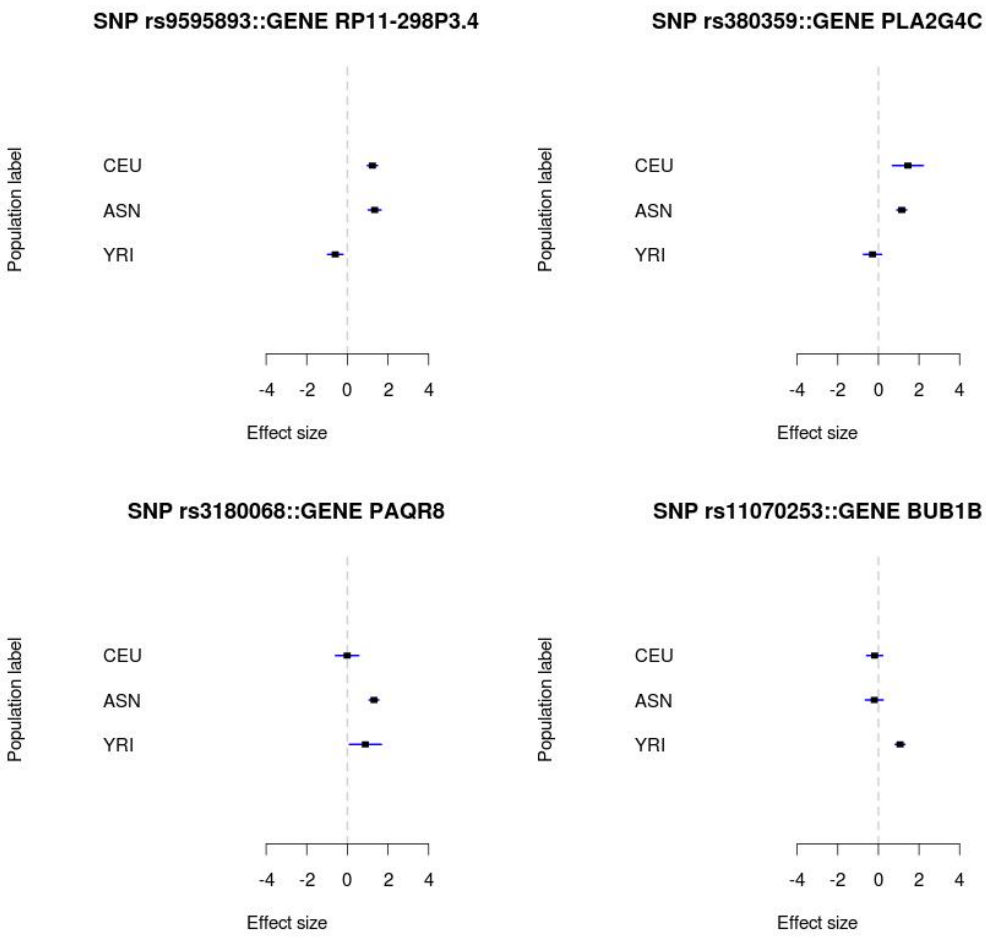

Figure 2.6: Examples of eQTL SNPs appearing to show strong heterogeneity of genetic effects in different populations. In each panel, the forest plot of a gene-SNP combination is shown: the estimated effect and its 95% confidence interval are plotted separately for each population.

Investigating Potential Population-specific eQTLs

The examples shown above (Figure 2.6 and Table 2.4) prompt us to further modify our proposed models and test scientific hypothesis attempting to explain the observed heterogeneities. One such hypothesis states that some eQTLs may behave in a population specific manner. For example, an active gene regulatory element in one population may become inactive in another population. As a result, genetic variants within or close by the regulatory element are no longer associated with expressions of the target, which could be possible as a consequence of natural selection (Kudaravalli *et al.* (2009)). Under this theory, the activities of the eQTL related regulatory elements in different populations may *explain* the observed heterogeneity.

To illustrate this, we modify models described above to explicitly assess the hypothesis that an eQTL is active in only some of the populations. Further, given an eQTL is active in some populations, we allow the effects to vary but assume the direction of the effect is always consistent.

Let an indicator denote a non-zero genetic effect of an eQTL (i.e. an eQTL being active) in a particular population. For the three populations in our data, we can enumerate and represent all 2^3 possible configurations of eQTL activity for a SNP by combining three indicators. A particular configuration denoted by $C = (110)$ indicates that genetic effects of the potential eQTL are non-zero in the first two populations and exactly 0 in the third population. To evaluate a particular configuration, we compute the Bayes Factor by contrasting marginal likelihood of the configuration of interest to the null model, i.e. $C = (000)$. For example, for $C = (110)$,

$$\begin{aligned} \mathrm{BF}_{C=(110)} &= \frac{P\left(\boldsymbol{y}_1, \boldsymbol{y}_2, \boldsymbol{y}_3 | \boldsymbol{g}_1, \boldsymbol{g}_2, \boldsymbol{g}_3, C = (110)\right)}{P\left(\boldsymbol{y}_1, \boldsymbol{y}_2, \boldsymbol{y}_3 | \boldsymbol{g}_1, \boldsymbol{g}_2, \boldsymbol{g}_3, C = (000)\right)} \\ &= \frac{P\left(\boldsymbol{y}_1, \boldsymbol{y}_2 | \boldsymbol{g}_1, \boldsymbol{g}_2\right)}{P(\boldsymbol{y}_1, \boldsymbol{y}_2 | H_0)}. \end{aligned} \tag{2.44}$$

(Note, The simplification is due to the assumption that the vectors of residual errors in (2.1) are independent across populations.) This shows that the Bayes Factor above depends only on the data where eQTL is assumed active. Here, we use a ES model with CENF prior and set $k = 0.314$ (Recall, it reflects a prior belief that

the probability of a particular effect has an opposite sign to the average effect is 0.001) to jointly model effects of a potential eQTL in active populations. For gene BUB1B and SNP rs11070253, we evaluate all eight activity configurations and show the resulting Bayes Factors in Table 2.5. Clearly, the data seems show strong support to the configuration where rs11070253 is associated with gene expression levels of BUB1B gene only in the Yoruban population.

CEU	ASN	YRI	$\log_{10}$ BF
0	0	0	0.000
1	0	0	−0.283
0	1	0	−0.285
0	0	1	**8.053**
1	1	0	−0.279
1	0	1	5.799
0	1	1	5.880
1	1	1	3.817

Table 2.5: Evaluation of population specificity for SNP rs11070253 and gene BUB1B. The log 10 Bayes Factors for all possible activity configurations (the left column) are shown.

2.4 Discussion

In this chapter, we have introduced a set of novel statistical methods to analyze potential-heterogeneous genetic association data in a Bayesian framework. With the proposed Bayesian models and easy-to-compute Bayes Factors, we are able to deal with potentially heterogeneous genetic association data in a systematic way. Through demonstrations and examples, we have shown the proposed Bayesian methods enable us to

1. identify interesting genetic variants that are associated with phenotype of interest in some subgroups.

2. investigate association signals that exhibit strong heterogeneities in effects and identify potential gene-by-environment interactions.

3. follow up on strongly heterogeneous genetic association signals and explain the heterogeneity by further explicit modeling approach.

Within a single, unified statistical framework, we are able to bridge exploratory analysis and detailed investigation.

The hierarchical models we proposed are quite natural and have been widely employed in the context of meta-analysis. They are very similar to the mixed effect meta-analysis models in Frequentist approaches for studying of quantitative phenotypes, where the subgroup-specific intercept terms μ_s in (2.1) are regarded as fixed effect terms and genetic effect β_s (or b_s) are regarded as random effect terms.

In study of G×E interaction, the most frequently used models for quantitative phenotypes are different from what we proposed. The typical models used in this context is a linear model with marginal effect of subgroups and an gene-subgroup interaction term included, i.e.

$$y_i = \mu_i + \beta_e s_i + \beta_g g_i + \beta_{[g:e]} s_i g_i + e_i, \quad e_i \sim \mathrm{N}(0, \sigma^2), \tag{2.45}$$

where s_i is a dummy variable denoting the subgroup membership of individual i, and $\beta_{[g:e]}$ is the coefficient of the subgroup-genotype interaction term and often of interest. This model, in some level, is quite similar to 2.1). By re-arranging and grouping the terms, the linear model can be written as

$$y_i = (\mu_i + \beta_e s_i) + (\beta_g + \beta_{[g:e]} s_i) g_i + e_i, \quad e_i \sim \mathrm{N}(0, \sigma^2). \tag{2.46}$$

Essentially, each subgroup is described with its own intercept, $\mu_i + \beta_s s_i$, and its own genetic effect, $\beta_e + \beta_{[g:e]} s_i$. (Note, if a marginal effect of subgroup is not included, the model is making a much stronger assumption on equal intercepts for different subgroups, which can be dangerous in practice and may lead to Simpson's paradox (Bravata and Olkin (2001))). Nevertheless, the interaction model still makes stronger assumption by assuming error variance across subgroups are the same. In comparison, meta-analysis models allow this quantity to differ in subgroups, which is more robust in practice. This robust assumption is highly desirable in genetic association context,

because, most likely, neither model (2.1) nor model (2.45) captures all factors affecting the phenotype of interest, and confounding factors almost certainly exist. In model (2.1), the effect of the unaccounted confounding factors are "absorbed" by both the intercept and error variance terms, whereas the interaction model, lacking flexible error variance terms, does not possess this property.

Our main contribution to this topic is providing a comprehensive Bayesian framework to deal with potentially heterogeneous genetic data in a rather general and broad context. Our first goal is always to find genetic variants that are associated with phenotype of interest in some subgroup. We achieve this by testing a global null hypothesis. Similar arguments have been by advocated by various authors (Stephens and Balding (2009), Lebrec *et al.* (2010)) in both Bayesian and Frequentist context. Given a genetic variant shows signs of association, we then proceed to investigate the details of potential heterogeneity of effects in subgroups by methods of Bayesian model comparison. Further, this framework also provide us flexibilities to construct candidate models to explain observed heterogeneity, which is the ultimate goal of genetic association studies in this circumstance.

The building block of our Bayesian framework is the use of Bayes Factors. We have developed different computationally efficient approximations to obtain numerical results, which is critical for handling of large scale genetic association data. In addition, we have shown the intrinsic connection between the Bayes Factors and Frequentist test statistics in this context through proposition 1.

The three data examples we show are designed to be representative of a wide range of genetic applications involving potentially-heterogeneous genetic association data. The global lipids study is a typical modern day, large-scale genetic association meta-analysis; the deCODE recombination study is designed to identify gene-environment interactions and finally the eQTL study is an increasingly popular way to understand both how genetic variants affect gene regulations and the specificity of the gene regulation processes. We show how our unified Bayesian strategy can be applied to all three seemingly distinct applications.

In the future work, we hope to extend our methods to consider multiple SNPs simultaneously, which is a more powerful way to detect genetic association. Simi-

lar Bayesian approach has been proposed in a single group framework (Guan and Stephens (2011)), but not in the setting of multiple subgroups. Another direction to generalize our method is to allow correlation among phenotypes even under the null. Such data are typically generated in the genetic experiments where phenotypes are measured using the same set of samples but under different environmental conditions. A motivating example is the study of eQTLs from multiple tissue-types. Most commonly, this type of the experiment takes gene expression measurement from different tissues of the same set of individuals. We might suspect that expression levels are correlated within the same individual even they do not share the same genetic association.

2.5 Acknowledgements

We thank Yongtao Guan, Michael Stein and Peter McCullagh for the helpful discussions.

CHAPTER 3
A HIERARCHICAL MODEL APPROACH FOR MAPPING TISSUE-SPECIFIC EQTLS

3.1 Introduction

With the success of many genome-wide association studies, researchers have now identified thousands of genetic variants that link to complex diseases. Based on these findings, a new wave of research has started focusing on understanding the underlying gene regulation processes and the impact of genetic variants in these processes. One type of such study aims to identify genetic variants that are associated with variations in gene expression, i.e. expression Quantitative Trait Loci (eQTL) analysis. eQTL analysis enables the inspection of the most immediate consequences of heritable genetic variants, namely, their effects on gene expression through transcription and post-transcription controls.

We first illustrates the role of eQTL studies in understanding gene regulation by the following simple example. There are short DNA sequences known as enhancers present in human genome. Typically, enhancers are located outside genes, and they are not directly involved in transcribing DNA into RNA. However, during the gene expression process, some particular proteins, known as transcription factors, can bind with enhancer sequences and as a result, the bindings lead to greatly increased expression levels of target genes. It has been widely known that the binding affinity of transcription factors is sensitive to the DNA sequences in enhancers; therefore a point mutation (e.g. a SNP) in an enhancer region has some effect on expression level of the target gene by affecting the binding affinity of the transcription factors. Conversely, if we observe that a SNP is associated with the gene expression level of a target gene, it is then reasonable to hypothesize that the particular SNP is a part of, or in LD with, an active gene regulatory element (e.g. enhancer).

In this chapter, we focus on developing statistical methods for analyzing eQTL data across different cell (tissue) types with the motivation of understanding cell-type specific gene regulation processes. As we have known, all different types of cells in

a same human body carry almost identical DNA sequences, yet their appearance, behaviors and functionality differ greatly, mainly due to differential gene regulations in different cell environments. Going back to the enhancer example, if in some types of cells, the required transcription factors are not manufactured or some strong prohibitive mechanism prevents the binding between the transcription factor and the enhancer sequence, the potential enhancers then become deactivated. Consequently, for these cell types, genetic variants in or near the potential enhancer regions become unassociated with the expression levels of the target gene. Hence, by identifying inconsistent genetic association patterns across different cell types from eQTL data, we may be able to identify mechanisms of differential gene regulation.

To investigate differential gene regulation by mapping eQTLs across tissues, for each individual gene, we are interested in identifying potential eQTLs that are either shared among tissues or behave in a tissue specific manner. By treating tissues as different subgroups, we can directly apply the methods described in the previous chapter. More interestingly, since we have simultaneous measurements of expressions from a large number of genes, we are able to pool information from multiple genes and

- Make statistical statement about the scope of tissue specificity of eQTLs (e.g. what percentage of eQTLs behave in a tissue specific manner?) .
- Identify and measure the strength of biological features linked to tissue specificity of eQTLs

In this chapter, we propose statistical methods for solving these problems.

Some publications address these issues problem by *ad hoc* methods, for example, they first call eQTLs independently in each tissue and then count numbers of called eQTLs across tissues (Dimas *et al.* (2009), Nica *et al.* (2011)). Such method is unsatisfying in two aspects: firstly, we typically do not have the same power to map eQTLs in different tissues and it is difficult to compare eQTLs that are called independently in different tissues; secondly, by simply counting the numbers of called eQTLs, the uncertainty of the calls is ignored and as a consequence, it is also difficult to attach a confidence interval to the inferred tissue specificity percentages.

The method we proposed overcomes both of these shortcomings of the *ad hoc* methods: we build our method on the groundwork from the previous chapter, which addresses the potential unequal power issue; further, we perform formal statistical inference which naturally incorporates interval estimation. In the following sections, we first introduce our notation, assumptions and models, then we discuss different methods for fitting the proposed model, finally we demonstrate our method by a real data application.

3.2 A Hierarchical Mixture Model

3.2.1 Assumptions and Notations

Our model is extended from the hierarchical model proposed by Veyrieras *et al.* (2008). To reduce computational burden, we only consider potential eQTLs in the *cis*-regulatory region of given genes. Same as in the previous chapter, we define the *cis*-region of a gene as the continuous genomic region that starts at 1Mb upstream of its transcription start site and ends at 1Mb downstream of its transcription end site. Further, we assume that all genes considered fall into two mutually exclusive classes: a gene either has no eQTL in its *cis*-region in any of the tissue type, or it can have exactly one eQTL in some tissues. Note, this assumption is mainly for reducing the computation complexity, and relaxing this assumption is an important area for our future work.

We consider a total number of g genes and their expression measurements are obtained from t tissues. For gene k, we denote the number of SNPs in its *cis*-region as m_k. The expression data $\boldsymbol{Y}_k$ is represented by an $n \times t$ matrix, with each column denoting the normalized expression levels for n individual samples in a particular tissue. Finally, we use $\boldsymbol{X}_k$ denote the observed relevant genotype data of all samples for gene k.

3.2.2 Basic Version of Hierarchical Mixture Model

For gene k, we use a latent binary indicator z_k to denote if there is any eQTL in its *cis*-region for any tissue type. We denote the prior probability that any gene has no eQTL in any tissue type by π_0, i.e.

$$\Pr(z_k = 1) = 1 - \pi_0. \tag{3.1}$$

We use a latent random indicator m_k-vector $\boldsymbol{s}_k$ to denote which SNP in the *cis*-region, conditional on $z_k = 1$, is the actual eQTL and let s_{kp} denote the p-th entry of $\boldsymbol{s}_k$. Our "one *cis* eQTL per gene" assumption restricts $\boldsymbol{s}_k$ can have at most one entry equaling 1 (with the remaining entries being 0). By this definition,

$$\Pr(\boldsymbol{s}_k = \mathbf{0} | z_k = 0) = 1, \tag{3.2}$$

and

$$\Pr(s_{kp} = 1 | z_k = 1) = \gamma_{kp}. \tag{3.3}$$

For simplicity, we further assume, *a priori*, every SNP is equally likely to be the true eQTL, i.e.

$$\gamma_{kp} = \frac{1}{m_k}. \tag{3.4}$$

This assumption will be relaxed later in this section.

To model the tissue specificity of eQTLs, we allow the association between gene expression and SNPs to behave in a tissue specific way: if a SNP is the assumed eQTL for a gene, we only require such association preserved in some (at least one), but not necessarily all, of the examined tissue types. Using a binary indicator to denote whether there is an association in a particular tissue, we can enumerate all possible mutually exclusive configurations. For example, for 2 tissue types, the set of all possible non-null configurations of a given SNP is $\{(10), (01), (11)\}$, where (11) indicates that the association is consistent in both tissues. In general, for t tissue types, the total number of possible activity configurations is $2^t - 1$. For gene k and SNP p, we index all configurations and use a $(2^t - 1)$-dimension latent indicator vector

$\boldsymbol{c}_{kp}$ to denote the actual configuration. In case the SNP is not the eQTL,

$$\Pr(\boldsymbol{c}_{kp} = \mathbf{0} | s_{kp} = 0) = 1. \tag{3.5}$$

Conditional on SNP p is the eQTL in gene k, we assume the jth configuration is active with prior probability

$$\Pr(c_{kpj} = 1 | s_{kp} = 1) = \eta_j. \tag{3.6}$$

Joining column vector $\boldsymbol{c}_{kp}$ for all p SNPs, we obtain a latent $(2^t - 1) \times p$ random matrix C_k.

Finally, for a given eQTL, across tissues where the association is preserved, we model the heterogeneous genetic effects use the Bayesian meta-analysis (ES) model introduced in the previous chapter. The ES model requires specification of a set of parameters (ϕ^2, ω^2) to reflect our prior belief on heterogeneity and average genetic effect size respectively. Instead of using a single set of parameter, we pre-define a grid of l possible parameter values and use latent l-vector $\boldsymbol{w}_{kp}$ indicate which particular parameter set is in use for the pair of gene k and SNP p. The m-th entry of the indicator is denoted by w_{kpm}, and we assume prior probability

$$\Pr(\boldsymbol{w}_{kp} = \mathbf{0} | s_{kp} = 0) = 1, \tag{3.7}$$

and

$$\Pr(w_{kpm} = 1 | s_{kp} = 1) = \lambda_m. \tag{3.8}$$

Joining column vector $\boldsymbol{w}_{kp}$ for all p SNPs, we obtain a latent $l \times p$ random matrix W_k

The basic version of the hierarchical model can be summarized by the graphical representation in Figure 3.1.

We assume the set of parameters denoted by $\Theta = (\pi_0, \boldsymbol{\eta}, \boldsymbol{\lambda},)$ is common across all genes. More specifically, $1 - \pi_0$ describes the percentage of all genes having an eQTL in its *cis*-region, $\boldsymbol{\eta}$ describes the distribution of tissue specificities among all eQTLs,

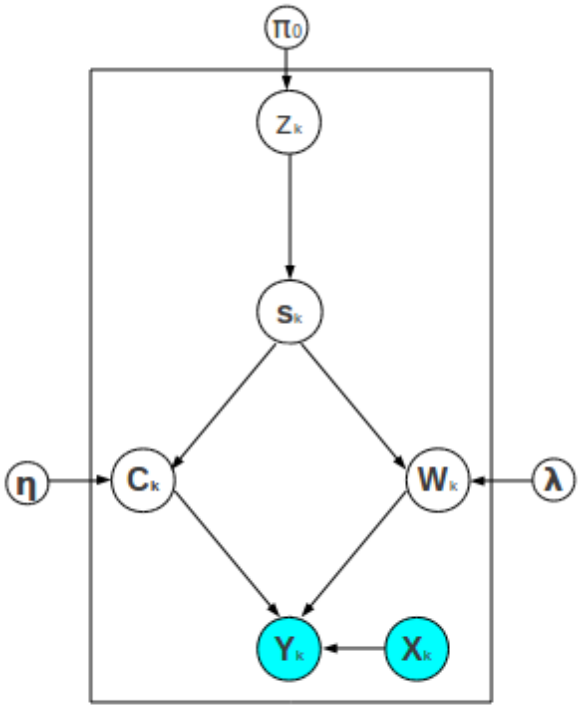

Figure 3.1: A graphical representation of the hierarchical model for modeling tissue specific expression eQTL data, where filled circles represent the data observed and unfilled circles represent latent quantities.

and $\boldsymbol{\lambda}$ describes the distribution of genetic effect sizes among all eQTLs.

3.3 Parameter Inference

Our primary interest is making inference on the parameter set $\Theta = (\pi_0, \boldsymbol{\eta}, \boldsymbol{\lambda})$, especially π_0 and $\boldsymbol{\eta}$. In this section, we discuss two distinct inference methods for Θ: maximum likelihood estimation based on EM algorithm and Bayesian inference based on MCMC.

3.3.1 Maximum Likelihood Inference

In the maximum likelihood framework, to deal with latent variables $z_k, \boldsymbol{s}_k, \boldsymbol{c}_k$ and $\boldsymbol{w}_k, k = 1, \ldots, g$, we treat them as missing data and apply the EM algorithm.

For a total number of g genes, let $\boldsymbol{z} = (z_1, \ldots, z_g), \boldsymbol{S} = (\boldsymbol{s}_1, \ldots, \boldsymbol{s}_g), \boldsymbol{C} = (C_1, \ldots, C_g)$ and $\boldsymbol{W} = (W_1, \ldots W_g)$ denote the complete set of latent variables. Let $\boldsymbol{Y} = (\boldsymbol{Y}_1, \ldots \boldsymbol{Y}_g)$ and $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_g)$ denote the complete set of observed data. Based on the hierarchical model described in previous section, we can write out the complete data log-likelihood as follows,

$$\begin{aligned}
\log\ & P(\boldsymbol{Y}, \boldsymbol{z}, \boldsymbol{S}, \boldsymbol{C}, \boldsymbol{W} | \boldsymbol{X}, \Theta) = \\
& \sum_k (1 - z_k) \log \pi_0 + \sum_k z_k \log(1 - \pi_0) \\
& + \sum_{k,p} z_k s_{kp} \log \frac{1}{m_k} + \sum_{k,p,j} z_k s_{kp} c_{kpj} \log \eta_j + \sum_{k,p,m} z_k s_{kp} w_{kpm} \log \lambda_m \\
& + \sum_{k,p,j,m} z_k s_{kp} c_{kpj} w_{kpm} \cdot \mathrm{BF}_{kpjm} + \sum_k \log p_k^0.
\end{aligned} \tag{3.9}$$

In (3.9), p_k^0 denotes the likelihood of the null model for gene k, which states no association between any SNP and expression levels in any tissue type, i.e.

$$P_k^0 := P(\boldsymbol{Y}_k | z_k = 0) \tag{3.10}$$

and

$$\mathrm{BF}_{kpjm} = \frac{P(\boldsymbol{Y}_k | z_k = 1, s_{kp} = 1, c_{kpj} = 1, w_{kpm} = 1, \boldsymbol{X}_k, \Theta)}{P_k^0} \tag{3.11}$$

is the Bayes Factor computed for a fully specified alternative model. We evaluate Bayes Factors using the numerical solutions discussed in the previous chapter.

The EM algorithm searches for maximum likelihood estimate of Θ, by iteratively performing an expectation (E) step and a maximization (M) step.

In the E-step, for the t-th iteration, we evaluate the expectation of complete data log-likelihood (3.9) conditional on current estimate of parameter $\Theta^{(t)}$, $\boldsymbol{X}$ and $\boldsymbol{Y}$. The

computation is straightforward, for examples,

$$\begin{aligned}
\mathrm{E}(z_k|\boldsymbol{Y}_k, \boldsymbol{X}_k, \Theta^{(t)}) &= \Pr(z_k = 1|\boldsymbol{Y}_k, \boldsymbol{X}_k, \Theta^{(t)}) \\
&= \frac{\Pr(z_k = 1|\Theta^{(t)}) \cdot p(\boldsymbol{Y}_k|z_k = 1, \boldsymbol{X}_k, \Theta^{(t)})}{p(\boldsymbol{Y}_k|\boldsymbol{X}_k, \Theta^{(t)})} \\
&= \frac{(1 - \pi_0^{(t)})\mathrm{BF}_k^{(t)}}{\pi_0^{(t)} + (1 - \pi_0^{(t)})\mathrm{BF}_k^{(t)}},
\end{aligned} \tag{3.12}$$

similarly,

$$\mathrm{E}(z_k s_{kp}|\boldsymbol{Y}_k, \boldsymbol{X}_k, \Theta^{(t)}) = \frac{(1 - \pi_0^{(t)})\frac{1}{m_k}\mathrm{BF}_{kp}^{(t)}}{\pi_0^{(t)} + (1 - \pi_0^{(t)})\mathrm{BF}_k^{(t)}}, \tag{3.13}$$

$$\mathrm{E}(z_k s_{kp} c_{kpj} w_{kpm}|\boldsymbol{Y}, \boldsymbol{X}, \Theta^{(t)}) = \frac{(1 - \pi_0^{(t)})\frac{1}{m_k}\eta_j^{(t)}\lambda_m^{(t)}\mathrm{BF}_{kpjm}}{\pi_0^{(t)} + (1 - \pi_0^{(t)})\mathrm{BF}_k^{(t)}}, \tag{3.14}$$

where

$$\begin{aligned}
\mathrm{BF}_k^{(t)} &= \frac{p(\boldsymbol{Y}_k|z_k = 1, \boldsymbol{X}_k, \Theta^{(t)})}{p_k^0} \\
&= \sum_{p,j,m} \frac{1}{m_k}\eta_j^{(t)}\lambda_m^{(t)}\mathrm{BF}_{kpjm},
\end{aligned} \tag{3.15}$$

and

$$\begin{aligned}
\mathrm{BF}_{kp}^{(t)} &= \frac{p(\boldsymbol{Y}_k|z_k = 1, s_{kp} = 1, \boldsymbol{X}_k, \Theta)}{p_k^0} \\
&= \sum_{j,m} \eta_j^{(t)}\lambda_m^{(t)}\mathrm{BF}_{kpjm},
\end{aligned} \tag{3.16}$$

In the M-step, we find a new set of parameters $\Theta^{(n+1)}$ to maximize the conditional expectation $\mathrm{E}\left(\log p(\boldsymbol{Y}, \boldsymbol{z}, \boldsymbol{S}, \boldsymbol{C}, \boldsymbol{W}|\boldsymbol{X}, \Theta)|\boldsymbol{Y}, \boldsymbol{X}, \Theta^{(t)}\right)$. In the basic setup of the hierarchical model, the simultaneous maximization can be performed analytically. In particular,

$$\pi_0^{(t+1)} = \frac{1}{g}\sum_{k=1}^{g} \frac{\pi_0^{(t)}}{\pi_0^{(t)} + (1-\pi_0^{(t)})\mathrm{BF}_k^{(t)}}, \tag{3.17}$$

$$\eta_j^{(t+1)} = \frac{\sum_{k,p,m} \frac{\gamma_{kp}^{(t)}\lambda_m^{(t)}\mathrm{BF}_{kpjm}}{\pi_0^{(t)}+(1-\pi_0^{(t)})\mathrm{BF}_k^{(t)}} \cdot \eta_j^{(t)}}{\sum_{j'}\left(\sum_{k,p,m} \frac{\gamma_{kp}^{(t)}\lambda_m^{(t)}\mathrm{BF}_{kpj'm}}{\pi_0^{(t)}+(1-\pi_0^{(t)})\mathrm{BF}_k^{(t)}} \cdot \eta_{j'}^{(t)}\right)}, \tag{3.18}$$

and

$$\lambda_m^{(t+1)} = \frac{\sum_{k,p,j} \frac{\gamma_{kp}^{(t)}\eta_j^{(t)}\mathrm{BF}_{kpjm}}{\pi_0^{(t)}+(1-\pi_0^{(t)})\mathrm{BF}_k^{(t)}} \cdot \lambda_m^{(t)}}{\sum_{m'}\left(\sum_{k,p,j} \frac{\gamma_{kp}^{(t)}\eta_j^{(t)}\mathrm{BF}_{kpjm'}}{\pi_0^{(t)}+(1-\pi_0^{(t)})\mathrm{BF}_k^{(t)}} \cdot \lambda_{m'}^{(t)}\right)}. \tag{3.19}$$

Typically, we initiate the EM algorithm by setting $\Theta^{(0)}$ to some random values and running iterations until some pre-defined convergence threshold is met. In practice, we monitor the increase of the the log-likelihood function between successive iterations, and stop the iterations as the increase becomes sufficiently small.

We construct profile likelihood confidence intervals for estimated parameters. For example, a $(1-\alpha)\%$ profile likelihood confidence set for π_0 is built as

$$\{\pi_0 : \log p(\boldsymbol{Y}|\pi_0, \hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\lambda}}, \boldsymbol{X}) > \log p(\boldsymbol{Y}|\hat{\pi}_0, \hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\lambda}}, \boldsymbol{X}) - \frac{1}{2}Z^2_{(1-\alpha)}\}, \tag{3.20}$$

where $\hat{\pi}_0, \hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\lambda}}$ are MLEs obtained from the EM algorithm. Note, profile likelihood confidence intervals are based on asymptotic theory which requires parameters are located in the interior of the parameter space. In boundary conditions, the resulting confidence intervals may not possess the advocated coverage probability.

3.3.2 Bayesian Inference

To perform Bayesian inference for the proposed hierarchical model, we set priors for parameters of interest $\pi_0, \boldsymbol{\eta}$ and $\boldsymbol{\lambda}$, and the result of the inference is summarized in

the posterior distributions $P(\pi_0|\boldsymbol{Y}, \boldsymbol{X})$, $P(\boldsymbol{\eta}|\boldsymbol{Y}, \boldsymbol{X})$ and $P(\boldsymbol{\lambda}|\boldsymbol{Y}, \boldsymbol{X})$. Specifically, we use the following prior distributions:

$$\pi_0 \sim \text{Beta}(\alpha_{\pi_0}, \alpha_{\pi_1}), \tag{3.21}$$

$$\boldsymbol{\eta} \sim \text{Dirichlet}(\alpha_{\eta_1}, \alpha_{\eta_2}, \dots), \tag{3.22}$$

$$\boldsymbol{\lambda} \sim \text{Dirichlet}(\alpha_{\lambda_1}, \alpha_{\lambda_2}, \dots). \tag{3.23}$$

In practice, we set $\alpha_{\pi_0} = \alpha_{\pi_1} = 1$, $\alpha_{\eta_1} = \alpha_{\eta_2} = \cdots = 1$, and $\alpha_{\lambda_1} = \alpha_{\lambda_2} = \cdots = 1$.

To simulate from posterior distributions, we provide solutions in Gibbs sampling and Metropolis-Hastings algorithm below.

Gibbs Sampler

Gibbs sampling algorithm simulates from full conditional distributions. In the setting of our hierarchical model, these full conditionals are easy to obtain. For example,

$$\begin{aligned} P(\pi_0|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{z}, \boldsymbol{S}, \boldsymbol{C}, \boldsymbol{W}, \boldsymbol{\eta}, \boldsymbol{\lambda}) &\propto P(\pi_0) \prod_{k=1}^{g} \Pr(z_k|\pi_0) \\ &\propto \pi_0^{\alpha_{\pi_0}+g-\sum_k z_k-1}(1-\pi_0)^{\alpha_{\pi_1}+\sum_k z_k-1}. \end{aligned} \tag{3.24}$$

Therefore,

$$\pi_0|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{z}, \boldsymbol{S}, \boldsymbol{C}, \boldsymbol{W}, \boldsymbol{\eta}, \boldsymbol{\lambda} \sim \text{Beta}(\alpha_{\pi_0} + g - \sum_k z_k, \alpha_{\pi_1} + \sum_k z_k). \tag{3.25}$$

Similarly, it can be shown that

$$\boldsymbol{\eta}|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{z}, \boldsymbol{S}, \boldsymbol{C}, \boldsymbol{W}, \pi_0, \boldsymbol{\lambda} \sim \text{Dirichlet}(\alpha_{\eta_1} + \sum_{k,p} z_k s_{kp} c_{kp1}, \alpha_{\eta_2} + \sum_{k,p} z_k s_{kp} c_{kp2}, \dots) \tag{3.26}$$

$$\boldsymbol{\lambda}|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{z}, \boldsymbol{S}, \boldsymbol{C}, \boldsymbol{W}, \pi_0, \boldsymbol{\eta} \sim \text{Dirichlet}(\alpha_{\lambda_1} + \sum_{k,p} z_k s_{kp} w_{kp1}, \alpha_{\lambda_2} + \sum_{k,p} z_k s_{kp} w_{kp2}, \dots) \tag{3.27}$$

For latent variables $\boldsymbol{z}, \boldsymbol{S}, \boldsymbol{C}$ and $\boldsymbol{W}$, we update them as a batch for each gene k based on the following equation:

$$\begin{aligned}\Pr(z_k, \boldsymbol{s}_k, C_k, W_k|\boldsymbol{Y}, \boldsymbol{X}, \pi_0, \boldsymbol{\lambda}, \boldsymbol{\eta}) = &\Pr(z_k|\boldsymbol{Y}, \boldsymbol{X}, \pi_0, \boldsymbol{\lambda}, \boldsymbol{\eta}) \Pr(\boldsymbol{s}_k|z_k, \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\lambda}, \boldsymbol{\eta}) \\ &\cdot \Pr(C_k|\boldsymbol{s}_k, \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\lambda}, \boldsymbol{\eta}) \Pr(W_k|\boldsymbol{s}_k, \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\lambda}, \boldsymbol{\eta})\end{aligned} \tag{3.28}$$

These conditionals are also easy to obtain,

$$z_k|\boldsymbol{Y}, \boldsymbol{X}, \pi_0, \boldsymbol{\lambda}, \boldsymbol{\eta} \sim \text{Bernoulli}\left(\frac{(1-\pi_0)\text{BF}_k}{\pi_0 + (1-\pi_0)\text{BF}_k}\right), \tag{3.29}$$

$$\boldsymbol{s}_k|z_k = 1, \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\lambda}, \boldsymbol{\eta} \sim \text{Multinomial}(1; \boldsymbol{p}_k), \tag{3.30}$$

$$\boldsymbol{c}_{kp}|s_{kp} = 1, \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\lambda}, \boldsymbol{\eta} \sim \text{Multinomial}(1; \boldsymbol{q}_{kp}), \tag{3.31}$$

$$\boldsymbol{w}_{kp}|s_{kp} = 1, \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\lambda}, \boldsymbol{\eta} \sim \text{Multinomial}(1; \boldsymbol{r}_{kp}), \tag{3.32}$$

where $\boldsymbol{p}_k, \boldsymbol{q}_{kp}$ and $\boldsymbol{r}_{kp}$ are probability vectors, with

$$p_{kp} = \frac{\text{BF}_{kp}}{\sum_p \text{BF}_{kp}},\ p = 1, \ldots, m_k \tag{3.33}$$

$$q_{kpj} = \frac{\eta_j \sum_m \lambda_m \text{BF}_{kpjm}}{\text{BF}_{kp}},\ j = 1, \ldots, 2^t - 1 \tag{3.34}$$

and

$$r_{kpm} = \frac{\lambda_m \sum_j \eta_j \text{BF}_{kpjm}}{\text{BF}_{kp}},\ m = 1, \ldots, l \tag{3.35}$$

The definition of $\text{BF}_{kpjm}, \text{BF}_{kp}$ and BF_k follows from (3.11), (3.16) and (3.15) respectively.

Given the above conditional distributions, the Gibbs sampling algorithm proceeds as follows:

1. *initialize* $(\pi_0, \boldsymbol{\eta}, \boldsymbol{\lambda}, \boldsymbol{z}, \boldsymbol{S}, \boldsymbol{C}, \boldsymbol{W})$ *at some random values.*

2. *repeat following steps for* $t = 0, 1, 2, \ldots$

    2.1 for each gene $k = 1, \ldots, g$

- *sample* $z_k^{(t+1)}$ *from* Bernoulli $\left(\frac{(1-\pi_0^{(t)})\mathrm{BF}_k^{(t)}}{\pi_0^{(t)}+(1-\pi_0^{(t)})\mathrm{BF}_k^{(t)}}\right)$.
  - *if* $z_k^{(t+1)}=0$, *set* $\boldsymbol{s}_k^{(t+1)}=\mathbf{0}, C_k^{(t+1)}=\mathbf{0}, W_k^{(t+1)}=\mathbf{0}$.
  - *otherwise*
    - *i. sample* $\boldsymbol{s}_k^{(t+1)}$ *from* Multinomial$(1;\boldsymbol{p}_k^{(t)})$.
    - *ii. for each SNP* $p=1,\ldots,m_k$
      - *if* $s_{kp}^{(t+1)}=0$, *set* $\boldsymbol{c}_{kp}^{(t+1)}=\mathbf{0}$ *and* $\boldsymbol{w}_{kp}^{(t+1)}=\mathbf{0}$.
      - *if* $s_{kp}^{(t+1)}=1$, *then*
        - *A. sample* $\boldsymbol{c}_{kp}^{(t+1)}$ *from* Multinomial$(1;\boldsymbol{q}_{kp}^{(t)})$.
        - *B. sample* $\boldsymbol{w}_{kp}^{(t+1)}$ *from* Multinomial$(1;\boldsymbol{r}_{kp}^{(t)})$.

2.2 sample $\pi_0^{(t+1)}$ *from* Beta$(\alpha_{\pi_0}+g-\sum_k z_k^{(t+1)}, \alpha_{\pi_1}+\sum_k z_k^{(t+1)})$.

2.3 sample $\boldsymbol{\eta}^{(t+1)}$ *from* Dirichlet$(\alpha_{\eta_1}+\sum_{k,p} z_k^{(t+1)} s_{kp}^{(t+1)} c_{kp1}^{(t+1)},\ldots)$

2.4 sample $\boldsymbol{\lambda}^{(t+1)}$ *from* Dirichlet$(\alpha_{\lambda_1}+\sum_{k,p} z_k^{(t+1)} s_{kp}^{(t+1)} w_{kp1}^{(t+1)},\ldots)$

Metropolis-Hastings Algorithm

If our interest for inference is solely on parameter $\Theta=(\pi_0,\boldsymbol{\eta},\boldsymbol{\lambda})$, we can avoid sampling steps on the latent indicator variables by a direct implementation of the Metropolis-Hastings algorithm. From the previous section, we have noticed that the latent variables $(\boldsymbol{z},\boldsymbol{S},\boldsymbol{C},\boldsymbol{W})$ can be analytically integrated out and the likelihood of the basic hierarchical model is available in closed form:

$$P(\boldsymbol{Y}|\Theta,\boldsymbol{X})=\prod_{k=1}^{g}\left(\pi_0+(1-\pi_0)\mathrm{BF}_k\right)p_k^0, \tag{3.36}$$

where p_k^0 and BF_k follow from (3.10) and (3.15) respectively.

Therefore, we can apply a very generic version of Metropolis-Hastings algorithm. Let $q(\Theta'|\Theta)$ denote the proposal distribution, we outline the algorithm as follows,

1. *initialize* $\Theta^{(0)}=(\pi_0^{(0)},\boldsymbol{\eta}^{(0)},\boldsymbol{\lambda}^{(0)})$ *at some random values.*

2. *repeat following steps for* $t=0,1,2,\ldots$:

2.1 generate $\tilde{\Theta}$ *from* $q(\tilde{\Theta}|\Theta^{(t)})$.

2.2 compute

$$r = \frac{P(\tilde{\Theta})P(\boldsymbol{Y}|\tilde{\Theta}, \boldsymbol{X})q(\Theta^{(t)}|\tilde{\Theta})}{P(\Theta^{(t)})P(\boldsymbol{Y}|\Theta^{(t)}, \boldsymbol{X})q(\tilde{\Theta}|\Theta^{(t)})}.$$

2.3 generate $u \sim \text{Uniform}(0, 1)$,

- *if* $u < r$, *set* $\Theta^{(t+1)} = \tilde{\Theta}$;
- *else* $\Theta^{(t+1)} = \Theta^{(t)}$.

The only difficulty in the above algorithm is the choice of proposal distribution q. For commonly used random walk Metropolis-Hastings algorithm, the proposal distribution proposes new moves in the following fashion:

$$\tilde{\pi}_0 = \pi_0^{(t)} + \Delta_{\pi_0}, \tag{3.37}$$

where $\Delta_{\pi_0} \sim \text{N}(0, \sigma^2)$. However in our settings, all parameters of interest are constrained and located in some probability simplex, this type of the random walk proposal becomes very inefficient, especially when $\Theta^{(t)}$ are close to the boundary of the parameter space.

To overcome this problem, we follow Cappe *et al.* (2003) to over-parameterize $\pi_0, \boldsymbol{\eta}$ and $\boldsymbol{\lambda}$ using a set of independent Gamma random variables. For example, we let

$$\eta_j = \frac{\omega_j}{\sum_i \omega_i}, \ \omega_i \sim \text{Gamma}(\alpha_{\eta_i}, 1) \tag{3.38}$$

Under this parameterization, we maintain the prior distribution $\boldsymbol{\eta} \sim \text{Dirichlet}(\alpha_{\eta_1}, \alpha_{\eta_2}, \dots)$ as desired. Although, ω_{η_i}'s are obviously unidentifiable, this is not a problem for our purpose: η_i's still remain identifiable. To further relax the positivity restrictions on Gamma random variables, we conveniently apply random walk proposals on $\log \omega_{\eta_i}$'s whose induced prior distributions are very easy to obtain.

3.4 Model Extensions

Under the hierarchical model described above, we are able to evaluate the scope of tissue specificity for eQTLs. However, in order to identify biological features that are linked to eQTLs and/or their tissue specificity, we need extend our model from the basic version.

To formulate the problem, let us restrict ourselves to R categories of genomic annotations. These annotations either give simple measurements of a genetic variant with respect to some genomic landmarks, e.g. the distance of a SNP to the transcription start site of a gene, or denote the relationship between a genetic variant and a known genomic functional unit, e.g. if a SNP locates in a known enhancer region. For our purpose, we assume these R categories are pre-defined and the measurements/annotations of these R features are available for each SNP. With multiple-tissue eQTL data, our goal is to investigate, among those R genomic features,

- which features are associated with SNPs being eQTLs
- which features are associated with tissue specificity of eQTLs

Recall, in the basic hierarchical model, for gene k and SNP p, we assume

$$\Pr(s_{kp} = 1 | z_k = 1) = \gamma_{kp} = \frac{1}{m_k}, \tag{3.39}$$

$$\Pr(c_{kpj} = 1 | s_{kp} = 1) = \eta_j. \tag{3.40}$$

Here, we modify above prior specifications and connect them to annotation information through logistic functions. Let $(\delta_{kp1}, \ldots, \delta_{kpR})$ denote the complete annotations of SNP p with respect to gene k. To formulate the prior probability of SNP p being eQTL in gene k , we first define

$$d^s_{kp} = \sum_{r=1}^{R} \delta_{kpr} \beta^s_r, \tag{3.41}$$

and then let

$$\Pr(s_{kp} = 1 | z_k = 1) = \frac{\exp(d^s_{kp})}{\sum_{p'=1}^{m_k} \exp(d^s_{kp'})}. \tag{3.42}$$

Here, the parameter vector $\boldsymbol{\beta}^s = (\beta^s_1, \ldots, \beta^s_R)$ characterizes the strength of association between the genomic features and the plausibility of a SNP being eQTL. In the special case of $\boldsymbol{\beta}^s = 0$, this prior assumes each SNP is equally likely to be the only eQTL in the *cis*-region.

The general idea also applies in re-parameterizing prior configuration probability $\boldsymbol{\eta}$. We note, in most cases, the main interest is in distinguishing the consistent configuration from the tissue-specific configurations. Let us use the convention that the first entry of the indicator vector $\boldsymbol{c}$ always denotes the consistent configuration. For SNP p in gene k, we define

$$d^c_{kp} = \sum_{r=1}^{R} \delta_{kpr} \beta^c_r, \tag{3.43}$$

and assume the prior probability of a tissue consistent configuration is

$$\Pr(c_{kp1} = 1 | s_{kp} = 1) = \frac{\exp(\mu^c + d^c_{kp})}{1 + \exp(\mu^c + d^c_{kp})}. \tag{3.44}$$

For tissue-specific configurations,

$$\Pr(c_{kpj'} = 1 | s_{kp} = 1) = \frac{\eta'_{j'}}{1 + \exp(\mu^c + d^c_{kp})}, j' = 2, \ldots, 2^t - 1, \tag{3.45}$$

or equivalently,

$$\Pr(c_{kpj'} = 1 | s_{kp} = 1, c_{kp1} = 0) = \eta'_{j'}, j' = 2, \ldots, 2^t - 1. \tag{3.46}$$

The parameter vector $\boldsymbol{\beta}^c = (\beta^c_1, \ldots, \beta^c_R)$ captures the impacts of annotation features to the tissue specificity of eQTLs. In the special case of $\boldsymbol{\beta}^c = 0$, the probability of tissue specificity is solely parameterized by the single intercept term μ^c, which is very

similar as the treatment in the basic hierarchical model.

In the extended hierarchical model, the inference on $\Theta = (\pi_0, \boldsymbol{\lambda}, \boldsymbol{\beta}^s, \boldsymbol{\beta}^c, \boldsymbol{\eta}')$ becomes more challenging. Although the general maximum likelihood and Bayesian framework still apply, the numerical solutions inevitably are more complicated and computationally intensive.

For implementation of the EM algorithm, the E-step is intact. However, we no longer have closed form solutions for simultaneous maximizing Θ in the M-step. It is possible to use numerical maximization routines to resolve this issue, nevertheless, when the number of annotation features considered, R, is large, all numerical maximization routines attempting simultaneous optimization become unstable. An efficient alternative strategy is to apply conditional maximization (CM) procedure proposed by Meng and Rubin (1993), in which maximization is performed with respect to a single parameter at a time while conditional on current values of the other parameters.

For Bayesian solutions, given priors for $\boldsymbol{\beta}^s$ and $\boldsymbol{\beta}^c$, the extended model has very little impact on the implementation of the Metropolis-Hastings algorithm (other than complicating the computation of likelihood function). Howeve, for the Gibbs sampler, the full conditional distributions $\boldsymbol{\beta}^s|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\beta}^c, \pi_0, \boldsymbol{\eta}', \boldsymbol{\lambda}$ and $\boldsymbol{\beta}^c, \mu^c|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\beta}^s, \pi_0, \boldsymbol{\eta}', \boldsymbol{\lambda}$ can not be obtained from the known distribution families. We solve this issue by replacing the full conditional sampling steps by embedding Metropolis-Hastings steps.

The choice of priors for $\boldsymbol{\beta}^s$ and $\boldsymbol{\beta}^c$ are worth discussing. If the purpose of fitting the model is to estimate the effect size of each component, it is reasonable to assign normal priors, or even improper flat priors, for these parameters. On the other hand, if the goal is to determine from the data if a particular feature has some impact on eQTLs, the problem is better framed as a model/feature selection problem. In this scenario, we prefer a "spike and slat" type prior, for example, a mixture of point mass at 0 and a normal distribution centered at 0. We then can use the posterior inclusion probability to quantify the importance of a particular genomic feature. On the same note, if feature selection, instead of effect estimation, is the goal in this context, the current implementation of EM algorithm might not be sufficient to address the problem.

3.5 Data Application

We apply the proposed hierarchical model to the multiple-tissue eQTL data collected in Dimas *et al.* (2009). In their experiment, 85 unrelated individuals of West European origins were investigated for *cis*-eQTLs in three cell types: primary fibroblasts, Epstein-Barr virus-immortalized B cells (lymphoblastoid cell lines or LCLs), and T cells. Gene expression profiling was performed for 48,804 probes with the IlluminaWG-6 v3 expression array and all 85 individuals were genotyped on the Illumina 550K SNP array. The gene expression data went through quality controls and normalization steps by the original authors and were made available through NIH Gene Expression Omnibus (GEO) project.

From all available genes, we select a subset that is highly likely to be expressed in all three cell types. The rationale of this selection step is that if a gene is not expressed in certain cell type, the expression measurement in that cell for this gene is merely noise and therefore in principle is not associated with any genetic variant. This scenario complicates the biological explanation of a finding of no association in certain cell types: for example, an inconsistent association pattern across cell types might be due to *differential expressions*. To focus on identifying patterns of *differential regulation*, we limit ourselves to analyzing genes that are actually expressed in all cell types studied. In particular, we rank probes according to their average expression intensities across individuals in three cell types separately and select only probes whose average intensities are higher than the corresponding median values in *all* three cell types. After this selection step, we end up with probes mapping to 5,970 unique genes. We then perform additional quantile normalization of expression measurements for each selected gene within each cell type. As in the original paper by Dimas *et al.* (2009), we focus on the genetic variants located in the *cis*-region of the genes, for which we define as the genomic region enclosed by 1Mb upstream of the transcription start site (TSS) and 1Mb downstream of the transcription end site (TES) of a gene.

In the experiment, the same set of individual samples are used for expression measurements of all three cell types. This type of the design may cause issue for

applying our Bayesian meta-analysis methods, because even under the null model of no eQTL, gene expression levels in different cell types within same individual might be correlated. To check this, we examine the Pearson's correlations of gene expressions within an individual for each gene between each pair of tissue types. We find average correlation between Fibroblast cell and B-cell is 0.019 (median 0.016), between Fibroblast cell and T-cell is 0.021 (median 0.018) and between B-cell and T-cell is 0.029 (median 0.027). These results suggest empirically gene expression levels in different cell types are approximately uncorrelated in our data for a typical gene.

3.5.1 Use of the Basic Hierarchical Model

We first apply the basic hierarchical model to investigate the scope of the tissue-specificity of eQTLs. We pre-calculate the Bayes Factors for each gene and each SNP under all possible alternative scenarios (i.e. BF_{kpmj} in (3.11)). These values serve as common inputs for all three fitting methods (EM, Gibbs sampling and Metropolis-Hastings). To compute these Bayes Factors, we apply the methods described in the previous chapter. In particular, for 3 cell types, we enumerate all 7 possible non-null eQTL activity configurations and apply the ES model with CEFN prior with $k = 0.1$ and $\omega = 0.4, 0.6, 0.8, 1.0, 1.2$, which corresponds to $l = 5$ different effect size priors.

We run all three proposed algorithms on this data set. For the EM algorithm, we start the iteration at some random starting point and stop the iteration when the increment of log-likelihood is less than 0.005. For Bayesian inference, we use prior distributions in (3.21)-(3.23) and set $\alpha_{\pi_0} = \alpha_{\pi_1} = 1$, $\alpha_{\eta_1} = \alpha_{\eta_2} = \cdots = 1$, and $\alpha_{\lambda_1} = \alpha_{\lambda_2} = \cdots = 1$. For the Gibbs sampling and Metropolis-Hastings algorithms, we run both samplers for 10,000 iterations with some random starting points and discard the first 5,000 iterations as burnin. We also repeat the procedures for all three algorithms by changing to different starting points, all results from different runs are quite consistent.

Figure 3.2 shows log-likelihood values explored by EM and Metropolis-Hastings algorithms. The EM run takes 121 iterations to reach pre-defined convergence threshold, the MCMC simulation also takes around 100 iterations to settle at the same

magnitude of likelihood surface. In terms of time efficiency, the EM run takes about 2 hours to complete, and both Metropolis-Hastings and Gibbs sampler take about 20 hours to complete.

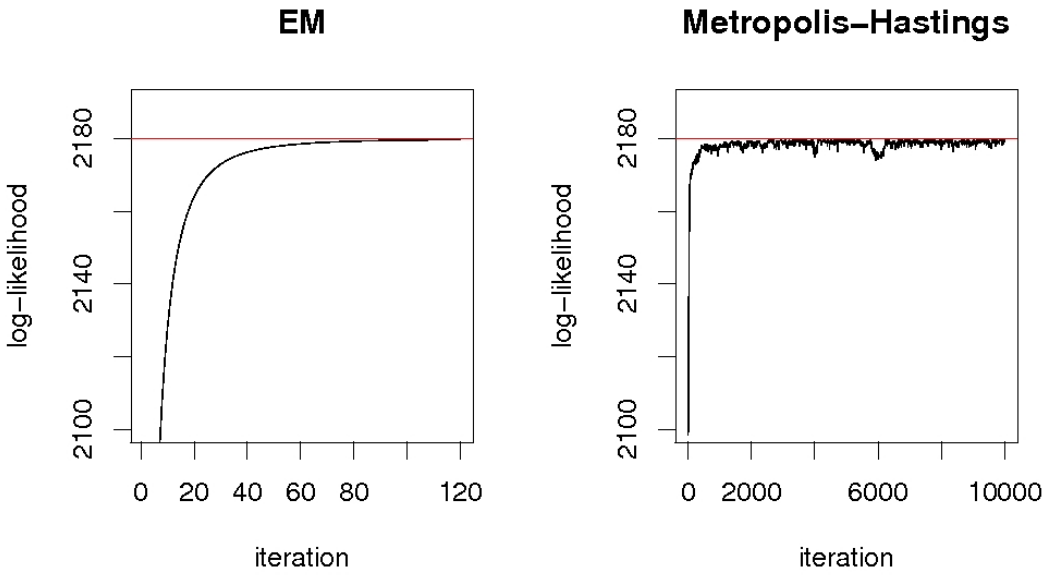

Figure 3.2: Trace plot of log-likelihood values explored by EM and Metropolis-Hastings algorithms.

Figure 3.3 and Figure 3.4 show trace plots and histograms of posterior samples for π_0 and η value corresponding to tissue-consistent configuration (i.e. eQTL associations are preserved in all three cell types) from the Gibbs sampler and the Metropolis-Hasting algorithm respectively. In both occasions, Markov chains settle down at values around eventual posterior mode very quickly.

We show the inference results of π_0 and $\boldsymbol{\eta}$ from all three fitting methods in Table 3.1. For the EM algorithm, we report the MLEs and corresponding 95% profile likelihood confidence intervals; for the Gibbs sampler and the Metropolis-Hastings algorithm, we report the posterior means and 95% credible intervals. We observe that not only the point estimates are very close from different methods, the estimated intervals are also aligned quite well.

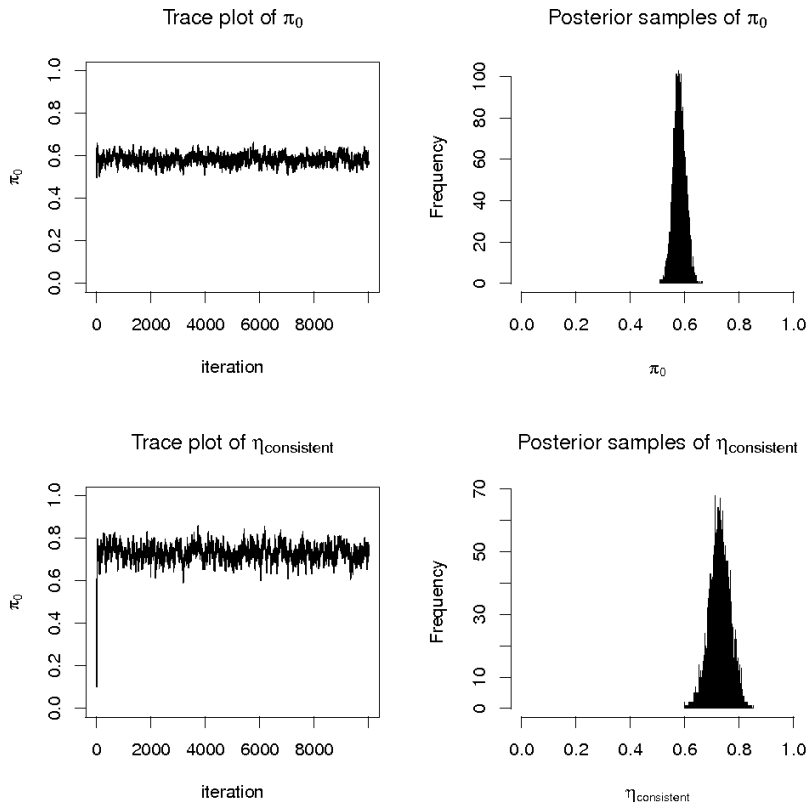

Figure 3.3: Trace plots and histograms of posterior samples of π_0 and $\eta_{\text{consistent}}$ (η value corresponding to the consistent configuration) from a Gibbs sampler run.

Based on these inference results, we conclude there are many eQTLs (around 25% among all eQTLs) behaving in a tissue specific manner. Among three examined cell types, B-cell and T-cell share much larger percentage of common eQTLs than either Fibroblast cell and T-cell or Fibroblast cell and B-cell share. From biological point of view, this result is not surprising: B-cell and T-cell both belong to the category of immune cells and Fibroblast cell is functionally more distant than both of them.

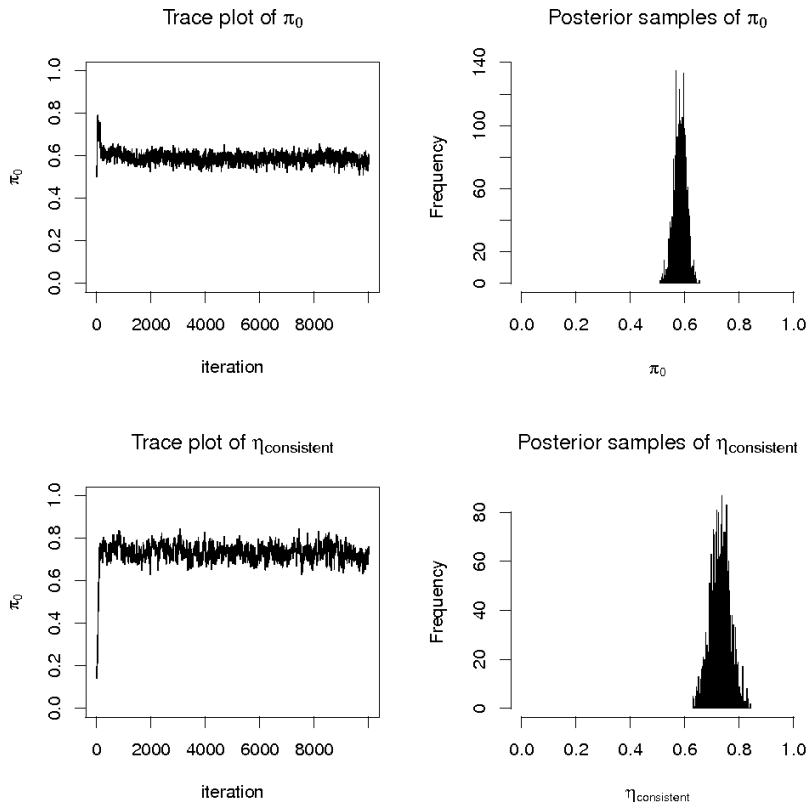

Figure 3.4: Trace plots and histograms of posterior samples of π_0 and $\eta_{\text{consistent}}$ from a Metropolis-Hastings algorithm run.

3.5.2 Impact of Genomic Features on eQTLs

In this section, we apply the extended hierarchical model described in section 3.4 to study the association between genomic features of the property of eQTLs. For demonstration purpose, we focus only on the SNP distance to the TSS of the target gene. In what follows, we call this specific genomic feature DTSS in brief. In particular, we are interested in investigating the impact of DTSS on the following properties of eQTLs:

1. the probability of a SNP being an eQTL

2. the tissue specificity of an eQTL

	EM	Gibbs Sampler	Metropolis-Hastings	Notes
π_0	0.579 (0.544, 0.611)	0.581 (0.540, 0.624)	0.584 (0.542, 0.621)	–
$\eta_{(100)}$	0.086 (0.044, 0.138)	0.082 (0.037, 0.136)	0.079 (0.036, 0.135)	F only
$\eta_{(010)}$	0.063 (0.026, 0.111)	0.063 (0.023, 0.112)	0.062 (0.022, 0.108)	B only
$\eta_{(001)}$	0.001 (0.000, 0.023)	0.009 (0.000, 0.031)	0.008 (0.001, 0.030)	T only
$\eta_{(110)}$	0.002 (0.000, 0.026)	0.010 (0.000, 0.034)	0.014 (0.003, 0.043)	F and B
$\eta_{(101)}$	0.000 (0.000, 0.014)	0.006 (0.000, 0.020)	0.006 (0.000, 0.024)	F and T
$\eta_{(011)}$	0.102 (0.057, 0.152)	0.101 (0.052, 0.151)	0.103 (0.060, 0.152)	B and T
$\eta_{(111)}$	0.746 (0.685, 0.805)	0.729 (0.653, 0.801)	0.729 (0.662, 0.796)	Consistent

Table 3.1: Inference results for π_0 and $\boldsymbol{\eta}$ from EM, Gibbs sampler and Metropolis-Hastings algorithm. For EM algorithm, MLE and 95% profile likelihood confidence intervals are reported; for MCMC methods posterior mean and 95% credible intervals are shown. The subscript for η indicates the eQTL activity configuration in Fibroblast cell, B-cell and T-cell respectively. e.g. subscript (011) indicates the type of eQTLs that are active in T-cell and B-cell but inactive in Fiborblast cell.

The impact of DTSS on eQTLs has been previously studied. Veyrieras *et al.* (2008), Stranger *et al.* (2007) showed the enrichment of eQTL signals in the immediate neighborhood of TSS using a dataset of a single cell type. In particular, Veyrieras *et al.* (2008) quantitatively evaluated the strength of the impact by fitting a hierarchical model. Dimas *et al.* (2009) also examined this genomic feature using the data described in the beginning of this section. By visual inspection of called eQTLs, they confirmed the eQTL abundance in the near TSS region. Moreover, they also claim (also by visual inspection) that their data suggest that tissue-consistent eQTLs "tend to cluster tightly around the TSS".

We use the same subset of Dimas *et al.* (2009) data as in the basic hierarchical model to examine the association between DTSS and the properties of eQTLs. We first perform the exploratory analyses discussed in the previous chapter. For this purpose, we use the ES model with five levels of $\sqrt{\phi^2+\omega^2}$ values: 0.4, 0.6, 0.8, 1.0, 1.2, and seven degrees of heterogeneities characterized by ϕ^2/ω^2 values: $0, 1/4, 1/2, 1, 2, 4, \infty$. We assign these 35 grid values equal prior weight and compute the Bayes Factor to represent the overall evidence of a SNP being an eQTL . For the 5,970 genes considered, we select the top associated SNP for each gene based on the value of $\mathrm{BF}^{\mathrm{ES}}_{\mathrm{all}}$. We further selected a subset of SNPs for which $\mathrm{BF}^{\mathrm{ES}}_{\mathrm{all}} > 100$ and exam-

ine the heterogeneities of eQTL effects across cell types by computing $\mathrm{BF}^{\mathrm{ES}}_{\mathrm{maxH}}/\mathrm{BF}^{\mathrm{ES}}_{\mathrm{fix}}$ (we use this simple statistic as a proxy for tissue specificity). We plot resulting $\log_{10}\left(\mathrm{BF}^{\mathrm{ES}}_{\mathrm{all}}\right)$ and $\log_{10}\left(\mathrm{BF}^{\mathrm{ES}}_{\mathrm{maxH}}/\mathrm{BF}^{\mathrm{ES}}_{\mathrm{fix}}\right)$ against DTSS respectively in Figure 3.5. There is a clear pattern indicating strong eQTLs tend to cluster in the immediate neighborhood of TSS. To show this pattern is not an artifact due to the sampling bias of the SNPs, we include a histogram of DTSS of all examined SNPs in the same figure. The relationship between the heterogeneity of the effects across tissue types and DTSS is less conclusive from the plot: there seems to be excessive of consistent eQTLs ($\log_{10}\left(\mathrm{BF}^{\mathrm{ES}}_{\mathrm{maxH}}/\mathrm{BF}^{\mathrm{ES}}_{\mathrm{fix}}\right) \leq 0$) that are close to TSS, however the strongest tissue specific eQTL signals are also near respective TSS.

Although the exploratory analyses are informative, they do not sufficiently quantify the impact of DTSS. We then apply the extended hierarchical model to evaluate the association between DTSS and the eQTL properties of interest. We include the measurement of DTSS as the only genomic feature and perform the inference on β^s_{DTSS} (measuring the effect of DTSS on the probability of a SNP being an eQTL) and β^c_{DTSS}(measuring the effect of DTSS on the probability of an eQTL being tissue consistent). Instead of using the genomic distance (in base pairs) as the measure of DTSS, we follow Veyrieras *et al.* (2008) and subdivide the *cis*-region into discrete bins and represent the DTSS of a particular SNP with its corresponding bin number (with longer distance, the bin number is larger).

We fit the hierarchical model using the Metropolis-Hastings algorithm. Moreover, we assume improper flat priors for $\beta^s_{\mathrm{DTSS}}, \beta^c_{\mathrm{DTSS}}$ and μ^c (we also experiment with the proper normal priors, the inference result does not seem to be sensitive to particular prior choices), and keep priors for other parameters in the same form as in the basic model. We run the sampler for 10,000 iterations and discard the first 5,000 iterations as burnin.

The estimated posterior mean for β^s_{DTSS} is -0.303 with 95% credible interval $(-0.316, -0.290)$, which indicates a strong effect of DTSS on the probability of a SNP being an eQTL: the odds ratio are generally higher for SNPs nearby TSS. The effect of DTSS on tissue specificity of an eQTL is much weaker: with posterior mean for β^c_{DTSS} is estimated as -0.004 and corresponding 95% credible interval

$(-0.035,\ 0.035)$. In comparison, parameter μ^c has posterior mean 0.834 and 95% credible interval $(0.496, 1.114)$. The conclusions we obtained here are mostly consistent with the exploratory analyses.

Our conclusions are seemingly inconsistent with the findings reported by Dimas *et al.* (2009). However, it is worth re-emphasizing that our analysis is only based on the subset of the genes that are highly expressed in all three cell types (for better biological interpretation of results), while in Dimas *et al.* (2009), the complete set of the experimented genes is used. This aspect alone perhaps explains the most of the difference in results. Furthermore, Dimas *et al.* (2009) called eQTLs in separate cell types independently and they visually inspected the patterns of correlation between DTSS and tissue-specifities of called eQTLs. Besides the drawbacks of potential unequal power in separately calling eQTLs in different tissues, the fundamental difficulty in this type of *post-hoc* analysis is to incorporate the uncertainties of eQTL callings and corresponding tissue-specificity assessments. In comparison, our hierarchical model naturally resolves this issue.

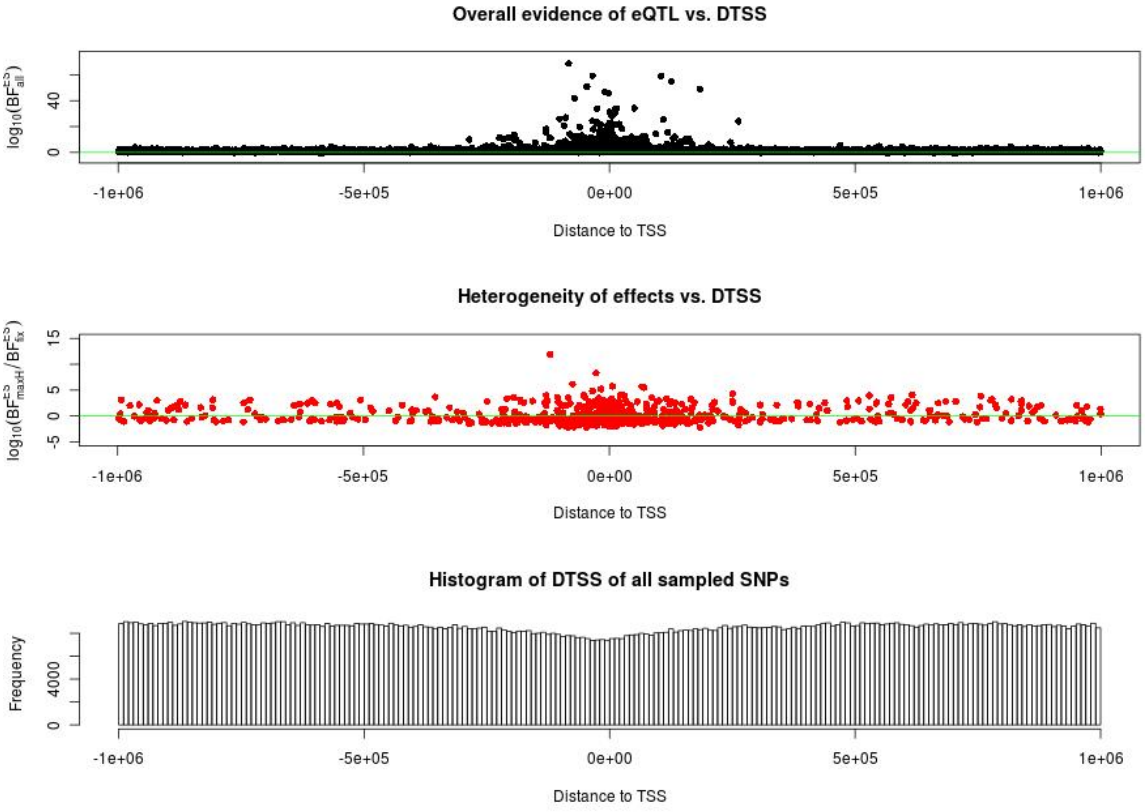

Figure 3.5: The exploratory analysis of the distance to the transcription start site (DTSS) of the target gene with respect to eQTL properties. On the top panel, the plot shows the relationship of overall evidence of an eQTL vs. DTSS: clearly, stronger signals tend to cluster in the close region of TSS. The middle panel shows the measure of effect heterogeneity vs. DTSS for relatively strong eQTL signals ($\mathrm{BF}^{\mathrm{ES}}_{\mathrm{all}} > 100$). The histogram of DTSS of all examined *cis*-SNPs for 5,490 selected genes are plotted in the bottom panel: there is no pattern of over-sampling of SNPs that are close to TSS.

3.6 Discussion and Future works

In this chapter, we introduce hierarchical model approaches to study tissue specific eQTLs. With the statistical tools built in the previous chapter, we are equipped to evaluate tissue specificity of a single potential eQTL for a single gene. Built on these tools, the essence of the hierarchical models is effective pooling of information across genes that are simultaneously studied in the experiment. With the proposed hierarchical models, we are able to evaluate the scope of tissue specificity of eQTLs and investigate common biological features that are linked to differential gene regulations. In fact, the proposed hierarchical models are very general to study gene-environment interactions in the genomics context: as long as the data can be clustered into pre-defined subgroups, the hierarchical model applies. For example, it can be readily used in studies of population-specific, sex-specific, and more interestingly, disease-specific gene regulations. Most recently, we have applied the basic version of the hierarchical model to interrogate potential drug effects on change of gene regulation landscape.

We show both maximum likelihood and Bayesian inference can be performed in the proposed hierarchical model framework. For the basic version, through simulations (not shown) and real data applications, we observe that both types of inference yield very similar results on point and interval estimation. In fact, the EM algorithm seems to have higher computational efficiency in practice. This perhaps is because the data we have been using are generally quite informative and the likelihood surfaces are typically unimodal. However, when dealing with feature selection problem in the extended hierarchical model, Bayesian solutions can be very naturally formulated by incorporating appropriate "spike and slab" priors and the results from the inference are more natural to interpret.

There are a few aspects of current hierarchical model that we hope to improve through future works:

- Large number of tissue types. The current model relies on enumerating all possible eQTL activity configurations. This strategy works fine for small number of tissue types, however when tissue type increases, the number of configurations grows exponentially fast, which make enumeration of all possibilities computa-

tionally daunting and practically impossible.

- Allowing multiple eQTLs for a given gene. The current assumption of "one *cis*-eQTL per gene" is unlikely realistic but rather convenient for reducing computational burden. In practice, we expect that imposing this assumption has the side effect of potentially under-estimate percentage of tissue-specific eQTLs: imagine a gene having two distinct eQTLs in its *cis*-region, with one being tissue-consistent and the other being tissue-specific. If the genetic effects of the eQTLs are similar while they are active, it is typically the case that the tissue-consistent eQTL has a larger single SNP Bayes Factor than the tissue-specific eQTL. With the "one *cis*-eQTL per gene" assumption, each potential eQTL SNP is essentially weighted by its single SNP Bayes Factor for being the only true eQTL (see, for example, (3.30) and (3.33)). As a result, the tissue-consistent eQTL gets up-weighted (and ideally we hope the two eQTLs should be equally weighted).

We hope to address these issues in the immediate future: these efforts could greatly increase the flexibility and soundness of the current hierarchical model.

3.7 Acknowledgements

We thank Manolis Dermitzakis and Stylianos Antonarakis for sharing the data of Dimas *et al.* (2009) (data entry EGAD00000000027 at EGA EBI). We are also grateful to Jean-Baptiste Veyrieras, Kevin Bullaughey and Jonathan Pritchard for helpful discussions.

CHAPTER 4
USING LINEAR PREDICTORS TO IMPUTE ALLELE FREQUENCIES FROM SUMMARY OR POOLED GENOTYPE DATA

4.1 Introduction

In this chapter, we discuss a rather specific statistical issue often encountered in analysis of potentially heterogeneous genetic association data: the handling of the missing genotypes. Genotype imputation (Servin and Stephens (2008), Guan and Stephens (2008), Marchini *et al.* (2007), Howie *et al.* (2009), Browning and Browning (2007), Huang *et al.* (2009)) has recently emerged as a useful tool in the analysis of genetic association studies as a way of performing tests of association at genetic variants (specifically SNPs) that were not actually measured in the association study. In brief, the idea is to exploit the fact that untyped SNPs are often correlated, in a known way, with one or more typed SNPs. Imputation-based approaches exploit these correlations, using observed genotypes at typed SNPs to estimate, or impute, genotypes at untyped SNPs, and then test for association between the imputed genotypes and phenotype, taking account of uncertainty in the imputed genotypes. (Although in general statistics applications the term "imputation" may imply replacing unobserved data with a single point estimate, in the genetic context it is often used more broadly to include methods that consider the full conditional distribution of the unobserved genotypes, and this is the way we use it here.) These approaches have been shown to increase overall power to detect associations by expanding the number of genetic variants that can be tested for association (Servin and Stephens (2008), Marchini *et al.* (2007)), but perhaps their most important use has been in performing meta-analysis of multiple studies that have typed different, but correlated, sets of SNPs (e.g. Zeggini *et al.* (2008)).

Existing approaches to imputation in this context have been developed to work with individual-level data: given genotype data at typed SNPs in each individual they attempt to impute the genotypes of each individual at untyped SNPs. From

a general statistical viewpoint, one has a large number of correlated discrete-valued random variables (genotypes), whose means and covariances can be estimated, and the aim is to predict values of a subset of these variables, given observed values of all the other variables. Although one could imagine applying off-the-shelf statistical methods to this problem (e.g. Yu and Schaid (2007) consider approaches based on linear regression), in practice the most successful methods in this context have used purpose-built methods based on discrete Hidden Markov Models (HMMs) that capture ideas from population genetics (e.g. Li and Stephens (2003), Scheet and Stephens (2005)).

In this chapter we consider a related, but different, problem: given the *frequency* of each allele at all typed SNPs, we attempt to impute the *frequency* of each allele at each untyped SNP. We have two main motivations for considering this problem. The first is that, although most large-scale association studies collect individual-level data, it is often the case that, for reasons of privacy (Homer *et al.* (2008b), Sankararaman *et al.* (2009)) or politics, only the allele frequency data (e.g. in cases vs controls) are made available to the research community at large. The second motivation is an experimental design known as DNA pooling (Homer *et al.* (2008a), Meaburn *et al.* (2006)), in which individual DNA samples are grouped into "pools" and high-throughput genotypings are performed on each pool. This experimental design can be considerably cheaper than separately genotyping each individual, but comes at the cost of providing only (noisy) estimates of the allele frequencies in each pool. In this setting the methods described here can provide not only estimates of the allele frequencies at untyped SNPs, but also more accurate estimates of the allele frequencies at typed SNPs.

From a general statistical viewpoint this problem of imputing frequencies is not so different from imputing individual genotypes: essentially it simply involves moving from discrete-valued variables to continuous ones. However, this change to continuous variables precludes direct use of the discrete HMM-based methods that have been applied so successfully to impute individual genotypes. The methods we describe here come from our attempts (appendix I) to extend and modify these HMM-based approaches to deal with continuous data. In doing so we end up with a considerably simplified method that might be considered an off-the-shelf statistical approach: in

essence, we model the allele frequencies using a multivariate normal distribution, which results in unobserved frequencies being imputed using linear combinations of the observed frequencies (as in Kriging, for example). Some connection with the HMM based approaches remains though, in how we estimate the mean and variance-covariance matrix of the allele frequencies. In particular, consideration of the HMM-based approaches lead to a natural way to regularize the estimated variance-covariance matrix, making it sparse and banded: something that is important here for both computational and statistical reasons. The resulting methods are highly computationally efficient, and can easily handle very large panels (phased or unphased). They are also surprisingly accurate, giving estimated allele frequencies that are similar in accuracy to those obtained from state-of-the-art HMM-based methods applied to individual genotype data. That is, one can estimate allele frequencies at untyped SNPs almost as accurately using only the *frequency* data at typed SNPs as using the *individual* data at typed SNPs. Furthermore, when individual-level data are available one can also apply our method to imputation of individual genotypes (effectively by treating each individual as a pool of 1), and this results in imputation accuracy very similar to that of state-of-the-art HMM-based methods, at a fraction of the computational cost. Finally, in the context of noisy data from pooling experiments, we show via simulation that the method can produce substantially more accurate estimated allele frequencies at genotyped markers.

4.2 Methods and Models

In this chapter we consider the following form of imputation problem. We assume that data are available on p SNPs in a reference panel of data on m individuals samples from a population, and that a subset of these SNPs are typed on a further study sample of individuals taken from a similar population. The goal is to estimate data at untyped SNPs in the study sample, using the information on the correlations among typed and untyped SNPs that is contained in the reference panel data.

We let $\boldsymbol{M}$ denote the panel data, and $\boldsymbol{h}_1, \ldots, \boldsymbol{h}_{2n}$ denote the $2n$ haplotypes in the study sample. In the simplest case, the panel data will be a $2m \times p$ binary matrix,

and the haplotypes $\boldsymbol{h}_1, \ldots, \boldsymbol{h}_{2n}$ can be thought of as additional rows of this matrix with some missing entries whose values need "imputing". Several papers have focused on this problem of "individual-level" imputation (Servin and Stephens (2008), Scheet and Stephens (2005), Marchini *et al.* (2007), Browning and Browning (2007), Li *et al.* (2006)).

In this chapter, we consider the problem of performing imputation when only summary-level data are available for $\boldsymbol{h}_1, \ldots, \boldsymbol{h}_{2n}$. Specifically, let

$$\boldsymbol{y} = (y_1 \ \ldots \ y_p)' = \frac{1}{2n} \sum_{i=1}^{2n} \boldsymbol{h}_i, \tag{4.1}$$

denote the vector of allele frequencies in the study sample. We assume that these allele frequencies are measured at a subset of *typed* SNPs, and consider the problem of using these measurements, together with information in $\boldsymbol{M}$, to estimate the allele frequencies at the remaining *untyped* SNPs. More formally, if $(\boldsymbol{y}_t, \boldsymbol{y}_u)$ denotes the partition of $\boldsymbol{y}$ into typed and untyped SNPs, our aim is to estimate the conditional distribution $\boldsymbol{y}_u | \boldsymbol{y}_t, \boldsymbol{M}$.

Our approach is based on the assumption that $\boldsymbol{h}_1, \ldots, \boldsymbol{h}_{2n}$ are independent and identically distributed (i.i.d.) draws from some conditional distribution $\Pr(\boldsymbol{h}|\boldsymbol{M})$. Specifically, in common with many existing approaches to individual-level imputation (Stephens and Scheet (2005), Marchini *et al.* (2007), Li *et al.* (2006)), we use the HMM-based conditional distribution from Li and Stephens (2003), although other choices could be considered. It then follows by the central limit theorem, provided that the sample size $2n$ is large, the distribution of $\boldsymbol{y}|\boldsymbol{M}$ can be approximated by a multivariate normal distribution:

$$\boldsymbol{y}|\boldsymbol{M} \sim \mathrm{N}_p(\boldsymbol{\mu}, \Sigma), \tag{4.2}$$

where $\boldsymbol{\mu} = \mathrm{E}(\boldsymbol{h}|\boldsymbol{M})$ and $\Sigma = \frac{1}{2n}\mathrm{Var}(\boldsymbol{h}|\boldsymbol{M})$.

From this joint distribution, the required conditional distribution is easily obtained. Specifically, by partitioning $\boldsymbol{\mu}$ and Σ in the same way as $\boldsymbol{y}$, according to

SNPs' typed/untyped status, (4.2) can be written,

$$\begin{pmatrix} \boldsymbol{y}_u \\ \boldsymbol{y}_t \end{pmatrix} \Big| \boldsymbol{M} \Bigg) \sim \mathrm{N}_p \left(\begin{pmatrix} \boldsymbol{\mu}_u \\ \boldsymbol{\mu}_t \end{pmatrix}, \begin{pmatrix} \Sigma_{uu} & \Sigma_{ut} \\ \Sigma_{tu} & \Sigma_{tt} \end{pmatrix} \right), \tag{4.3}$$

and

$$\boldsymbol{y}_u | \boldsymbol{y}_t, \boldsymbol{M} \sim \mathrm{N}_q(\boldsymbol{\mu}_u + \Sigma_{ut}\Sigma_{tt}^{-1}(\boldsymbol{y}_t - \boldsymbol{\mu}_t), \Sigma_{uu} - \Sigma_{ut}\Sigma_{tt}^{-1}\Sigma_{tu}). \tag{4.4}$$

The mean of this last distribution can be used as a point estimate for the unobserved frequencies $\boldsymbol{y}_u$, while the variance gives an indication of the uncertainty in these estimates. (In principle the mean can lie outside the range $[0, 1)$, in which case we use 0 or 1, as appropriate, as the point estimate; however this happens very rarely in practice).

The parameters $\boldsymbol{\mu}$ and Σ must be estimated from the panel data. It may seem natural to estimate these using the empirical mean $\boldsymbol{f}^{\text{panel}}$ and the empirical covariance matrix Σ^{panel} from the panel. However, Σ^{panel} is highly rank deficient because the sample size m in the panel is far less than the number of SNPs p, and so this empirical matrix cannot be used directly. Use of the conditional distribution from Li and Stephens solves this problem. Indeed, under this conditional distribution $\mathrm{E}(\boldsymbol{h}|\boldsymbol{M}) = \hat{\boldsymbol{\mu}}$ and $\mathrm{Var}(\boldsymbol{h}|\boldsymbol{M}) = \hat{\Sigma}$ can be derived analytically (appendix F) as:

$$\hat{\boldsymbol{\mu}} \quad = (1-\theta)\boldsymbol{f}^{\text{panel}} + \tfrac{\theta}{2}\mathbf{1}, \tag{4.5}$$

$$\hat{\Sigma} \quad = (1-\theta)^2 S + \tfrac{\theta}{2}(1-\tfrac{\theta}{2})I, \tag{4.6}$$

where θ is a parameter relating to mutation, and S is obtained from Σ^{panel} by shrinking off-diagonal entries towards 0. Specifically,

$$S_{ij} = \begin{cases} \Sigma_{ij}^{\text{panel}} & i = j \\ \exp(-\frac{\rho_{ij}}{2m})\Sigma_{ij}^{\text{panel}} & i \neq j \end{cases} \tag{4.7}$$

where ρ_{ij} is an estimate of the population-scaled recombination rate between SNPs i and j (e.g. Hudson (2001), Li and Stephens (2003), McVean *et al.* (2002)). We use

the value of θ suggested by Li and Stephens (2003),

$$\theta = \frac{(\sum_{i=1}^{2m-1} \frac{1}{i})^{-1}}{2m + (\sum_{i=1}^{2m-1} \frac{1}{i})^{-1}}, \tag{4.8}$$

and values of ρ_{ij} obtained by applying the software PHASE (Stephens and Scheet (2005)) to the HapMap CEU data, which are conveniently distributed with the IMPUTE software package. For SNPs i and j that are distant, $\exp(-\frac{\rho_{ij}}{2m}) \approx 0$. To exploit the benefits of sparsity we set any value that was less than 10^{-8} to be 0, which makes $\hat{\Sigma}$ sparse and banded: see Figure 4.1 for illustration. This makes matrix inversion in (4.4) computationally feasible and fast, using standard Gaussian elimination.

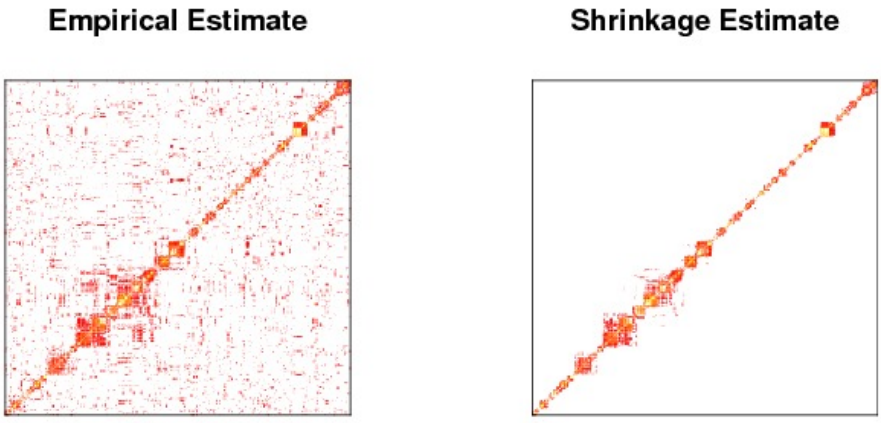

Figure 4.1: Comparison of empirical and shrinkage estimates (based on Li and Stephens Model) of squared correlation matrix from the panel. Both of them are estimated using Hapmap CEU panel with 120 haplotypes. The region plotted is on chromosome 22 and contains 1000 Affymetrix SNPs which cover a 15Mb genomic region. Squared correlation values in $[0.05, 1.00]$ are displayed using R's heat.colors scheme, with gold color representing stronger correlation and red color representing weaker correlation.

4.2.1 Incorporating Measurement Error and Over-dispersion

Our treatment above assumes that the allele frequencies of typed SNPs, $\boldsymbol{y}_t$, are observed without error. In some settings, for example in DNA pooling experiments, this is not the case. We incorporate measurement error by introducing a single parameter ϵ^2, and assume

$$\boldsymbol{y}_t^{\text{obs}} | \boldsymbol{y}_t^{\text{true}} \sim \text{N}_{p-q}(\boldsymbol{y}_t^{\text{true}}, \epsilon^2 I), \tag{4.9}$$

where random vectors $\boldsymbol{y}_t^{\text{obs}}$ and $\boldsymbol{y}_t^{\text{true}}$ represent the observed and true sample allele frequencies for typed SNPs respectively, and subscript $p-q$ denotes the number of typed SNPs. We assume that, given $\boldsymbol{y}_t^{\text{true}}$, the observations $\boldsymbol{y}_t^{\text{obs}}$ are conditionally independent of the panel data ($\boldsymbol{M}$) and the allele frequencies at untyped SNPs ($\boldsymbol{y}_u^{\text{true}}$).

Our treatment in the previous section also makes several other implicit assumptions: for example, that the panel and study individuals are sampled from the same population, and that the parameters ρ and θ are estimated without error. Deviations from these assumptions will cause over-dispersion: the true allele frequencies will lie further from their expected values than the model predicts. To allow for this, we modify (4.2) by introducing an over-dispersion parameter σ^2:

$$\boldsymbol{y}^{\text{true}} | \boldsymbol{M} \sim \text{N}_p(\hat{\boldsymbol{\mu}}, \sigma^2 \hat{\Sigma}). \tag{4.10}$$

Over-dispersion models like this are widely used for modeling binomial data (McCullagh and Nelder (1989)).

In our applications below, for settings involving measurement error (i.e. DNA pooling experiments), we estimate σ^2, ϵ^2 by maximizing the multivariate normal likelihood:

$$\boldsymbol{y}_t^{\text{obs}} | \boldsymbol{M} \sim \text{N}_{p-q}(\hat{\boldsymbol{\mu}}_t, \sigma^2 \hat{\Sigma}_{tt} + \epsilon^2 I). \tag{4.11}$$

For settings without measurement error, we set $\epsilon^2 = 0$ and estimate σ^2 by maximum likelihood.

From the hierarchical model defined by (4.9) and (4.10), the conditional distribu-

tions of allele frequencies at untyped and typed SNPs are given by:

$$\begin{aligned}\boldsymbol{y}_u^{\text{true}}|\boldsymbol{y}_t^{\text{obs}}, \boldsymbol{M} \sim \mathrm{N}_q\Big(&\hat{\boldsymbol{\mu}}_u + \hat{\Sigma}_{ut}(\hat{\Sigma}_{tt} + \frac{\epsilon^2}{\sigma^2}I)^{-1}(\boldsymbol{y}_t^{\text{obs}} - \hat{\boldsymbol{\mu}}_t), \\ &\sigma^2(\hat{\Sigma}_{uu} - \hat{\Sigma}_{ut}(\hat{\Sigma}_{tt} + \frac{\epsilon^2}{\sigma^2}I)^{-1}\hat{\Sigma}_{tu})\ \Big),\end{aligned} \tag{4.12}$$

and

$$\begin{aligned}\boldsymbol{y}_t^{\text{true}}|\boldsymbol{y}_t^{\text{obs}}, \boldsymbol{M} \sim \mathrm{N}_{p-q}\Big(&(\frac{1}{\sigma^2}\hat{\Sigma}_{tt}^{-1} + \frac{1}{\epsilon^2}I)^{-1}(\frac{1}{\sigma^2}\hat{\Sigma}_{tt}^{-1}\hat{\boldsymbol{\mu}}_t + \frac{1}{\epsilon^2}\boldsymbol{y}_t^{\text{obs}}), \\ &(\frac{1}{\sigma^2}\hat{\Sigma}_{tt}^{-1} + \frac{1}{\epsilon^2}I)^{-1}\Big).\end{aligned} \tag{4.13}$$

We use (4.12) to impute allele frequencies at untyped SNPs. In particular, we use the conditional mean

$$\hat{\boldsymbol{y}}_u^{\text{true}} = \hat{\boldsymbol{\mu}}_u + \hat{\Sigma}_{ut}(\hat{\Sigma}_{tt} + \frac{\epsilon^2}{\sigma^2}I)^{-1}(\boldsymbol{y}_t^{\text{obs}} - \hat{\boldsymbol{\mu}}_t), \tag{4.14}$$

as a natural point estimate for these allele frequencies. In settings involving measurement error, we use (4.13) to estimate allele frequencies at typed SNPs, again using the mean

$$\hat{\boldsymbol{y}}_t^{\text{true}} = (\frac{1}{\sigma^2}\hat{\Sigma}_{tt}^{-1} + \frac{1}{\epsilon^2}I)^{-1}(\frac{1}{\sigma^2}\hat{\Sigma}_{tt}^{-1}\hat{\boldsymbol{\mu}}_t + \frac{1}{\epsilon^2}\boldsymbol{y}_t^{\text{obs}}), \tag{4.15}$$

as a point estimate. Note that this mean has an intuitive interpretation as a weighted average of the observed allele frequency at that SNP and information from other nearby, correlated, SNPs. For example, if two SNPs are perfectly correlated, then in the presence of measurement error, the average of the measured frequencies will be a better estimator of the true frequency than either of the single measurements (assuming measurement errors are uncorrelated). The lower the measurement error, ϵ^2, the greater the weight given to the observed frequencies; and when $\epsilon^2 = 0$ the estimated frequencies are just the observed frequencies.

Remark. For both untyped and typed SNPs, our point estimates for allele frequencies , (4.14) and (4.15)) are linear functions of the observed allele frequencies. Although these linear predictors were developed based on an appeal to the Central

Limit Theorem, and resultant normality assumption, there are alternative justifications for use of these particular linear functions that do not rely on normality. Specifically, assuming that the two haplotypes making up each individual are i.i.d draws from a conditional distribution $\Pr(\boldsymbol{h}|\boldsymbol{M})$ with mean $\hat{\boldsymbol{\mu}}$ and variance covariance matrix $\sigma^2\hat{\Sigma}$, then the linear predictors (4.14) and (4.15)) minimize the integrated risk (assuming squared error loss) among all linear predictors (West and Harrison (1997)). In this sense they are the best linear predictors, and so we refer to this method of imputation as Best Linear IMPutation or BLIMP.

4.2.2 Extension to Imputing Genotype Frequencies

The development above considers imputing unobserved allele frequencies. In some settings one might also want to impute genotype frequencies. A simple way to do this is to use an assumption of Hardy–Weinberg equilibrium: that is, to assume that if y is the allele frequency at the untyped SNP, then the three genotypes have frequencies $(1-y)^2, 2y(1-y)$ and y^2. Under our normal model, the expected values of these three quantities can be computed:

$$\begin{aligned}
\mathrm{E}((1-y)^2|\boldsymbol{y}_t^{\mathrm{obs}},\boldsymbol{M}) &= (1-\mathrm{E}(y|\boldsymbol{y}_t^{\mathrm{obs}},\boldsymbol{M}))^2 + \mathrm{Var}(y|\boldsymbol{y}_t^{\mathrm{obs}},\boldsymbol{M}) \\
\mathrm{E}(y^2|\boldsymbol{y}_t^{\mathrm{obs}},\boldsymbol{M}) &= (\mathrm{E}(y|\boldsymbol{y}_t^{\mathrm{obs}},\boldsymbol{M}))^2 + \mathrm{Var}(y|\boldsymbol{y}_t^{\mathrm{obs}},\boldsymbol{M}) \\
\mathrm{E}(2y(1-y)|\boldsymbol{y}_t^{\mathrm{obs}},\boldsymbol{M}) &= 1-\mathrm{E}((1-y)^2|\boldsymbol{y}_t^{\mathrm{obs}},\boldsymbol{M}) - \mathrm{E}(y^2|\boldsymbol{y}_t^{\mathrm{obs}},\boldsymbol{M}),
\end{aligned} \tag{4.16}$$

where $\mathrm{E}(y|\boldsymbol{y}_t^{\mathrm{obs}},\boldsymbol{M})$ and $\mathrm{Var}(y|\boldsymbol{y}_t^{\mathrm{obs}},\boldsymbol{M})$ are given in (4.12). These expectations can be used as estimates of the unobserved genotype frequencies.

The method above uses only *allele* frequency data at typed SNPs. If data are also available on *genotype* frequencies, as might be the case if the data are summary data from a regular genome scan in which all individuals were individually genotypes, then an alternative approach that does not assume HWE is possible. In brief, we write the unobserved genotype frequencies as means of the genotype indicators, $1_{[g=0]}$ and $1_{[g=2]}$ (analogous to expression (4.1)), and then derive expressions for the means and covariances of these indicators both within SNPs and across SNPs. Imputation can then be performed by computing the appropriate conditional distributions using the

joint normal assumption, as in (4.4). See appendix G for more details.

In practice, we have found these two methods give similar average accuracy (results not shown), although this could be because our data conform well to Hardy–Weinberg equilibrium.

4.2.3 Individual-level Genotype Imputation

Although we developed the above model to tackle the imputation problem when individual genotypes are not available, it can also be applied to the problem of individual-level genotype imputation when individual-level data *are* available, by treating each individual as a pool of two haplotypes (application of these methods to small pool sizes is justified by the **Remark** above). For example, doubling (4.14) provides a natural estimate of the posterior mean genotype for an untyped SNP. For many applications this posterior mean genotype may suffice; see Guan and Stephens (2008) for the use of such posterior means in downstream association analyses. If an estimate that takes a value in {0,1,2} is desired, then a simple *ad hoc* procedure that we have found works well in practice is to round the posterior mean to the nearest integer. Alternatively, a full posterior distribution on the three possible genotypes can be computed by using the genotypic version of our approach (appendix G).

4.2.4 Using Unphased Genotype Panel

Our method can be readily adapted to settings where the panel data are unphased. To do this, we note that the estimates (4.5, 4.6) for $\boldsymbol{\mu}$ and Σ depend on the panel data only through the empirical mean and variance covariance matrix of the panel haplotypes. When the panel data are unphased, we simply replace these with 0.5 times the empirical mean and variance covariance matrix of the panel genotypes (since, assuming random mating, genotypes are expected to have twice the mean and twice the (co)variance of haplotypes); see Weir (1979) for related discussion.

4.2.5 Imputation without a Panel

In some settings, it may be desired to impute missing genotypes in a sample where no individuals are typed at all SNPs (i.e. there is no panel $\boldsymbol{M}$), and each individual is typed at a different subset of SNPs. For example, this may arise if many individuals are sequenced at low coverage, as in the currently-ongoing 1000 genomes project (Durbin *et al.* (2010)). In the absence of a panel we cannot directly obtain the mean and variance-covariance estimates $\hat{\boldsymbol{\mu}}$ and $\hat{\Sigma}$ as in (4.5) and (4.6). An alternative way to obtain these estimates is to treat each individual genotype vector as a random sample from multivariate normal distribution $\mathrm{N}_p(\hat{\boldsymbol{\mu}}, \hat{\Sigma})$, and apply the ECM algorithm (Meng and Rubin (1993)) to perform maximum likelihood estimation. However, this approach does not incorporate shrinkage. We therefore modify the algorithm in an *ad-hoc* way to incorporate shrinkage in the conditional maximization step. See Appendix H for details.

4.3 Data Application

We evaluate methods described above by applying them to data from a subset of the WTCCC Birth Cohort, consisting of 1376 unrelated British individuals genotyped on the Affymetrix 500K platform (Wellcome Trust Case Control Consortium (2007)). For demonstration purpose, we use only the 4329 SNPs from chromosome 22. We impute data at these SNPs using the 60 unrelated HapMap CEU parents (The International HapMap Consortium (2005)) as the panel. For the recombination parameters required in (4.7) we use the estimates distributed in the software package IMPUTE v1 (Marchini *et al.* (2007)), which were estimated from the same panel using the software package PHASE (Stephens and Scheet (2005)).

In our evaluations, we consider three types of application: frequency imputation using summary-level data, individual-level genotype imputation, and noise reduction in DNA pooling experiments. We examine both the accuracy of point estimates, and calibration of the credible intervals.

4.3.1 Frequency Imputation using Summary-level Data

In this section, we evaluate the performance of (4.14) for imputing frequencies at untyped SNPs. The observed data consist of the marginal allele frequencies at each SNP, which we compute from the WTCCC individual-level genotype data. To assess imputation accuracy, we perform the following cross-validation procedure: we mask the observed data at every 25th SNP, then treat the remaining SNPs as typed and use them to impute the frequencies of masked SNPs and compare the imputation results with the actual observed frequencies. We repeat this procedure 25 times by shifting the position of the first masked SNP. Because in this case, the observed frequencies are obtained through high quality individual-level genotype data, we assume the experimental error parameter $\epsilon^2 = 0$.

To provide a basis for comparison, we also perform the same experiment using the software package IMPUTE v1 (Marchini *et al.* (2007)), which is among the most accurate of existing methods for this problem. IMPUTE requires individual-level genotype data, and outputs posterior genotype probabilities for each unmeasured genotype. We therefore input the individual-level genotype data to IMPUTE and estimate the allele frequency at each untyped SNP using the posterior expected frequency computed from the posterior genotype probabilities. Like our method, IMPUTE performs imputation using the conditional distribution from Li and Stephens (Li and Stephens (2003)); however, it uses the full conditional distribution whereas our method uses an approximation based on the first two moments. Furthermore, IMPUTE uses individual-level genotype data. For both these reasons, we would expect IMPUTE to be more accurate than our method, and our aim is to assess how much we lose in accuracy by our approximation and by using summary-level data.

To assess accuracy of estimated allele frequencies, we use the Root Mean Squared Error (RMSE),

$$\text{RMSE} = \sqrt{\frac{1}{J}\sum_{j=1}^{J}(y_j - \hat{y}_j)^2}, \tag{4.17}$$

where J is the number of SNPs tested (4329) and $y_j, \hat{y}_j$ are observed and imputed allele frequencies for SNP j respectively.

The RMSE from our method was 0.0157 compared with 0.0154 from IMPUTE (Table 4.1). Thus, for these data, using only summary-level data sacrifices virtually nothing in accuracy of frequency imputation. Furthermore, we found that using an unphased panel (replacing the phased HapMap CEU haplotypes with the corresponding unphased genotypes) resulted in only a very small decrease in imputation accuracy: RMSE = 0.0159. In all cases the methods are substantially more accurate than a "naive method" that simply estimates the sample frequency using the panel frequency (Table 4.1).

We also investigated the calibration of the estimated variances of the imputed frequencies from (4.12). To do this, we constructed a Z-statistic for each test SNP j,

$$Z_j = \frac{y_j - \mathrm{E}(y_j|\boldsymbol{y}^t, \boldsymbol{M})}{\sqrt{\mathrm{Var}(y_j|\boldsymbol{y}^t, \boldsymbol{M})}}, \tag{4.18}$$

where y_j is the true observed frequency, and the conditional mean and variance are as in (4.12). If the variances are well calibrated, the Z-scores should follow a standard normal distribution (with slight dependence among Z-scores of neighboring SNPs due to LD). Figure 4.2a shows that, indeed, the empirical distribution of Z-scores is close to standard normal (results are shown for phased panel; results for unphased panel are similar). Note that the over-dispersion parameter plays a crucial role in achieving this calibration. In particular, the Z-scores produced by the model without over-dispersion (4.2) do not follow a standard normal distribution, with many more observations in the tails (Figure 4.2b) indicating that the variance is under-estimated.

Comparison with Unregularized Linear Frequency Estimator

To assess the role of regularization, we compared the accuracy of BLIMP with simple unregularized linear frequency estimators based on a small number of near-by "predicting" SNPs. (In fact, selecting a small number of SNPs can be viewed as a kind of regularization, but we refer to it as un-regularized for convenience.) The un-regularized linear estimator has the same form as in (4.4), but uses the unregu-

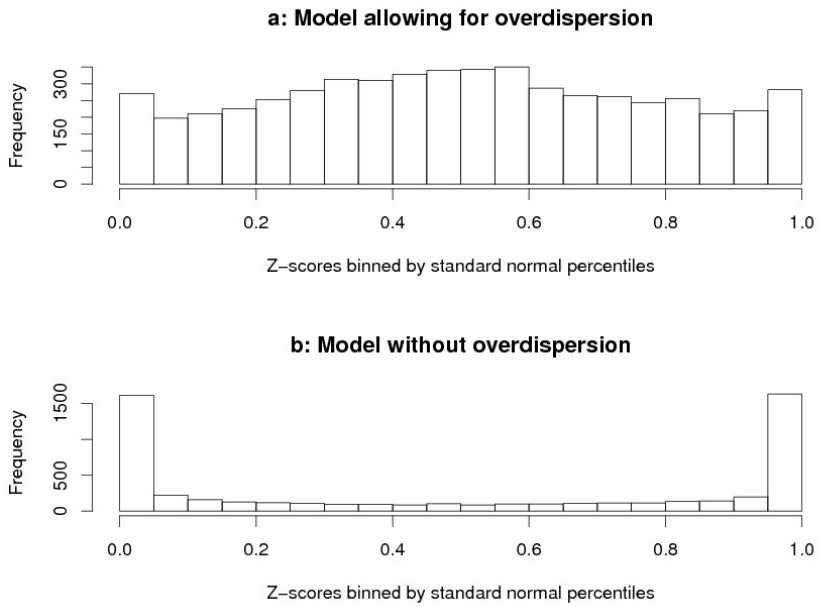

Figure 4.2: Comparison of variance estimation in models with and without over-dispersions. The Z-scores are binned according to the standard normal percentiles, e.g. the first bin (0 to 0.05) contains Z-score values from $-\infty$ to -1.645. If the Z-scores are i.i.d. and strictly follows standard normal distribution, we expect all the bins having approximately equal height.

larized estimates $\boldsymbol{f}^{\text{panel}}$ and Σ^{panel} for $\boldsymbol{\mu}$ and Σ. We consider two schemes to select predictors: the first scheme selects k flanking SNPs on either side of the target SNP (so $2k$ predictors in total); the second scheme selects the $2k$ SNPs with the highest marginal correlation with the target SNP. Figure 4.3 shows RMSE as predicting SNPs increases from 0 to 50. We find that the best performance of the unregularized methods is achieved by the first scheme, with a relatively large number of predicting SNPs (20-40); however its RMSE is larger than that of IMPUTE and BLIMP.

4.3.2 Individual-level Genotype Imputation

Although very satisfactory methods already exist for individual-level genotype imputation, BLIMP has the potential advantage of being extremely fast and low on memory-usage (see computational comparisons, below). We therefore also assessed

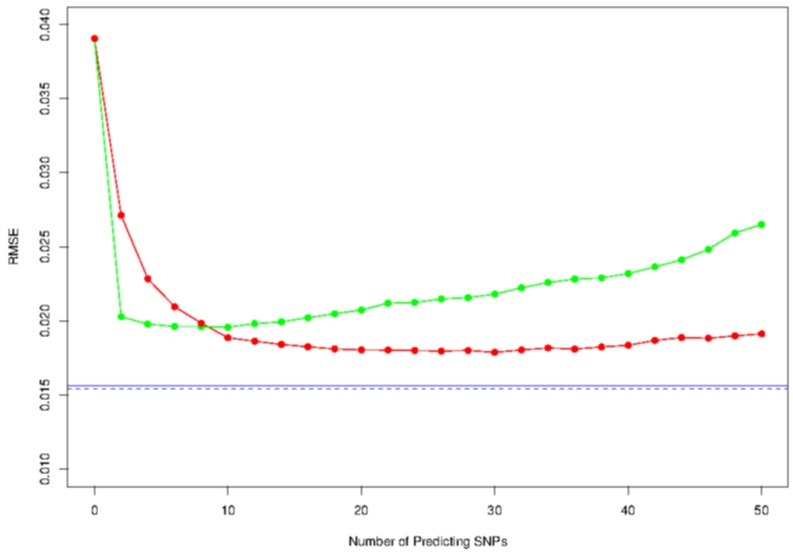

Figure 4.3: Comparison between BLIMP estimator and un-regularized linear estimators. The lines show the RMSE of each allele frequency estimator vs. number of predicting SNPs. Results are shown for two schemes for selecting predicting SNPs: flanking SNPs (red line) and correlated SNPs (green line). Neither scheme is as accurate as BLIMP (blue solid line) or IMPUTE (blue dashed line).

its performance for individual-level genotype imputation. We used the same cross-validation procedure as in frequency imputation, but using individual-level data as input. As above, we compared results from our approach with those obtained using IMPUTE v1.

We again use RMSE to measure accuracy of imputed (posterior mean) genotypes:

$$\text{RMSE} = \sqrt{\frac{1}{mp}\sum_{j=1}^{p}\sum_{i=1}^{m}(g_j^i - \hat{g}_j^i)^2} \tag{4.19}$$

where m is the number of the individuals (1376), p is the total number of tested SNPs (4329) and g_j^i, $\hat{g}_j^i$ are observed and estimated (posterior mean) genotypes for individual i at SNP j respectively.

For comparison purpose, we also use a different measure of accuracy that is commonly used in this setting: the genotype error rate, which is the number of wrongly

imputed genotypes divide by the total number of imputed genotypes. To minimize the expected value of this metric, one should use the posterior mode genotype as the estimated genotype. Thus, for IMPUTE v1 we used the posterior mode genotype for this assessments with this metric. However, for simplicity, for our approach we used the posterior mean genotype rounded to the nearest integer. (Obtaining posterior distributions on genotypes using our approach, as outlined in appendix G, is considerably more complicated, and in fact produced slightly less accurate results, not shown).

We found that, under either metric BLIMP provides only very slightly less accurate genotype imputations than IMPUTE (Table 4.1). Further, as before, replacing the phased panel with an unphased panel produces only a small decrease in accuracy (Table 4.1).

Frequency imputation

	RMSE	Error Rate
naive method	0.0397	NA
BLIMP (phased panel)	0.0157	NA
BLIMP (unphased panel)	0.0159	NA
IMPUTE	0.0154	NA

Individual genotype imputation

	RMSE	Error Rate
BLIMP (phased panel)	0.2339	6.46%
BLIMP (unphased panel)	0.2407	6.77%
IMPUTE	0.2303	6.30%

Table 4.1: Comparison of accuracy of BLIMP and IMPUTE for frequency and individual-level genotype imputations. The RMSE and Error rate, defined in the text, provide different metrics for assessing accuracy; in all cases BLIMP was very slightly less accurate than IMPUTE. The "naive method" refers to the strategy of estimating the sample frequency of each untyped SNP by its observed frequency in the panel; this ignores information in the observed sample data, and provides a baseline level of accuracy against which the other methods can be compared.

These results show average accuracy when all untyped SNPs are imputed. However, it has been observed previously (e.g. Marchini *et al.* (2007), Guan and Stephens (2008)) that accuracy of calls at the most confident SNPs tends to be considerably

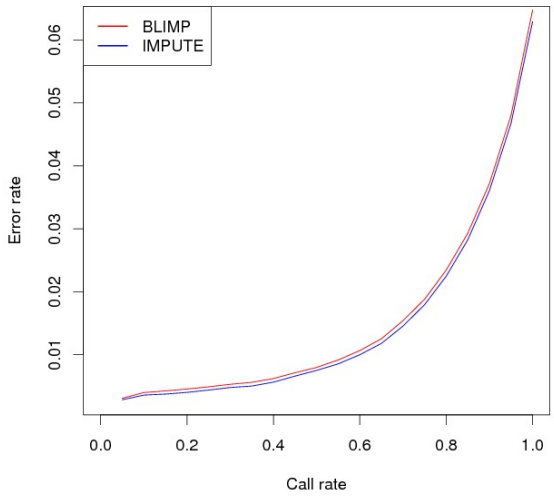

Figure 4.4: Controlling individual-level genotype imputation error rate on a per-SNP basis. For BLIMP, the error rate is controlled by thresholding on the estimated variance for imputed SNP frequencies; for IMPUTE the call threshold is determined by average maximum posterior probability.

higher than the average. We checked that this is also true for BLIMP. To obtain estimates of the confidence of imputations at each SNP we first estimated σ by maximum likelihood using the summary data across all individuals, and then compute the variance for each SNP using (4.12); note that this variance does not depend on the individual, only on the SNP. We then considered performing imputation only at SNPs whose variance was less than some threshold, plotting the proportion of SNPs imputed ("call rate") against their average genotype error rate as this threshold varies. The resulting curve for BLIMP is almost identical to the corresponding curve for IMPUTE (Figure 4.4).

4.3.3 Individual-level Genotype Imputation without a Panel

We use the same WTCCC Birth Cohort data to assess our modified ECM algorithm for performing individual-level genotype imputation without using a panel. To create a data set with data missing at random we mask each genotype, independently, with

probability m. We create multiple data sets by varying m from 5% to 50%. For each data set we run our ECM algorithm for 20 iterations. (Results using different starting points for the ECM algorithm were generally very consistent, and so results here are shown for a single starting point.)

We compare the imputation accuracy with the software package BIMBAM (Guan and Stephens (2008)) which implements the algorithms from Scheet and Stephens (2005).

BIMBAM requires the user to specify a number of "clusters", and other parameters related to the EM algorithm it uses: after experimenting with different settings we applied BIMBAM on each data set assuming 20 clusters, with 10 different EM starting points, performing 20 iterations for each EM run. (These settings produced more accurate results than shorter runs.)

Overall, imputation accuracy of the two methods was similar (Table 4.2), with BLIMP being slightly more accurate with larger amounts of missing data and BIMBAM being slightly more accurate for smaller amounts of missing data.

We note that in this setting, some of the key computational advantages of our method are lost. In particular, when each individual is missing genotypes at different SNPs, one must effectively invert a different covariance matrix for each individual. Furthermore, this inversion has to be performed multiple times, due to the iterative scheme. For small amounts of missing data the results from BLIMP we present here took less time than the results for BIMBAM, but for larger amounts the run times are similar.

	Missing Rate			
	5%	10%	20%	50%
BIMBAM	5.79%	6.35%	7.15%	9.95%
BLIMP ECM	6.07%	6.49%	7.31%	9.91%

Table 4.2: Comparison of imputation error rates from BLIMP and BIMBAM for individual genotype imputation without a panel.

4.3.4 Noise Reduction in Pooled Experiment

We used simulations to assess the potential for our approach to improve allele frequency estimates from noisy data in DNA pooling experiments (equation (4.13)). To generate noisy observed data we took allele frequencies of 4329 genotyped SNP from the WTCCC Birth Cohort chromosome 22 data as true values, and added independent and identically distributed $N(0, \epsilon^2)$ noise terms to each true allele frequency. Real pooling data will have additional features not captured by these simple simulations (e.g. biases towards one of the alleles), but our aim here is simply to illustrate the potential for methods like ours to reduce noise in this type of setting. We varied ϵ from $0.01 - 0.18$ to to examine different noise levels. Actual noise levels in pooling experiments will depend on technology and experimental protocol; to give a concrete example, Meaburn *et al.* (2006) found differences between allele frequency estimates from pooled genotyping and individual genotyping of the same individuals, at 26 SNPs, in the range 0.008 to 0.077 (mean 0.036).

We applied our method to the simulated data by first estimating σ and ϵ using (4.12), and then, conditional on these estimated parameters, estimating the allele frequency at each observed SNP using the posterior mean given in equation (4.13). We assessed the accuracy (RMSE) of these allele frequency estimates by comparing them with the known true values.

We found that our method was able to reliably estimate the amount of noise present in the data: the estimated values for the error parameter ϵ show good correspondence with the standard deviation used to simulate the data (Figure 4.5a), although for high noise levels we underestimate the noise because some of the errors are absorbed by the parameter σ.

More importantly, we found our estimated allele frequency estimates were consistently more accurate than the direct (noisy) observations, with the improvement being greatest for higher noise levels (Figure 4.5b). For example, with $\epsilon = 0.05$ our method reduced the RMSE by more than half, to 0.024.

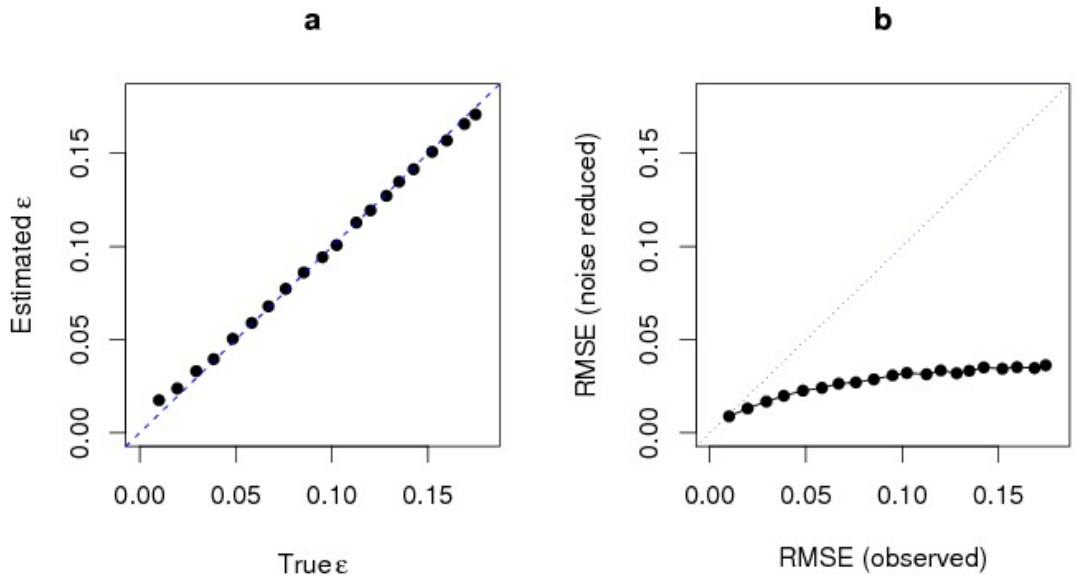

Figure 4.5: **a**. Detection of experimental noise in simulated data. The simulated data sets are generated by adding Gaussian noise $N(0, \epsilon^2)$ to the actual observed WTCCC frequencies. The estimated ϵ values are plotted against the true ϵ values used for simulation. We estimate ϵ using maximum likelihood by (4.11). **b**. An illustration on the effect of noise reduction in varies noise levels. RMSE from noise reduced estimates are plotted against RMSE from direct noisy observations. The noise reduced frequency estimates are posterior means obtained from model (4.13).

4.3.5 Computational Efficiency

Imputation using our implementation of BLIMP is highly computationally efficient. The computational efficiency is especially notable when dealing with large panels: the panel is used only to estimate $\hat{\boldsymbol{\mu}}$ and $\hat{\Sigma}$, which is both quick and done just once, after which imputation computations do not depend on panel size. Our implementations also take advantage of the sparsity of $\hat{\Sigma}$ to increase running speed and reduce memory usage. To give a concrete indication of running speed, we applied BLIMP to the WTCCC Birth Cohort data on chromosome 22, containing 4,329 genotyped and 29,697 untyped Hapmap SNPs on 1376 individuals, using a Linux system with eight-core Intel Xeon 2.66GHz processors (although only one processor is used by our implementation). The running time is measured by 'real' time reported by the Unix

"time" command. For frequency imputation, BLIMP took 9 minutes and 34 seconds, with peak memory usage of 162 megabytes; for individual-level genotype imputation BLIMP took 25 minutes, using under 300 megabytes of memory. As a comparison, IMPUTE v1 took 195 minutes for individual-level genotype imputation, with memory usage exceeding 5.1 gigabytes.

Since these comparisons were done we note that a new version of IMPUTE (v2) has been released (Howie *et al.* (2009)). This new version gives similar imputation accuracy in the settings we described above; it runs more slowly than v1 but requires less memory.

4.4 Conclusion and Discussion

Imputation has recently emerged as an important and powerful tool in genetic association studies. In this chapter, we propose a set of statistical tools that help solve the following problems:

1. Imputation of allele frequencies at untyped SNPs when only summary-level data are available at typed SNPs.

2. Noise reduction for estimating allele frequencies from DNA pooling-based experiments.

3. Fast and accurate individual-level genotype imputation.

The proposed methods are simple, yet statistically elegant, and computationally extremely efficient. For individual-level genotype imputation the imputed genotypes from this approach are only very slightly less accurate than state-of-the-art methods. When only summary-level data are available we found that imputed allele frequencies were almost as accurate as when using full individual genotype data.

The linear predictor approach to imputation requires only an estimate of the mean and the covariance matrix among SNPs. Our approach to obtaining these estimates is based on the conditional distribution from Li and Stephens (2003); however, it would certainly be possible to consider other estimates, and specifically to use other

approaches to shrink the off-diagonal terms in the covariance matrix. An alternative, closely-related, approach is to obtain a linear predictor directly by training a linear regression to predict each SNP, using the panel as a training set, and employing some kind of regularization scheme to solve potential problems with over-fitting caused by large p small n. This approach has been used in the context of individual-level genotype imputation by A. Clark (personal communication), and Yu and Schaid (2007). However, the choice of appropriate regularization is not necessarily straightforward, and different regularization schemes can provide different results (Yu and Schaid (2007)). Our approach of regularizing the covariance matrix using the conditional distribution from Li and Stephens (2003) has the appeal that this conditional distribution has already been shown to be very effective for individual genotype imputation, and for modeling patterns of correlation among SNPs more generally (Li and Stephens (2003), Stephens and Scheet (2005), Servin and Stephens (2008), Marchini *et al.* (2007)). Furthermore, the fact that, empirically, BLIMP's accuracy is almost as good as the best available purpose-built methods for this problem suggests that alternative approaches to regularization are unlikely to yield considerable improvements in accuracy.

The accuracy with which linear combinations of typed SNPs can predict untyped SNPs is perhaps somewhat surprising. That said, theoretical arguments for the use of linear combinations have been given in previous work. For example, Clayton *et al.* (2004) showed by example that, when SNP data are consistent with no recombination (as might be the case for markers very close together on the genome), each SNP can be written as a linear regression on the other SNPs. Conversely, it is easy to construct hypothetical examples where linear predictors would fail badly. For example, consider the following example from Nicolae (2006a): 3 SNPs form 4 haplotypes, 111, 001, 100 and 010, each at frequency 0.25 in a population. Here the correlation between every pair of SNPs is 0, but knowing any 2 SNPs is sufficient to predict the third SNP precisely. Linear predictors cannot capture this "higher order" interaction information, so produce sub-optimal imputations in this situation. In contrast, other methods (including IMPUTE) could use the higher-order information to produce perfect imputations. The fact that, empirically, the linear predictor works well

suggests that this kind of situation is rare in real human population genotype data. Indeed, this is not so surprising when one considers that, from population genetics theory, SNPs tend to be uncorrelated only when there is sufficiently high recombination rate between them, and recombination will tend to break down any higher-order correlations as well as pairwise correlations.

Besides HMM-based methods, another type of approach to genotype imputation that has been proposed is to use "multi-marker" tagging (de Bakker *et al.* (2005), Purcell *et al.* (2007), Nicolae (2006b)). A common feature of these methods is to preselect a (relatively small) *subset* of "tagging" SNPs or haplotypes from all typed SNPs based on some LD measure threshold, and then use a possibly non-linear approach to predicting untyped SNPs from this subset. Thus, compared with our approach, these methods generally use a more complex prediction method based on a smaller number of SNPs. Although we have not compared directly with these methods here, published comparisons (Howie *et al.* (2009)) suggest that they are generally noticeably less accurate than HMM-based methods like IMPUTE, and thus by implication less accurate than BLIMP. That is, it seems from these results that, in terms of average accuracy, it is more important to make effective use of low-order correlations from all available SNPs that are correlated with the target untyped SNP, than to take account of unusual higher-order correlations that may occasionally exist.

Our focus here has been on the accuracy with which untyped SNP allele frequencies can be imputed. In practice an important application of these imputation methods is to test untyped alleles for association with an outcome variable (e.g. case-control status). Because our allele frequency predictors are linear combinations of typed SNP frequencies, each test of an untyped SNP is essentially a test for differences between a given linear combination of typed SNPs in case and control groups. Several approaches to this are possible; for example the approach in Nicolae (2006b) could be readily applied in this setting. The resulting test would be similar to the test suggested in Homer *et al.* (2008a) which also uses a linear combination of allele frequencies at typed SNPs to test untyped SNPs for association with case-control status in a pooling context. The main difference is that their proposed linear combinations are *ad hoc*, rather than being chosen to be the best linear imputations; as such we

expect that appropriate use of our linear imputations should result in more powerful tests, although a demonstration of this lies outside the scope of this dissertation.

4.5 Acknowledgments

We thank Yongtao Guan and Bryan Howie for helpful discussions. This work was supported by NIH grants HG02585 and HL084689.

CHAPTER 5
CONCLUSIONS

In this dissertation research, we are motivated by an important problem in statistical genetics: the heterogeneity of genetic effects. In chapter 2, we show a Bayesian framework that systematically deals with potentially-heterogeneous genetic data. In particular, we separate the problems of identifying genetic variants that are associated with a phenotype of interest and investigating the heterogeneities of identified genetic variants. In chapter 3, we apply these methods in an important genomics application of mapping tissue-specific eQTLs and develop a hierarchical approach to efficiently pooling information across many simultaneously measured genes. In chapter 4, we propose a linear imputation method for handling missing data in meta-analysis of genetic association studies. This method is unique in the sense that it can perform accurate imputation when only summary statistics are available.

One of the biggest challenges in statistical genetics research is to formulate a scientific question into an appropriate statistical question. In identifying genetic variants with potential gene-environment interaction, we find it is very helpful to re-frame the problem as two related inquires: first, is a particular genetic variant associated with the phenotype of interest at all? If yes, how the association behaves in different environmental conditions. We naturally translate the first question into a hypothesis testing problem and address the second question in a model comparison framework. It is a consequence of this formulation that we choose to solve the problem in a Bayesian framework: Bayesian model comparison is natural and convenient to assess models in a discrete model space (and the hypothesis testing can be regarded as a special case in this framework).

The size and scale of modern date genetic/genomic data are fast-growing and has reached to a level that computational feasibility has become critical in many applications. Computational efficiency is an important consideration for us in solving almost all the problems addressed in this dissertation. To ease computational burden, we adopt strategies from two distinct directions:

1. Performing approximate numerical computations based on the exact model.

2. Simplify the complicated models by imposing stronger assumptions.

The example for the first strategy is our use of Laplace approximation in computing complicated Bayes Factors; the normal approximation of allele frequency distribution. The "one *cis*-eQTL per gene" assumption in the hierarchical model of mapping tissue-specific eQTLs serve as the examples for the second strategy. However, it is also important to investigate the effects of these approximations and simplifications, especially their impacts on interpreting the final numerical results, and improve the methods based on the findings.

APPENDIX A
COMPUTING BAYES FACTORS

In this section, we show the detailed calculation of various Bayes Factors.

A.1 Computation in the ES Model

A particular ES model, describing an alternative hypothesis H_a, is fully specified by setting values for (ϕ, ω) and hyper-parameters $(v_1, \ldots, v_S, l_1, m_1, \ldots, l_S, m_S)$. Under the contrasting null model H_0, we set $\phi = \omega = 0$ while keeping other hyper-parameter the same.

Let $\boldsymbol{\beta}_s = (\mu_s, \beta_s)$, $\tau_s = \sigma_s^{-2}$ and $\boldsymbol{\theta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_s, \tau_1, \ldots, \tau_s, \bar{b})$, the marginal likelihood under model H_a can be written as

$$\begin{aligned}
P(\boldsymbol{Y}|\boldsymbol{G}, H_a) &= \int P(\boldsymbol{Y}|\boldsymbol{G}, \boldsymbol{\theta}, H_a) p(\boldsymbol{\theta}|H_a) d\boldsymbol{\theta} \\
&= \int \left(\prod_s P(\boldsymbol{y}_s|\boldsymbol{g}_s, \boldsymbol{\beta}_s, \tau_s) \prod_s P(\boldsymbol{\beta}_s|\tau_s, \bar{b}, H_a) \prod_s P(\tau_s|H_a) P(\bar{b}|H_a) \right) d\boldsymbol{\beta}_1 \cdots d\boldsymbol{\beta}_S d\tau_1 \cdots d\tau_S d\bar{b} \\
&= \int \left(\int \left(\prod_s \int P(\boldsymbol{y}_s|\boldsymbol{g}_s, \boldsymbol{\beta}_s, \tau_s) P(\boldsymbol{\beta}_s|\tau_s, \bar{b}, H_a) d\boldsymbol{\beta}_s \right) p(\bar{b}|H_a) d\bar{b} \right) \prod_s P(\tau_s|H_a) d\tau_1 \cdots d\tau_S
\end{aligned} \tag{A.1}$$

Let $\boldsymbol{X}_s = (\mathbf{1}\ \boldsymbol{g}_s)$ denote the design matrix of regression model (2.1) for subgroup s, it follows that

$$\begin{aligned}
P(\boldsymbol{y}_s|\boldsymbol{g}_s, \boldsymbol{\beta}_s, \tau_s) &= (\frac{2\pi}{\tau_s})^{-n_s/2} \exp\left(-\frac{\tau_s}{2} (\boldsymbol{y}_s - \boldsymbol{X}_s \boldsymbol{\beta}_s)' (\boldsymbol{y}_s - \boldsymbol{X}_s \boldsymbol{\beta}_s) \right) \\
&= (\frac{2\pi}{\tau_s})^{-n_s/2} \exp\left(-\frac{1}{2} (\tilde{\boldsymbol{y}}_s - \boldsymbol{X}_s \boldsymbol{b}_s)' (\tilde{\boldsymbol{y}}_s - \boldsymbol{X}_s \boldsymbol{b}_s) \right),
\end{aligned} \tag{A.2}$$

where $\tilde{\boldsymbol{y}}_s = \sqrt{\tau_s} \boldsymbol{y}_s$ and $\boldsymbol{b}_s = \sqrt{\tau_s} \boldsymbol{\beta}_s = (\sqrt{\tau_s}\mu_s, b_s)$. We further denote

$$\bar{\boldsymbol{b}} = \begin{pmatrix} 0 \\ \bar{b} \end{pmatrix} \quad \text{and} \quad \Phi_s = \begin{pmatrix} v_s^2 & 0 \\ 0 & \phi^2 \end{pmatrix}, \tag{A.3}$$

and write prior distribution $P(\boldsymbol{b}_s|\bar{\boldsymbol{b}}, H_a)$ in following matrix form,

$$\boldsymbol{b}_s|\bar{\boldsymbol{b}}, H_a \sim \mathrm{N}(\bar{\boldsymbol{b}}, \Phi_s). \tag{A.4}$$

We compute the marginal likelihood by sequentially evaluating the following integrals,

$$\begin{aligned} F_{H_a,s} &= \int P(\boldsymbol{y}_s|\boldsymbol{X}_s, \boldsymbol{b}_s, \tau_s) P(\boldsymbol{b}_s|\bar{\boldsymbol{b}}, H_a) d\boldsymbol{b}_s \\ &= (\frac{2\pi}{\tau_s})^{-n_s/2} |\Phi_s|^{-\frac{1}{2}} \cdot |\boldsymbol{X}'_s\boldsymbol{X}_s + \Phi_s^{-1}|^{-\frac{1}{2}} \\ &\cdot \exp\left(-\frac{1}{2}\left(\tilde{\boldsymbol{y}}'_s\tilde{\boldsymbol{y}}_s - (\boldsymbol{X}'_s\tilde{\boldsymbol{y}}_s + \Phi_s^{-1}\bar{\boldsymbol{b}})'(\boldsymbol{X}'_s\boldsymbol{X}_s + \Phi_s^{-1})^{-1}(\boldsymbol{X}'_s\tilde{\boldsymbol{y}}_s + \Phi_s^{-1}\bar{\boldsymbol{b}}) + \bar{\boldsymbol{b}}'\Phi_s^{-1}\bar{\boldsymbol{b}}\right)\right). \end{aligned} \tag{A.5}$$

Let $J_{H_a} = \int (\prod_s F_{H_a,s})\ P(\bar{b}|H_a) d\bar{b}$; this quantity is also analytically computable by straightforward algebra.

To compute Bayes Factor of H_a versus H_0 under the ES model, we take limits with respect to hyper-parameters $(v_1, \ldots, v_S, l_1, m_1, \ldots, l_S, m_S)$ according to (2.13), that is,

$$\begin{aligned} \mathrm{BF}^{\mathrm{ES}}(\phi, \omega) &= \lim \frac{\int J_{H_a} \prod_s P(\tau_s) d\tau_1 \cdots d\tau_S}{\int J_{H_0} \prod_s P(\tau_s) d\tau_1 \cdots d\tau_S} \\ &= \frac{\int K_{H_a} d\tau_1 \cdots d\tau_S}{\int K_{H_0} d\tau_1 \cdots d\tau_S}. \end{aligned} \tag{A.6}$$

Let us denote

$$\mathrm{RSS}_{0,s} = \boldsymbol{y}'_s\boldsymbol{y}_s - n_s\bar{y}_s^2, \tag{A.7}$$

$$\mathrm{RSS}_{1,s} = \boldsymbol{y}'_s\boldsymbol{y}_s - \boldsymbol{y}'_s\boldsymbol{X}_s(\boldsymbol{X}'_s\boldsymbol{X}_s)^{-1}\boldsymbol{X}'_s\boldsymbol{y}_s, \tag{A.8}$$

$$\delta_s^2 = \frac{1}{\boldsymbol{g}'_s\boldsymbol{g}_s - n_s\bar{g}_s^2}, \tag{A.9}$$

$$\hat{\beta}_s = \frac{\boldsymbol{y}'_s\boldsymbol{g}_s - n_s\bar{y}_s\bar{g}_s}{\boldsymbol{g}'_s\boldsymbol{g}_s - n_s\bar{g}_s^2}, \tag{A.10}$$

$$\zeta^2 = \frac{1}{\sum_s(\delta_s^2 + \phi^2)^{-1}}, \tag{A.11}$$

where $\bar{y}_s$ and $\bar{g}_s$ are the sample means of phenotypes and genotypes in subgroup s.

It can be shown that,

$$K_{H_0} = \prod_s \tau_s^{\frac{n_s}{2}-1} \exp\left(-\frac{1}{2}\sum_s \tau_s \cdot \mathrm{RSS}_{0,s}\right), \tag{A.12}$$

and

$$\begin{aligned} K_{H_a} =& \sqrt{\frac{\zeta^2}{\zeta^2+\omega^2}} \prod_s \sqrt{\frac{\delta_s^2}{\delta_s^2+\phi^2}} \\ & \cdot \prod_s \tau_s^{\frac{n_s}{2}-1} \exp\left(-\frac{1}{2}\sum_s \tau_s \left(\frac{\phi^2}{\delta_s^2+\phi^2}\cdot \mathrm{RSS}_{1,s} + \frac{\delta_s^2}{\delta_s^2+\phi^2}\cdot \mathrm{RSS}_{0,s}\right)\right) \\ & \cdot \exp\left(\frac{1}{2}\frac{\omega^2\zeta^2}{\zeta^2+\omega^2}\left(\sum_s \frac{\hat{\beta}_s\sqrt{\tau_s}}{\delta_s^2+\phi^2}\right)^2\right). \end{aligned} \tag{A.13}$$

The multidimensional integral $\int K_{H_a} d\tau_1 \cdots d\tau_S$ generally does not have a simple analytic form (although it can be represented as finite sums of complicated hypergeometric functions). Next, we show two different approximations, both based on Laplace's method, to evaluate this integral. The first approximation is a direct application of Butler and Wood (2002) and the second one yields a simple analytic expression. To compute the Bayes Factor, we also use Laplace's method to evaluate the integral $\int K_{H_0} d\tau_1 \cdots d\tau_S$. Although this integral can be computed analytically, we find the approximate form by Laplace's method (which essentially uses Stirling's formula to approximate a gamma function) yields more accurate result for the final Bayes Factor: in particular, when there is only one subgroup ($S = 1$, where the Bayes Factor can be analytically computed), we obtain the exact result by applying the first approximation recipe.

Laplace's method approximates a multivariate integral in the following way,

$$\int_D h(\boldsymbol{\tau}) e^{g(\boldsymbol{\tau})} d\boldsymbol{\tau} \approx (2\pi)^{S/2} |H_{\hat{\boldsymbol{\tau}}}|^{-1/2} h(\hat{\boldsymbol{\tau}}) e^{g(\hat{\boldsymbol{\tau}})} \tag{A.14}$$

where $\boldsymbol{\tau}$ is an S-vector,

$$\hat{\boldsymbol{\tau}} = \arg\max_{\boldsymbol{\tau}} g(\boldsymbol{\tau}), \tag{A.15}$$

and $|H_{\hat{\boldsymbol{\tau}}}|$ is the absolute value of the determinant of the Hessian matrix of the function g evaluated at $\hat{\boldsymbol{\tau}}$. Note that the factorization of the integrand is rather arbitrary, it only requires that function h is smooth and positively valued and the smooth function g has a unique maximum lying in the interior of D (for detailed discussion, see Butler (2007)).

Our first approach to apply Laplace's method sets $h(\boldsymbol{\tau}) \equiv 1$ and $g(\boldsymbol{\tau})$ equaling K_{H_a} and K_{H_0} respectively. Except for some trivial situations (e.g. $S = 1$), the maximization of K_{H_a} with respect to $\boldsymbol{\tau}$ is analytically intractable. In practice, we use the Broyden-Fletcher-Goldfarb-Shanno (BFGS2) algorithm, a gradient-based numerical optimization routine (implemented in the GNU Scientific Library), to numerically maximize g. This procedure leads to $\widehat{\mathrm{BF}}^{\mathrm{ES}}(\phi, \omega)$.

Alternatively, we apply Laplace's method by factoring the integrand in such a way that g can be analytically maximized. This approach results in a closed-form approximation. More specifically, we factor K_{H_a} into

$$K_{H_a} = h(\tau_1, \ldots, \tau_S) e^{g(\tau_1, \ldots, \tau_S)}, \tag{A.16}$$

where

$$\begin{aligned} h(\tau_1, \ldots, \tau_S) =& \sqrt{\frac{\zeta^2}{\zeta^2 + \omega^2}} \prod_s \sqrt{\frac{\delta_s^2}{\delta_s^2 + \phi^2}} \\ & \cdot \prod_s \tau_s^{\frac{n_s}{2}-1} \exp\left(-\frac{1}{2} \sum_s \frac{\delta_s^2}{\delta_s^2 + \phi^2} \cdot (\mathrm{RSS}_{0,s} - \mathrm{RSS}_{1,s}) \right) \\ & \cdot \exp\left(\frac{1}{2} \frac{\omega^2 \zeta^2}{\zeta^2 + \omega^2} \left(\sum_s \frac{\hat{\beta}_s \sqrt{\tau_s}}{\delta_s^2 + \phi^2} \right)^2 \right) \end{aligned} \tag{A.17}$$

and

$$e^{g(\tau_1, \ldots, \tau_S)} = \prod_s \tau_s^{\frac{n_s}{2}-1} \cdot \exp\left(-\frac{1}{2} \sum_s \tau_s \cdot \mathrm{RSS}_{1,s} \right). \tag{A.18}$$

It is straightforward to show that the unique maximum of $g(\tau_1, \ldots, \tau_S)$ is attained at

$$\hat{\tau}_s = \frac{n_s - 2}{\mathrm{RSS}_{1,s}}, \ s = 1, \ldots, S, \tag{A.19}$$

which coincides with the REML estimate of τ_s in subgroup-level regression model (2.1). Following the notations in section 2.2.4 and noting the relationship between t and F statistics in simple linear regression,

$$T_s^2 = \frac{\mathrm{RSS}_{0,s} - \mathrm{RSS}_{1,s}}{\mathrm{RSS}_{1,s}/(n_s - 2)}. \tag{A.20}$$

Applying (A.14) results in

$$\mathrm{BF}^{\mathrm{ES}}(\phi, \omega) \simeq \sqrt{\frac{\zeta^2}{\zeta^2 + \omega^2}} \exp\left(\frac{\mathcal{T}_{\mathrm{es}}^2}{2} \frac{\omega^2}{\zeta^2 + \omega^2}\right) \tag{A.21}$$

$$\cdot \prod_s \left(\sqrt{\frac{\delta_s^2}{\delta_s^2 + \phi^2}} \left(\frac{\mathrm{RSS}_{0,s}}{\mathrm{RSS}_{1,s}}\right)^{\frac{n_s}{2}} \exp\left(-\frac{T_s^2}{2} \frac{\delta_s^2}{\delta_s^2 + \phi^2}\right)\right). \tag{A.22}$$

To further simplify the above expression, we use

$$\left(\frac{\mathrm{RSS}_{0,s}}{\mathrm{RSS}_{1,s}}\right)^{n_s/2} = \left(1 + \frac{T_s^2}{n_s - 2}\right)^{\frac{n_s}{2}} = e^{\frac{T_s^2}{2}} \left(1 + O\left(\frac{1}{n_s}\right)\right), \tag{A.23}$$

and (A.21) simplifies to

$$\mathrm{ABF}^{\mathrm{ES}}(\phi, \omega) = \sqrt{\frac{\zeta^2}{\zeta^2 + \omega^2}} \exp\left(\frac{\mathcal{T}_{\mathrm{es}}^2}{2} \frac{\omega^2}{\zeta^2 + \omega^2}\right) \prod_s \left(\sqrt{\frac{\delta_s^2}{\delta_s^2 + \phi^2}} \exp\left(\frac{T_s^2}{2} \frac{\phi^2}{\delta_s^2 + \phi^2}\right)\right). \tag{A.24}$$

Remark. Note, in case $\tau_1, \ldots, \tau_S$ are known, we can directly compute the exact Bayes Factor using

$$\mathrm{BF}^{\mathrm{ES}}(\phi, \omega) = \lim \frac{J_{H_a}}{J_{H_0}} \tag{A.25}$$

without evaluating the multi-dimensional integrals in (A.6). In this particular case, it is easy to show that the exact Bayes Factor has the exact functional form as in (2.23), only with all the $\hat{\tau}_s$'s replaced with the corresponding true values of τ_s's.

Finally, we give the proof for Proposition 1:

Proof of Proposition 1. The derivation above serves as a proof. An alternative proof can be obtained by noting that the REML esitmate of $\hat{\boldsymbol{\tau}}$ asymptotically converges to

the true value of $\boldsymbol{\tau}$ with probablity 1. From the remark above, by applying continuous mapping theorem, we conclude that $\mathrm{ABF}^{\mathrm{ES}}(\phi, \omega)$ converges to $\mathrm{BF}^{\mathrm{ES}}(\phi, \omega)$ with probability 1. □

A.2 Computation in the EE Model

The procedure for computing Bayes Factor assuming an EE model is essentially the same, we omit repeating the details but only show the final results of Bayes Factor of an EE model H_b, specified by (ψ, w), versus the null model H_0,

$$\mathrm{BF}^{\mathrm{EE}}(\psi, w) = \frac{\int K_{H_b} d\tau_1 \cdots d\tau_S}{\int K_{H_0} d\tau_1 \cdots d\tau_S}. \tag{A.26}$$

The expression of K_{H_0} remains the same as (A.12). We denote

$$\eta^2 = \left(\sum_s \frac{\tau_s}{\delta_s^2 + \tau_s \psi^2} \right)^{-1}. \tag{A.27}$$

It can be shown

$$\begin{aligned} K_{H_b} =& \sqrt{\frac{\eta^2}{\eta^2 + w^2}} \prod_s \sqrt{\frac{\delta_s^2}{\delta_s^2 + \tau_s \psi^2}} \\ & \cdot \prod_s \tau_s^{\frac{n_s}{2} - 1} \exp\left(-\frac{1}{2} \sum_s \tau_s \left(\frac{\tau_s \psi^2}{\delta_s^2 + \tau_s \psi^2} \cdot \mathrm{RSS}_{1,s} + \frac{\delta_s^2}{\delta_s^2 + \tau_s \psi^2} \cdot \mathrm{RSS}_{0,s} \right) \right) \\ & \cdot \exp\left(\frac{1}{2} \frac{w^2}{\eta^2 + w^2} \frac{\left(\sum_s \frac{\tau_s}{\delta_s^2 + \tau_s \psi^2} \cdot \hat{\beta}_s \right)^2}{\eta^2} \right). \end{aligned} \tag{A.28}$$

We use the similar numerical procedure to obtain $\widehat{\mathrm{BF}}^{\mathrm{EE}}(\psi, w)$ as in the ES model. To derive $\mathrm{ABF}^{\mathrm{EE}}$, we factor K_{H_b} into

$$K_{H_b} = h(\tau_1, \ldots, \tau_S) e^{g(\tau_1, \ldots, \tau_S)}, \tag{A.29}$$

where,

$$h(\tau_1,\ldots,\tau_S) = \sqrt{\frac{\eta^2}{\eta^2+w^2}}\prod_s \sqrt{\frac{\delta_s^2}{\delta_s^2+\tau_s\psi^2}} \cdot \prod_s \tau_s^{\frac{n_s}{2}-1} \exp\left(-\frac{1}{2}\sum_s \frac{\tau_s\delta_s^2}{\delta_s^2+\tau_s\psi^2}\cdot(\mathrm{RSS}_{0,s}-\mathrm{RSS}_{1,s})\right) \cdot \exp\left(\frac{1}{2}\frac{w^2}{\eta^2+w^2}\frac{\left(\sum_s \frac{\tau_s}{\delta_s^2+\tau_s\psi^2}\cdot\hat{\beta}_s\right)^2}{\eta^2}\right) \tag{A.30}$$

and

$$e^{g(\tau_1,\ldots,\tau_S)} = \prod_s \tau_s^{\frac{n_s}{2}-1}\cdot\exp\left(-\frac{1}{2}\sum_s \tau_s\cdot\mathrm{RSS}_{1,s}\right). \tag{A.31}$$

Again, function $g(\tau_1,\ldots,\tau_S)$ is maximized at

$$\hat{\tau}_s = \frac{n_s-2}{\mathrm{RSS}_{1,s}}\ ,\ s=1,\ldots,S. \tag{A.32}$$

We denote

$$d_s^2 = \frac{1}{\hat{\tau}_s}\delta_s^2 = \frac{\hat{\sigma}_s^2}{\boldsymbol{g}_s'\boldsymbol{g}_s - n_s\bar{g}_s^2}, \tag{A.33}$$

$$T_s^2 = \frac{\hat{\beta}_s}{d_s^2}, \tag{A.34}$$

$$\xi^2 = \left(\sum_s \frac{\hat{\tau}_s}{\delta_s^2+\hat{\tau}_s\psi^2}\right)^{-1} = \frac{1}{\sum_s (d_s^2+\psi^2)^{-1}}, \tag{A.35}$$

$$\hat{\bar{\beta}} = \frac{\sum_s (d_s^2+\psi^2)^{-1}\hat{\beta}_s}{\sum_s (d_s^2+\psi^2)^{-1}}, \tag{A.36}$$

and

$$\mathcal{T}_{\mathrm{ee}}^2 = \frac{\hat{\bar{\beta}}^2}{\xi^2} = \frac{\left(\sum_s \frac{\hat{\beta}}{d_s^2+\psi^2}\right)^2}{\eta^2}. \tag{A.37}$$

Using the similar procedure as in the ES model, we obtain

$$\mathrm{ABF}^{\mathrm{EE}}(\psi, w) = \sqrt{\frac{\xi^2}{\xi^2 + w^2}} \exp\left(\frac{\mathcal{T}_{\mathrm{ee}}^2}{2} \frac{w^2}{\xi^2 + w^2}\right) \prod_s \left(\sqrt{\frac{d_s^2}{d_s^2 + \psi^2}} \exp\left(\frac{T_s^2}{2} \frac{\psi^2}{d_s^2 + \psi^2}\right)\right). \tag{A.38}$$

Same as we have discussed in **Remarks** of section A.1, if $\tau_1, \ldots, \tau_S$ are known, the exact Bayes Factor of the EE model has the same function form as in (A.38), only with $\hat{\tau}_s$'s replaced by corresponding τ_s's.

A.3 Computation using CEFN Priors

Using curved exponential family normal prior, the computation of Bayes Factors is slightly different than what we show in previous sections. Here, we use ES model as a demonstration, the procedure for EE model is very similar.

To compute the Bayes Factor of a CEFN-ES model defined by parameters (k, ω) vs. the null model, we can carry out the same and exact calculation up to (A.5). However, due to the nature of CEFN prior, we can no longer perform analytic calculation to integrate out $\bar{b}$. Instead, we exchange the order of integrations by first analytically approximate the multi-dimensional integration with respect to $\tau_1, \ldots, \tau_S$ using the second procedure of Laplace's method described in previous sections. As a result, we obtain a approximate Bayes Factor as a one-dimensional integral

$$\begin{aligned}\mathrm{ABF}^{\mathrm{ES}}_{\mathrm{CEFN}}(k, \omega) = &\frac{1}{\sqrt{2\pi}\omega} \prod_s \left(\frac{\mathrm{RSS}_{0,s}}{\mathrm{RSS}_{1,s}}\right)^{n_s/2} \int_{-\infty}^{\infty} \prod_s \sqrt{\frac{\delta_s^2}{\delta_s^2 + k\bar{b}^2}} \\ &\cdot \exp\left[-\frac{1}{2}\left(\left(\sum_s \frac{1}{\delta_s^2 + k\bar{b}^2} + \frac{1}{\omega^2}\right)\bar{b}^2 - 2\sum_s \left(\frac{\hat{b}_s}{\delta_s^2 + k\bar{b}^2}\right)\bar{b} + \sum_s \frac{\delta_s^2}{\delta_s^2 + k\bar{b}^2} T_s^2\right)\right] d\bar{b}.\end{aligned} \tag{A.39}$$

We apply an adaptive Gaussian quadrature method, QAGI, implemented in GNU scientific library. Essentially, this method first maps the integral to the semi-open interval $[0, 1)$ using the transformation $y = (1-\bar{b})/\bar{b}$, then apply the standard adaptive Gaussian quadrature routine for the finite interval integration.

For EE model with CEFN prior, the final one-dimensional integral can be shown

as

$$\mathrm{ABF}^{\mathrm{EE}}_{\mathrm{CEFN}}(k,w) = \frac{1}{\sqrt{2\pi}w}\prod_s \left(\frac{\mathrm{RSS}_{0,s}}{\mathrm{RSS}_{1,s}}\right)^{n_s/2} \int_{-\infty}^{\infty} \prod_s \sqrt{\frac{d_s^2}{d_s^2 + k\bar{\beta}^2}}$$
$$\cdot \exp\left[-\frac{1}{2}\left((\sum_s \frac{1}{d_s^2 + k\bar{\beta}^2} + \frac{1}{w^2})\bar{\beta}^2 - 2\sum_s (\frac{\hat{\beta}_s}{d_s^2 + k\bar{\beta}^2})\bar{\beta} + \sum_s \frac{d_s^2}{d_s^2 + k\bar{\beta}^2} T_s^2\right)\right] d\bar{\beta}. \tag{A.40}$$

APPENDIX B
BAYES FACTOR FOR BINARY REGRESSION MODELS

In this section, we show the computation of Bayes Factor for case-control data.

Let us denote $\boldsymbol{\beta}_s = (\mu_s, \beta_s)$. The key component in our computation is to approximate subgroup-level log-likelihood function $l(\boldsymbol{\beta}_s)$ with a quadratic form expanding around its maximum likelihood estimates, i.e.

$$\log P(\boldsymbol{y}_s|\boldsymbol{g}_s, \boldsymbol{\beta}_s) = l(\boldsymbol{\beta}_s) \simeq l(\hat{\boldsymbol{\beta}}_s) - \frac{1}{2}(\boldsymbol{\beta}_s - \hat{\boldsymbol{\beta}}_s)' I_s(\hat{\boldsymbol{\beta}}_s)(\boldsymbol{\beta}_s - \hat{\boldsymbol{\beta}}_s), \tag{B.1}$$

where $I_s(\hat{\boldsymbol{\beta}}_s) = \begin{pmatrix} i_{\hat{\mu}_s\hat{\mu}_s} & i_{\hat{\mu}_s\hat{\beta}_s} \\ i_{\hat{\beta}_s\hat{\mu}_s} & i_{\hat{\beta}_s\hat{\beta}_s} \end{pmatrix}$ is the expected Fisher information evaluated at $\hat{\boldsymbol{\beta}}_s$. Although this type of approximation generally requires the observed Fisher information in (B.1), the observed and expected Fisher information indeed coincide as we use the canonical (logistic) link for binary regression model.

Further, we note

$$\gamma_s^2 := \mathrm{Var}(\hat{\beta}_s) = (i_{\hat{\beta}_s\hat{\beta}_s} - i_{\hat{\beta}_s\hat{\mu}_s} i^{-1}_{\hat{\mu}_s\hat{\mu}_s} i_{\hat{\mu}_s\hat{\beta}_s})^{-1} \tag{B.2}$$

is the estimated asymptotic variance of MLE $\hat{\beta}_s$.

Given approximate log-likelihood function (B.1) and a model H_c specified by (ψ, w), the prior distribution for $\boldsymbol{\beta}_s$ is given by

$$\boldsymbol{\beta}_s|\bar{\boldsymbol{\beta}}, H_c \sim \mathrm{N}(\bar{\boldsymbol{\beta}}, \Psi_s), \tag{B.3}$$

where

$$\bar{\boldsymbol{\beta}} = \begin{pmatrix} 0 \\ \bar{\beta} \end{pmatrix} \quad \text{and} \quad \Psi_s = \begin{pmatrix} v_s^2 & 0 \\ 0 & \psi^2 \end{pmatrix}. \tag{B.4}$$

It follows that

$$
\begin{aligned}
F_{H_c,s} &= \int P(\boldsymbol{y}_s|\boldsymbol{g}_s,\boldsymbol{\beta}_s)P(\boldsymbol{\beta}_s|\bar{\boldsymbol{\beta}},H_c)d\boldsymbol{\beta}_s \\
&= \exp\big(l(\hat{\boldsymbol{\beta}}_s)\big)\cdot|\Psi_s|^{-\frac{1}{2}}\cdot|I_s+\Psi_s^{-1}|^{-\frac{1}{2}}\cdot\exp\left(-\frac{1}{2}\hat{\boldsymbol{\beta}}_s'\big(I_s-I_s(I_s+\Psi_s^{-1})^{-1}I_s\big)\hat{\boldsymbol{\beta}}_s\right) \\
&\quad\cdot\exp\left(-\frac{1}{2}\Big(\bar{\boldsymbol{\beta}}'\big(\Psi_s^{-1}-\Psi_s^{-1}(I_s+\Psi_s^{-1})^{-1}\Psi_s^{-1}\big)\bar{\boldsymbol{\beta}}-\bar{\boldsymbol{\beta}}'\eta_s-\eta_s'\bar{\boldsymbol{\beta}}\Big)\right),
\end{aligned}
\tag{B.5}
$$

with $\eta_s=\Psi^{-1}(I_s+\Psi^{-1})^{-1}I_s\hat{\boldsymbol{\beta}}_s$.

Under contrasting null model H_0, the parameter space is restricted to $\beta_s=0$, for $\boldsymbol{\beta}_s$ satisfies this restriction

$$
(\boldsymbol{\beta}_s-\hat{\boldsymbol{\beta}}_s)'I_s(\hat{\boldsymbol{\beta}}_s)(\boldsymbol{\beta}_s-\hat{\boldsymbol{\beta}}_s)=i_{\hat{\mu}_s\hat{\mu}_s}\cdot(\mu_s-\hat{m}_s)^2+\frac{\hat{\beta}_s^2}{\gamma_s^2},
\tag{B.6}
$$

where $\hat{m}_s=\hat{\mu}_s+\frac{i_{\hat{\mu}_s\hat{\beta}_s}}{i_{\hat{\mu}_s\hat{\mu}_s}}\hat{\beta}_s$. It can be shown that

$$
\begin{aligned}
F_{H_0,s} &= \int P(\boldsymbol{y}_s|\boldsymbol{g}_s,\boldsymbol{\beta}_s)P(\boldsymbol{\beta}_s|\bar{\boldsymbol{\beta}},H_0)d\boldsymbol{\beta}_s \\
&= \exp\big(l(\hat{\boldsymbol{\beta}}_s)\big)\cdot v_s^{-1}(i_{\hat{\mu}\hat{\mu}}+v_s^{-2})^{-\frac{1}{2}}\cdot\exp(-\frac{\hat{\beta}_s^2}{2\gamma_s^2}) \\
&\quad\cdot\exp\left(-\frac{1}{2}\big(\ \hat{m}_s'i_{\hat{\mu}\hat{\mu}}\hat{m}_s-(i_{\hat{\mu}\hat{\mu}}\hat{m}_s)'(i_{\hat{\mu}\hat{\mu}}+v_s^{-2})^{-1}(i_{\hat{\mu}\hat{\mu}}\hat{m}_s)\ \big)\right).
\end{aligned}
\tag{B.7}
$$

Finally, we compute

$$
\mathrm{ABF}^{\mathrm{CC}}(\psi,w)=\lim\frac{\int(\prod_s F_{H_c,s})P(\bar{\beta}|H_c)\ d\bar{\beta}}{\prod_s F_{H_0,s}},
\tag{B.8}
$$

where the limit is taken as $v_s\to\infty,\ \forall s$. By straightforward algebra, we obtain the final result (2.36).

APPENDIX C
SMALL SAMPLE SIZE CORRECTION FOR APPROXIMATE BAYES FACTORS

In this section, we show in details that the impact of small sample size on approximate Bayes Factors and discuss our working solution (2.25) in dealing this issue.

Firstly, we show that for a valid Bayes Factor, (2.24) holds. This is because,

$$\mathrm{E}(\mathrm{BF}|H_0) = \int \frac{P(\boldsymbol{Y}|H_1)}{P(\boldsymbol{Y}|H_0)} \cdot P(\boldsymbol{Y}|H_0)\, d\boldsymbol{Y} = 1 \tag{C.1}$$

Secondly, we demonstrate this property does not generally hold in the approximate Bayes Factors. In particular, we consider a special case in which there is only a single subgroup. The approximate Bayes Factor assuming an ES model with paramters (ϕ, ω) can be reduced to

$$\mathrm{ABF}^{\mathrm{ES}}_{\mathrm{single}}(\phi, \omega) = \sqrt{1-\lambda}\exp(\frac{\lambda}{2}T_s^2), \tag{C.2}$$

and,

$$\log \mathrm{ABF}^{\mathrm{ES}}_{\mathrm{single}}(\phi, \omega) = \frac{1}{2}\log(1-\lambda) + \frac{\lambda}{2}T_s^2, \tag{C.3}$$

where $\lambda = \frac{\phi^2+\omega^2}{\phi^2+\omega^2+\delta_s^2}$ and takes values from $[0, 1]$.

Under H_0, T_s follows t-distribution with $n_s - 2$ degree of freedom and

$$\mathrm{E}(T_s^2|H_0) = \frac{n_s - 2}{n_s - 4} > 1. \tag{C.4}$$

Now consider the continuous function $f(\lambda) = \frac{1}{\lambda}\log(\frac{1}{1-\lambda})$ for $\lambda \in [0, 1]$, it can be shown that

$$\lim_{\lambda \to 0} f(\lambda) = 1 \tag{C.5}$$

$$\lim_{\lambda \to 1} f(\lambda) = \infty. \tag{C.6}$$

Hence, there must exist values of $\lambda \in (0, 1)$, such that

$$f(\lambda) < \mathrm{E}(T_s^2 | H_0). \tag{C.7}$$

Consequently, by Jensen's inequality, for those λ values

$$\log\left(\mathrm{E}\left(\mathrm{ABF}^{\mathrm{ES}}_{\mathrm{single}} | H_0\right)\right) \geq \mathrm{E}\left(\log\left(\mathrm{ABF}^{\mathrm{ES}}_{\mathrm{single}}\right) | H_0\right) > 0. \tag{C.8}$$

This shows property (2.24) does not generally hold for the approximate form of Bayes Factors and when sample size n_s is small, the inaccuracy becomes severe.

Ensuring the property (2.24) for approximate Bayes Factors also likely improves the accuracy of the approximation. To make simple corrections, we note that the approximate Bayes Factor (2.23) depends on data $\boldsymbol{Y}$ only through T_s (δ_s depneds on genotype data but not $\boldsymbol{Y}$). Further, from **Remark** in appendix A.1, we also notice the approximation becomes an exact Bayes Factor (for which property (2.24) is guaranteed) if estimated error variance terms $\hat{\sigma}_s^2$'s are replaced by their corresponding true values. When the true error variances are plugged in, under the H_0, T_s's follow the standard norm distribution. It is therefore sufficient to satisfy property (2.24) by quantile transforming each individual T_s in (2.23) from the t-distribution to the standard normal distribution. In essence, the correction can be viewed as a general strategy of providing a better point estimate of σ_s, therefore the similar strategy also likely improves the accuracy of approximate Bayes Factor when EE model and/or CEFN priors are used.

APPENDIX D
NUMERICAL ACCURACY OF BAYES FACTOR EVALUATIONS

In this section, we evaluate the numerical accuracy of various approximation methods for computing the Bayes Factors.

We use the population eQTL data (Stranger *et al.* (2007)) discussed in section 2.3.3. For each of the total 8,427 genes examined, we select the top associated *cis*-SNP based on the values of $\widehat{\text{BF}}^{\text{ES}}_{\text{meta}}$ and re-calculate the Bayes Factor directly based on (A.6) using a general adaptive Gaussian quadrature procedure (Note, because of its high computational cost in numerically evaluating multi-dimensional integrals, this numerical recipe does not apply in general practice). We treat these results as the "truth" and make comparison with $\widehat{\text{BF}}^{\text{ES}}_{\text{meta}}$ and $\text{ABF}^{\text{ES}}_{\text{all}}$ (with and without small sample corrections). Moreover, we convert various numerical results of Bayes Factors to log 10 scale and compute Root Mean Squared Errors (RMSE) for each approximation.

The results of the numerical evaluation for the ES model shown in Table D.1 and Figure D.1. Although the sample sizes in each subgroup are quite small in this dataset (41 Europeans, 59 Asians and 41 Africans), the numerical results of $\widehat{\text{BF}^{\text{ES}}}_{\text{all}}$ are almost identical to the results obtained from the adaptive Gaussian quadrature procedure (RMSE $= 1.2 \times 10^{-4}$ in log 10 scale). As expected, the approximate Bayes Factor, $\text{ABF}^{\text{ES}}_{\text{all}}$, has the worst numerical performance, mainly due to the small sample sizes in this dataset. Nevertheless, the ranking of the SNPs by $\text{ABF}^{\text{ES}}_{\text{all}}$ is quite consistent with what we obtain by the true values (rank correlation $= 0.99$). Figure D.1 suggests that under small sample situations, $\text{ABF}^{\text{ES}}_{\text{all}}$ tends to over-evaluate the true value and this over-evaluation can become quite severe when the true values are extremely large. On the other hand, the proposed small sample size correction method seems very effective: with this simple correction, the resulting $\text{A}^*\text{BF}^{\text{ES}}_{\text{all}}$ are quite accurate comparing with the true values.

We also perform a similar experiment for the EE model using the same dataset with five levels of $\sqrt{\psi^2 + w^2}$ values: 0.1, 0.2, 0.4, 0.8, 1.6, and seven degrees of

heterogeneities characterized by ψ^2/w^2 values: $0, 1/4, 1/2, 1, 2, 4, \infty$, and we assign these 35 grid values equal prior weight. The results are similar with the case in the EE model and shown in Table D.2.

	$\log_{10}(\widehat{\text{BF}^{\text{ES}}}_{\text{all}})$	$\log_{10}(\text{ABF}^{\text{ES}}_{\text{all}})$	$\log_{10}(\text{A}^*\text{BF}^{\text{ES}}_{\text{all}})$
RMSE	1.2×10^{-4}	4.95	0.14

Table D.1: Numerical accuracy of three approximations for evaluating Bayes Factors under the ES model. $\widehat{\text{BF}^{\text{ES}}}_{\text{all}}$ is based on the first approximation of Laplace's method discussed in appendix A, $\text{ABF}^{\text{ES}}_{\text{all}}$ is computed using (2.23) and $\text{A}^*\text{BF}^{\text{ES}}_{\text{all}}$ is based on (2.25) which is corrected for small sample sizes.

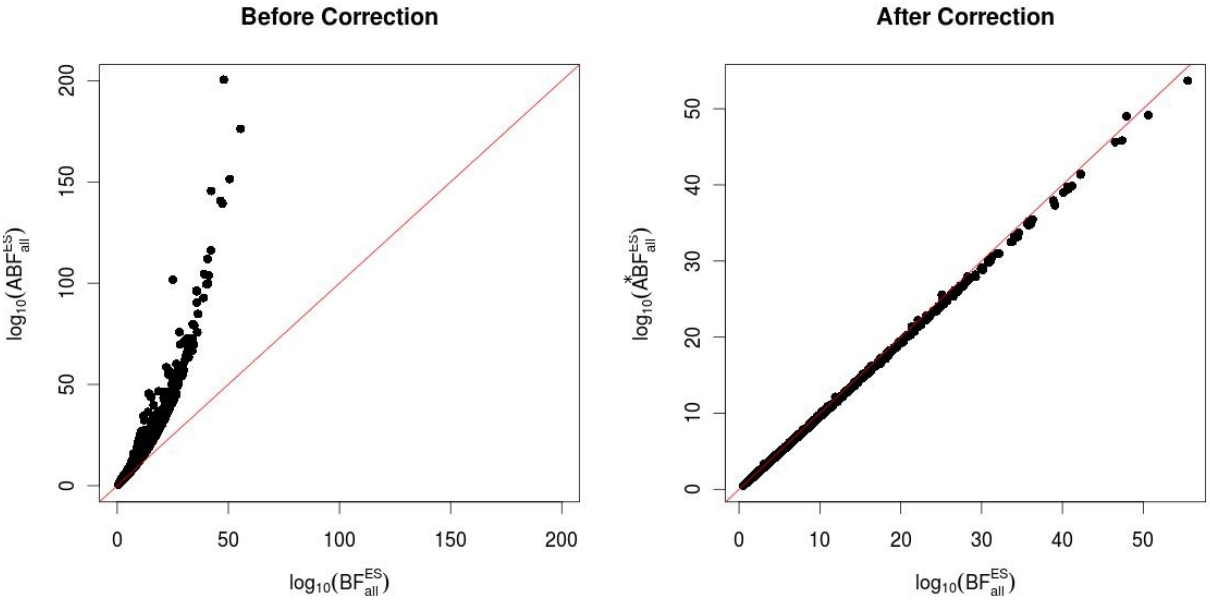

Figure D.1: Comparison of approximate Bayes Factors before and after applying small sample size corrections.

	$\log_{10}(\widehat{\mathrm{BF}^{\mathrm{EE}}}_{\mathrm{all}})$	$\log_{10}(\mathrm{ABF}^{\mathrm{EE}}_{\mathrm{all}})$	$\log_{10}(\mathrm{A^*BF}^{\mathrm{EE}}_{\mathrm{all}})$
RMSE	4.1×10^{-4}	5.03	0.09

Table D.2: Numerical accuracy of three approximations for evaluating Bayes Factors under the EE model.

APPENDIX E
USING IMPUTED GENOTYPES IN BAYESIAN ANALYSIS OF GENETIC ASSOCIATION DATA

In this section, we formalize the strategy proposed by Guan and Stephens (2008) in handling untyped SNPs in GWA studies and incorporate it into our Bayesian meta-analysis framework.

Suppose in study s, SNP t is untyped. The association between phenotype $\boldsymbol{y}_s$ and unobserved genotype $\boldsymbol{g}_{s,t}$ is described by linear model

$$\boldsymbol{y}_s = \mu_{s,t}\mathbf{1} + \beta_{s,t}\boldsymbol{g}_{s,t} + \boldsymbol{e}_{s,t}, \tag{E.1}$$

where the error term $\boldsymbol{e}_{s,t}$ has mean 0 and variance $\sigma^2_{s,t}I$. Information of $\boldsymbol{g}_{s,t}$ can be learned through imputation algorithm by estimating $\hat{\boldsymbol{g}}_{s,t} = \mathrm{E}(\boldsymbol{g}_{s,t}|\boldsymbol{D}_s, \boldsymbol{M})$ and we also use a linear model to connect the imputed and unobserved true genotypes, namely,

$$\boldsymbol{g}_{s,t} = \hat{\boldsymbol{g}}_{s,t} + \boldsymbol{d}_{s,t}, \tag{E.2}$$

where the error term $\boldsymbol{d}$ has mean 0 and diagonal variance-covariance matrix $V = \mathrm{Var}(\boldsymbol{g}_{s,t}|\boldsymbol{D}_s, \boldsymbol{M})$.

Combining (E.1) and (E.2), we obtain a new linear relationship between observed phenotypes and imputed "mean" genotypes,

$$\boldsymbol{y}_s = \mu_{s,t}\mathbf{1} + \beta_{s,t}\hat{\boldsymbol{g}}_{s,t} + \boldsymbol{\epsilon}_{s,t}, \tag{E.3}$$

where the compound error term,

$$\boldsymbol{\epsilon}_{s,t} = \beta_{s,t}\boldsymbol{d}_{s,t} + \boldsymbol{e}_{s,t}, \tag{E.4}$$

has mean 0 and variance $\beta^2_{s,t}V + \sigma^2_{s,t}I$. This formulation is known as Errors-in-Variables (EIV) model.

Because of the functional relationship between the mean and variance, the exact calculation based on this EIV model is generally difficult. The strategy proposed by

Guan and Stephens (2008) essentially ignores the complicated mean-variance relationship. If $\beta_{s,t} = 0$, i.e. the null model is true, (E.3) reduces to a regular linear model and this treatment is indeed correct. If $\beta_{s,t} \neq 0$, the proposed strategy yields naive least squares estimates of $\beta_{s,t}$ shrunk towards 0 and $\sigma^2_{s,t}$ overestimated (Hodges and Moore (1972)). Also we note that the value of $\delta^2_{s,t}$ computed using "mean" genotypes is larger than the value obtained using the true genotypes. Consequently, as we can see from (2.23), this strategy yields a conservative Bayes Factor under the alternative model. Nevertheless, if the imputation is accurate, that is, the magnitude of imputation error $\boldsymbol{d}$ is much less than the residual error $\boldsymbol{e}$, the resulting Bayes Factor is very close to the truth.

In the extreme situation when the untyped SNP t has little correlation with typed SNPs based on panel $\boldsymbol{M}$, it follows that

$$\mathrm{E}(\boldsymbol{g}_{s,t}|\boldsymbol{G}_s, \boldsymbol{M}) \simeq \mathrm{E}(\boldsymbol{g}_{s,t}|\boldsymbol{M}), \tag{E.5}$$

and the expected genotypes rely only the marginal information of SNP t in the panel. Consequently, the expected genotypes have very little individual variation ($\delta^2_{s,t} \to \infty$) and based on (2.28) this imputation result has little impact on the final outcome of meta-analysis Bayes Factor.

APPENDIX F
LEARNING FROM PANEL USING LI AND STEPHENS MODEL

In this section, we show the calculation of $\hat{\boldsymbol{\mu}}$ and $\hat{\Sigma}$ using phased population panel data. Following the Li and Stephens model, we assume that there are K template haplotypes in the panel. For a new haplotype $\boldsymbol{h}$ sampled from the same population, it can be modeled as an imperfect mosaic of existing template haplotypes in the panel. Let $\boldsymbol{e}_j$ denote the jth unit vector in K dimensions (1 in the jth coordinate, 0's elsewhere), we define random vector $\boldsymbol{Z}_t$ to be $\boldsymbol{e}_j$ if haplotype $\boldsymbol{h}$ at locus t copies from jth template haplotype. The model also assumes $\boldsymbol{Z}_1, ..., \boldsymbol{Z}_n$ form a Markov chain in state space $\{\boldsymbol{e}_1, ..., \boldsymbol{e}_K\}$ with transition probabilities

$$\Pr(\boldsymbol{Z}_t = \boldsymbol{e}_m | \boldsymbol{Z}_{t-1} = \boldsymbol{e}_n, \boldsymbol{M}) = (1 - r_t) 1_{[\boldsymbol{e}_m = \boldsymbol{e}_n]} + r_t \mathbf{e}'_m \boldsymbol{\alpha}, \tag{F.1}$$

where $\boldsymbol{\alpha} = \frac{1}{K} \cdot \mathbf{1}$ and $r_t = 1 - \exp(-\rho_t / K)$ is a parameter that controls the probability that $\boldsymbol{h}$ switches copying template at locus t. The initial-state probabilities of the Markov chain is given by

$$\pi(\boldsymbol{Z}_1 = \boldsymbol{e}_k | \boldsymbol{M}) = \frac{1}{K} \quad \text{for } k = 1, \cdots, K. \tag{F.2}$$

It is easy to check that the initial distribution π is also the stationary distribution of the described Markov chain. Because the chain is initiated at the stationary state, it follows that conditional on $\boldsymbol{M}$

$$\boldsymbol{Z}_1 =^d \boldsymbol{Z}_2 =^d \cdots =^d \boldsymbol{Z}_p =^d \pi. \tag{F.3}$$

Therefore marginally, the means and variances of $\boldsymbol{Z}_t$s have following simple forms:

$$\begin{aligned} \mathrm{E}(\boldsymbol{Z}_1 | \boldsymbol{M}) &= \cdots = \mathrm{E}(\boldsymbol{Z}_p | \boldsymbol{M}) = \boldsymbol{\alpha}, & \text{(F.4)} \\ \mathrm{Var}(\boldsymbol{Z}_1 | \boldsymbol{M}) &= \cdots = \mathrm{Var}(\boldsymbol{Z}_n | \boldsymbol{M}) = \mathrm{diag}(\boldsymbol{\alpha}) - \boldsymbol{\alpha}\boldsymbol{\alpha}'. & \text{(F.5)} \end{aligned}$$

Let K-dimensional vector $\boldsymbol{q}_t^{\text{panel}}$ denote the binary allelic state of panel haplotypes at locus t and scalar parameter θ represents mutation. The emission distribution in Li and Stephens model is given by

$$\Pr(h_t = 1|\boldsymbol{Z}_t = \boldsymbol{e}_k, \boldsymbol{M}) = (1-\theta)\boldsymbol{e}_k'\boldsymbol{q}_t^{\text{panel}} + \frac{1}{2}\theta, \tag{F.6}$$

that is, with probability $1-\theta$, $\boldsymbol{h}$ perfectly copies from k-th template in the panel at locus t, while with probability θ, a mutation occurs and h_t "mutates" to allele 0 or 1 equally likely. If we define $\boldsymbol{p}_t = (1-\theta)\boldsymbol{q}_t^{\text{panel}} + \frac{\theta}{2}\mathbf{1}$, then the emission distribution can be written as

$$\Pr(h_t = 1|\boldsymbol{Z}_t, \boldsymbol{M}) = \mathrm{E}(h_t|\boldsymbol{Z}_t, \boldsymbol{M}) = \boldsymbol{p}_t'\boldsymbol{Z}_t. \tag{F.7}$$

The goal here is to find the closed-form representations of first two moments of joint distribution $(h_1, h_2, ..., h_p)$ given the observed template panel $\boldsymbol{M}$. For marginal mean and variance of h_t, it follows that

$$\begin{aligned}\mathrm{E}(h_t|\boldsymbol{M}) &= \mathrm{E}(\mathrm{E}(h_t|\boldsymbol{Z}_t, \boldsymbol{M})|\boldsymbol{M}) \\ &= \boldsymbol{p}_t'\mathrm{E}(\boldsymbol{Z}_t|\boldsymbol{M}) \\ &= (1-\theta)\cdot f_t^{\text{panel}} + \frac{\theta}{2},\end{aligned} \tag{F.8}$$

$$\mathrm{Var}(h_t|\boldsymbol{M}) = (1-\theta)^2 f_t^{\text{panel}}(1 - f_t^{\text{panel}}) + \frac{\theta}{2}(1-\frac{\theta}{2}), \tag{F.9}$$

where $f_t^{\text{panel}} = \boldsymbol{q}_t'\cdot\boldsymbol{\alpha}$ is the observed allele frequency at locus t from panel $\boldsymbol{M}$. Finally, to compute $\mathrm{Cov}(h_s, h_t)$ for some loci $s < t$, we notice that conditional on $\boldsymbol{Z}_s$ and $\boldsymbol{M}$, h_s and h_t are independent and

$$\begin{aligned}\mathrm{E}(h_s \cdot h_t|\boldsymbol{M}) &= \mathrm{E}(\mathrm{E}(h_s \cdot h_t|\boldsymbol{Z}_s, \boldsymbol{M})|\boldsymbol{M}) \\ &= \mathrm{E}\big(\ \mathrm{E}(h_s|\boldsymbol{Z}_s, \boldsymbol{M}) \cdot \mathrm{E}(h_t|\boldsymbol{Z}_s, \boldsymbol{M})\ |\ \boldsymbol{M}\big).\end{aligned} \tag{F.10}$$

Let r_{st} denote the switching probability between s and t, and $\mathrm{E}(h_t|\boldsymbol{Z}_s, \boldsymbol{M})$ can be calculated from

$$\begin{aligned}\mathrm{E}(h_t|\boldsymbol{Z}_s, \boldsymbol{M}) &= \mathrm{E}(\ \mathrm{E}(h_t|\boldsymbol{Z}_t, \boldsymbol{Z}_s, \boldsymbol{M})|\boldsymbol{Z}_s, \boldsymbol{M}) \\ &= \mathrm{E}(\boldsymbol{Z}_t'\boldsymbol{p}_t|\boldsymbol{Z}_s, \boldsymbol{M}) \\ &= ((1-r_{st})\boldsymbol{Z}_s' + r_{st}\boldsymbol{\alpha}')\boldsymbol{p}_t.\end{aligned} \tag{F.11}$$

Therefore,

$$\begin{aligned}\mathrm{E}(h_s \cdot h_t|\boldsymbol{M}) &= \boldsymbol{p}_s'\mathrm{E}(\boldsymbol{Z}_s(\ (1-r_{st})\boldsymbol{Z}_s' + r_{st}\boldsymbol{\alpha}')|\boldsymbol{M})\boldsymbol{p}_t \\ &= (1-r_{st})\boldsymbol{p}_s'\mathrm{Var}(\boldsymbol{Z}_s)\boldsymbol{p}_t + \boldsymbol{p}_s'\boldsymbol{\alpha}\boldsymbol{\alpha}'\boldsymbol{p}_t \\ &= (1-r_{st}) \cdot \boldsymbol{p}_s'(\mathrm{diag}(\boldsymbol{\alpha}) - \boldsymbol{\alpha}\boldsymbol{\alpha}')\boldsymbol{p}_t + \mathrm{E}(h_s|\boldsymbol{M})\mathrm{E}(h_t|\boldsymbol{M}),\end{aligned} \tag{F.12}$$

It follows that

$$\mathrm{Cov}(h_s, h_t|\boldsymbol{M}) = (1-\theta)^2(1-r_{st})(f_{st}^{\mathrm{panel}} - f_s^{\mathrm{panel}} f_t^{\mathrm{panel}}), \tag{F.13}$$

where f_{st}^{panel} is the panel frequency of the haplotype "$1-1$" consisting of loci s and t.

In conclusion, under the Li and Stephens model , the distribution $\boldsymbol{h} \mid \boldsymbol{M}$ has expectation

$$\hat{\boldsymbol{\mu}} = \mathrm{E}(\boldsymbol{h}|\boldsymbol{M}) = (1-\theta)\boldsymbol{f}^{\mathrm{panel}} + \frac{\theta}{2}\mathbf{1}, \tag{F.14}$$

where $\boldsymbol{f}^{\mathrm{panel}}$ is the p-vector of observed frequencies of all p SNPs in the panel, and variance

$$\hat{\Sigma} = \mathrm{Var}(\boldsymbol{h}|\boldsymbol{M}) = (1-\theta)^2 S + \frac{\theta}{2}(1-\frac{\theta}{2})I, \tag{F.15}$$

where matrix S has the structure

$$S_{ij} = \begin{cases} f_i^{\mathrm{panel}}(1-f_i^{\mathrm{panel}}) & i = j \\ (1-r_{ij})(f_{ij}^{\mathrm{panel}} - f_i^{\mathrm{panel}} f_j^{\mathrm{panel}}) & i \neq j \end{cases} \tag{F.16}$$

APPENDIX G
DERIVATION OF JOINT GENOTYPE FREQUENCY DISTRIBUTION

In this section, we derive the joint genotype frequency distribution based on the Li and Stephens model.

Let g_{it} denote the genotype of individual i at locus t. The sample frequency of genotype 0 at locus t is given by

$$\Pr(g_t = 0) = p_0^{g_t} = \frac{1}{n}\sum_{i=1}^{n} 1_{[g_{it}=0]}. \tag{G.1}$$

Similarly, genotype frequencies $p_1^{g_t} = \Pr(g_t = 1)$ and $p_2^{g_t} = \Pr(g_t = 2)$ can be obtained by averaging indicators $1_{[g_{it}=1]}$ and $1_{[g_{it}=2]}$ over the samples respectively. Because of the restriction

$$1_{[g_{it}=0]} + 1_{[g_{it}=1]} + 1_{[g_{it}=2]} = 1, \tag{G.2}$$

given any two of the three indicators, the third one is uniquely determined. Let $\boldsymbol{g}_i$ denote $2p$-vector $(1_{[g_{i1}=0]}, 1_{[g_{i1}=2]}, \ldots, 1_{[g_{ip}=0]}, 1_{[g_{ip}=2]})$ and

$$\boldsymbol{y_g} = (p_0^{g_1}\ p_2^{g_1}\ \ldots\ p_0^{g_p}\ p_2^{g_p})' = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{g}_i. \tag{G.3}$$

Assuming that $\boldsymbol{g}_1, \ldots, \boldsymbol{g}_n$ are i.i.d draws from conditional distribution $\Pr(\boldsymbol{g}|\boldsymbol{M})$, by central limit theorem as sample size n is large, it follows

$$\boldsymbol{y_g}|\boldsymbol{M} \ \sim\ \mathrm{N}_{2p}(\boldsymbol{\mu_g}, \Sigma_g), \tag{G.4}$$

where $\boldsymbol{\mu_g} = \mathrm{E}(\boldsymbol{g}|\boldsymbol{M})$ and $\Sigma_g = \mathrm{Var}(\boldsymbol{g}|\boldsymbol{M})$.

For the remaining part of this section, we derive the closed-form expressions for $\boldsymbol{\mu_g}$ and Σ_g based on the Li and Stephens model.

Let $\boldsymbol{h}^a$ and $\boldsymbol{h}^b$ denote the two composing haplotypes for some genotype sampled

from population. Note that

$$\begin{aligned} 1_{[g_t=0]} &= (1-h_t^a)(1-h_t^b), \\ 1_{[g_t=2]} &= h_t^a h_t^b. \end{aligned} \tag{G.5}$$

Given panel $\boldsymbol{M}$, the two composing haplotypes are also assumed to be independent and identically distributed. Following the results from Appendix A, we obtain that

$$\begin{aligned} \mathrm{E}(1_{[g_t=0]}|\boldsymbol{M}) &= (1-\mathrm{E}(h_t|\boldsymbol{M}))^2, \\ \mathrm{E}(1_{[g_t=2]}|\boldsymbol{M}) &= \mathrm{E}(h_t|\boldsymbol{M})^2, \\ \mathrm{Var}(1_{[g_t=0]}|\boldsymbol{M}) &= (1-\mathrm{E}(h_t|\boldsymbol{M}))^2 \cdot (1-(1-\mathrm{E}(h_t|\boldsymbol{M}))^2), \\ \mathrm{Var}(1_{[g_t=2]}|\boldsymbol{M}) &= \mathrm{E}(h_t|\boldsymbol{M})^2 \cdot (1-\mathrm{E}(h_t|\boldsymbol{M})^2) \\ \mathrm{Cov}(1_{[g_t=0]}, 1_{[g_t=2]}|\boldsymbol{M}) &= -(1-\mathrm{E}(h_t|\boldsymbol{M}))^2 \cdot \mathrm{E}(h_t|\boldsymbol{M})^2 \end{aligned} \tag{G.6}$$

where $\mathrm{E}(h_t|\boldsymbol{M})$ is given by (F.8).

To compute covariance across different loci s and t, we note that

$$\begin{aligned} 1_{[g_s=0]}1_{[g_t=0]} &= (1-h_s^a)(1-h_s^b)(1-h_t^a)(1-h_t^b), \\ 1_{[g_s=2]}1_{[g_t=2]} &= h_s^a h_s^b h_t^a h_t^b, \\ 1_{[g_s=0]}1_{[g_t=2]} &= (1-h_s^a)(1-h_s^b)h_t^a h_t^b, \\ 1_{[g_s=2]}1_{[g_t=0]} &= h_s^a h_s^b (1-h_t^a)(1-h_t^b) \end{aligned} \tag{G.7}$$

Then all the covariance terms across different loci can be represented using $\mathrm{E}(h_s h_t|\boldsymbol{M})$, which is given by (F.12). For example,

$$\mathrm{Cov}(1_{[g_s=2]}, 1_{[g_t=2]}|\boldsymbol{M}) = \mathrm{Cov}(h_s, h_t|\boldsymbol{M})^2 + 2\mathrm{E}(h_s|\boldsymbol{M}) \cdot \mathrm{E}(h_t|\boldsymbol{M}) \cdot \mathrm{Cov}(h_s, h_t|\boldsymbol{M}). \tag{G.8}$$

APPENDIX H
MODIFIED ECM ALGORITHM FOR IMPUTING GENOTYPES WITHOUT A PANEL

In this section, we show our modified ECM algorithm for genotype imputation without a panel.

By our assumption, each individual genotype p-vector $\boldsymbol{g}^i$ is a random sample from the multivariate normal distribution $\mathrm{N}_p(\hat{\boldsymbol{\mu}}, \hat{\Sigma})$, and different individual vectors may have different missing entries. Suppose we have n individual samples, let $\boldsymbol{G}_{\text{obs}}$ denote the set of all typed genotypes across all individuals and $\boldsymbol{g}^i_{\text{obs}}$ denote the typed genotypes for individual i.

In the E step of ECM algorithm, we compute the expected values of the sufficient statistics $\sum_{i=1}^n g_j^i$ for $j = 1, \ldots, p$ and $\sum_{i=1}^n g_j^i g_k^i$ for $j, k = 1, \ldots, p$ conditional on $\boldsymbol{G}_{\text{obs}}$ and current estimate for $(\hat{\boldsymbol{\mu}}, \hat{\Sigma})$. Specifically, in t-th iteration,

$$\mathrm{E}(\sum_{i=1}^n g_j^i | \boldsymbol{G}_{\text{obs}}, \hat{\boldsymbol{\mu}}^{(t)}, \hat{\Sigma}^{(t)}) = \sum_{i=1}^n g_j^{i,(t)}, \tag{H.1}$$

$$\mathrm{E}(\sum_{i=1}^n g_j^i g_k^i | \boldsymbol{G}_{\text{obs}}, \hat{\boldsymbol{\mu}}^{(t)}, \hat{\Sigma}^{(t)}) = \sum_{i=1}^n (g_j^{i,(t)} g_k^{i,(t)} + c_{jk}^{i,(t)}), \tag{H.2}$$

where

$$g_j^{i,(t)} = \begin{cases} g_j^i & \text{if } g_j^i \text{ is typed} \\ \mathrm{E}(g_j^i | \boldsymbol{g}^i_{\text{obs}}, \hat{\boldsymbol{\mu}}^{(t)}, \hat{\Sigma}^{(t)}) & \text{if } g_j^i \text{ is untyped,} \end{cases} \tag{H.3}$$

and

$$c_{jk}^{i,(t)} = \begin{cases} 0 & \text{if } g_j^i \text{ or } g_k^i \text{ is typed} \\ \mathrm{Cov}(g_j^i, g_k^i | \boldsymbol{g}^i_{\text{obs}}, \hat{\boldsymbol{\mu}}^{(t)}, \hat{\Sigma}^{(t)}) & \text{if } g_j^i \text{ and } g_k^i \text{ are both untyped.} \end{cases} \tag{H.4}$$

The calculation of $\mathrm{E}(g_j^i | \boldsymbol{g}^i_{\text{obs}}, \hat{\boldsymbol{\mu}}^{(t)}, \hat{\Sigma}^{(t)})$ and $\mathrm{Cov}(g_j^i, g_k^i | \boldsymbol{g}^i_{\text{obs}}, \hat{\boldsymbol{\mu}}^{(t)}, \hat{\Sigma}^{(t)})$ follows directly from (4.4).

In the conditional maximization step, we first update the estimates for $\boldsymbol{f}^{\text{panel}}$ and

Σ^{panel} sequentially, i.e.

$$f_j^{\text{panel},(t+1)} = \frac{1}{n}\text{E}(\sum_{i=1}^{n} g_j^i | \boldsymbol{G}_{\text{obs}}, \hat{\boldsymbol{\mu}}^{(t)}), \text{ for} j = 1, \ldots, p, \tag{H.5}$$

and

$$\begin{aligned}\Sigma_{jk}^{\text{panel},(t+1)} &= \frac{1}{n}\text{E}(\sum_{i=1}^{n} g_j^i g_k^i | \boldsymbol{G}_{\text{obs}}, \hat{\boldsymbol{\mu}}^{(t)}, \hat{\Sigma}^{(t)}) \\ &\quad - f_j^{\text{panel},(t+1)} f_k^{\text{panel},(t+1)}, \text{ for} j, k = 1, \ldots, p.\end{aligned} \tag{H.6}$$

Finally, we update the shrinkage estimates $\hat{\boldsymbol{\mu}}$ and $\hat{\Sigma}$ using

$$\hat{\boldsymbol{\mu}}^{(t+1)} = (1-\theta)\boldsymbol{f}^{\text{panel},(t+1)} + \tfrac{\theta}{2}\mathbf{1}, \tag{H.7}$$

$$\hat{\Sigma}^{(t+1)} = (1-\theta)^2 S^{(t+1)} + \tfrac{\theta}{2}(1-\tfrac{\theta}{2})I, \tag{H.8}$$

where

$$S_{jk}^{(t+1)} = \begin{cases} \Sigma_{jk}^{\text{panel},(t+1)} & j = k \\ \exp(-\frac{\rho_{jk}}{2n})\Sigma_{jk}^{\text{panel},(t+1)} & j \neq k. \end{cases} \tag{H.9}$$

We initiated ECM algorithm by setting $\boldsymbol{f}^{\text{panel},(0)}$ to the marginal means from all observed data and $\Sigma^{\text{panel},(0)}$ to a diagonal matrix with diagonal entries being empirical variance computed from typed SNPs.

APPENDIX I
A GENERAL HIDDEN MARKOV MODEL APPROACH FOR ALLELE FREQUENCY IMPUTATION

This appendix summarizes our result on a general HMM approach for imputing allele frequencies based on a more general fastPHASE model (Scheet and Stephens (2005)). The input settings are the same as described in chapter 4.

I.1 The fastPHASE Model

Suppose we observe N haplotypes, each is genotyped at T markers. The fastPHASE model assumes that for each sampled haplotype at each locus, the allele at that locus is originated from one of K clusters. Let z_{it} denote the cluster origin at locus t for i-th haplotype, then latent variable $\boldsymbol{z}_i = (z_{i1}, z_{i2}, ..., z_{iT})$ forms a Markov chain on state space $\{1, ..., K\}$. The initial-state probabilities of this Markov chain are given by

$$\Pr(z_{i1} = k) = \alpha_{k1}, \quad k = 1, ..., K. \tag{I.1}$$

The transition probabilities are given by

$$\Pr(z_{it} = k' | z_{i(t-1)} = k) = (1 - r_t) 1_{\{k=k'\}} + r_t \alpha_{kt}. \tag{I.2}$$

Loosely speaking, parameter r_t measures the locus-specific probability of a haplotype switching cluster membership (or a "jump") at locus t. If a jump occurs, a locus-specific K-dimensional vector $\boldsymbol{\alpha}_t$ describes the probability distribution of the new cluster membership. Let x_{it} denote the allele at locus t for i-th haplotype. Given cluster origin at locus t, the probability of observing allele 1 is determined by

$$\Pr(x_{it} = 1 | z_{it} = k) = \theta_{kt}, \tag{I.3}$$

where $\boldsymbol{\theta}_t$ is a locus-specific K-dimensional vector, and it describes the cluster-specific allele frequency at locus t.

I.2 The State Equation

To extend the fastPHASE model with K clusters assuming only allele frequencies are observed, we adopt the following vector notation. Let $\boldsymbol{e}_j$ denote the jth unit vector in K dimensions (1 in the jth coordinate, 0's elsewhere), and define the random vector $\boldsymbol{Z}_{it}$ to be $\boldsymbol{e}_j$ if the ith haplotype at locus t is originated from the jth cluster. Then $\boldsymbol{Z}_{i1}, ..., \boldsymbol{Z}_{in}$ form a markov chain on $\{\boldsymbol{e}_1, ..., \boldsymbol{e}_K\}$ with the transition probabilities

$$\Pr(\boldsymbol{Z}_{it} = \boldsymbol{e}_m | \boldsymbol{Z}_{i(t-1)} = \boldsymbol{e}_n) = (1 - r_t) 1_{\{\boldsymbol{e}_m = \boldsymbol{e}_n\}} + r_t \mathbf{e}'_m \boldsymbol{\alpha}_t \tag{I.4}$$

and inital-state probabilities

$$\Pr(\boldsymbol{Z}_{i1} = \boldsymbol{e}_k) = \boldsymbol{e}'_k \boldsymbol{\alpha}_1 \tag{I.5}$$

Consequently, the expectation and variance-covariance matrix of conditional random vector $\boldsymbol{Z}_{i(t+1)} | \boldsymbol{Z}_{it}$ can be easily computed, i.e.

$$\mathrm{E}(\boldsymbol{Z}_{it} | \boldsymbol{Z}_{i(t-1)}) = (1 - r_t) \boldsymbol{Z}_{i(t-1)} + r_t \boldsymbol{\alpha}_t = \boldsymbol{p}_{it} \tag{I.6}$$

$$\mathrm{Var}(\boldsymbol{Z}_{it} | \boldsymbol{Z}_{i(t-1)}) = \mathrm{diag}(\boldsymbol{p}_{it}) - \boldsymbol{p}_{it} \boldsymbol{p}'_{it} \tag{I.7}$$

As we assume that the N haplotype samples are independently drawn from population,

$$\mathrm{E}\left(\sum_{i=1}^{N} \boldsymbol{Z}_{it} \;\middle|\; \boldsymbol{Z}_{1(t-1)}, ..., \boldsymbol{Z}_{N(t-1)} \right) = (1 - r_t) \sum_{i=1}^{N} \boldsymbol{Z}_{i(t-1)} + N r_t \boldsymbol{\alpha}_t, \tag{I.8}$$

and

$$\begin{aligned}\mathrm{Var}\left(\sum_{i=1}^{N}\boldsymbol{Z}_{it} \mid \boldsymbol{Z}_{1(t-1)},...,\boldsymbol{Z}_{N(t-1)}\right) &= \sum_{i=1}^{N}\mathrm{diag}(\boldsymbol{p}_{it}) - \sum_{i=1}^{N}\boldsymbol{p}_{it}\boldsymbol{p}'_{it} \\ &= \mathrm{diag}\big((1-r_t)\sum_{i=1}^{N}\boldsymbol{Z}_{i(t-1)} + Nr_t\boldsymbol{\alpha}_t\big) \\ &- (1-r_t)^2\mathrm{diag}(\sum_{i=1}^{N}\boldsymbol{Z}_{i(t-1)}) \\ &- r_t(1-r_t)\boldsymbol{\alpha}_t\big(\sum_{i=1}^{N}\boldsymbol{Z}_{i(t-1)}\big)' \\ &- r_t(1-r_t)\big(\sum_{i=1}^{N}\boldsymbol{Z}_{i(t-1)}\big)\boldsymbol{\alpha}'t \\ &- N{r_t}^2\boldsymbol{\alpha}_t\boldsymbol{\alpha}'t.\end{aligned} \tag{I.9}$$

In the above calculation, we note that $\sum_{i=1}^{N}\boldsymbol{Z}_{it}$ depends on $(\boldsymbol{Z}_{1(t-1)},...,\boldsymbol{Z}_{N(t-1)})$ only through the summary statistic $\sum_{i=1}^{N}\boldsymbol{Z}_{i(t-1)}$. Let $\boldsymbol{f}_t = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{Z}_{it}$, then

$$\mathrm{E}(\ \boldsymbol{f}_t \mid \boldsymbol{f}_{t-1}\ ) = (1-r_t)\boldsymbol{f}_{t-1} + r_t\boldsymbol{\alpha}_t, \tag{I.10}$$

and

$$\begin{aligned}V(\boldsymbol{f}_{t-1}) := \mathrm{Var}(\boldsymbol{f}_t|\boldsymbol{f}_{t-1}) =& \frac{1}{N}\Big(\mathrm{diag}\left((1-r_t)\boldsymbol{f}_{t-1} + r_t\boldsymbol{\alpha}_t\right) - (1-r_t)^2\mathrm{diag}(\boldsymbol{f}_{t-1}) \\ &- r_t(1-r_t)\boldsymbol{\alpha}_t\boldsymbol{f}'_{t-1} - r_t(1-r_t)\boldsymbol{f}_{t-1}\boldsymbol{\alpha}'t - {r_t}^2\boldsymbol{\alpha}_t\boldsymbol{\alpha}'t\Big).\end{aligned} \tag{I.11}$$

Naturally, $\boldsymbol{f}_t$ can be interpreted as the cluster membership configuration at locus t, e.g. the k-th entry of $\boldsymbol{f}_t$ is the fraction of alleles orinated from k-th cluster among all sampled alleles at locus t. By Central Limit Theorem, as sample size N is sufficiently large, the conditional distribution $\boldsymbol{f}_t \mid \boldsymbol{f}_{t-1}$ can be approximated by a normal

distribution , i.e.,

$$\boldsymbol{f}_t \mid \boldsymbol{f}_{t-1} \sim \mathrm{N}_K \left((1-r_t)\boldsymbol{f}_{t-1} + r_t \boldsymbol{\alpha}_t \ , \ V(\boldsymbol{f}_{t-1}) \right) \tag{I.12}$$

Written in form of a linear state evolution equation, we obtain

$$\boldsymbol{f}_t = \boldsymbol{T}_t \boldsymbol{f}_{t-1} + r_t \boldsymbol{\alpha}_t + \boldsymbol{v}_t, \quad \boldsymbol{v}_t \sim \mathrm{N}_K \left(\boldsymbol{0} \ , \ V(\boldsymbol{f}_{t-1}) \right) \tag{I.13}$$

where $\boldsymbol{T}_t = (1-r_t)\boldsymbol{I}$ and $\boldsymbol{v}_t$ is a K-dimenisonal random vector that accounts for the random noise in the process.

I.3 The Observation Equation

Let $Y_{1t}, ..., Y_{Kt}$ denote the counts of allele 1 from the samples in clusters $1, ..., K$ at locus t respectively. Given the cluster membership proportion vector $\boldsymbol{f}_t$ and $\boldsymbol{\theta}$, Y_{it} ~Binomial(Nf_{it} , θ_{it}), where f_{it} and θ_{it} are the ith entries of vectors $\boldsymbol{f}_t$ and $\boldsymbol{\theta}_t$ repectively. And the conditional expectation and variance can be calculated as following,

$$\mathrm{E}(Y_{it} \mid f_{it}) = N f_{it} \theta_{it} \tag{I.14}$$

$$\mathrm{Var}(Y_{it} \mid f_{it}) = N f_{it} \theta_{it} (1 - \theta_{it}) \tag{I.15}$$

Because Y_{it}'s are mutually independent, we have

$$\mathrm{E} \left(\sum_{i=1}^{K} Y_{it} \mid \boldsymbol{f}_t \right) = N \boldsymbol{\theta}_t' \boldsymbol{f}_t \tag{I.16}$$

$$\mathrm{Var} \left(\sum_{i=1}^{K} Y_{it} \mid \boldsymbol{f}_t \right) = N \boldsymbol{\theta}_t' (\boldsymbol{I} - \mathrm{diag}(\boldsymbol{\theta}_t)) \boldsymbol{\theta}_t \tag{I.17}$$

Let random variable $y_t = \frac{1}{N}\sum_{i=1}^{K} Y_{it}$ denote the frequency of allele 1 in the sample, then

$$\mathrm{E}(y_t \mid \boldsymbol{f}_t) = \boldsymbol{\theta}_t' \boldsymbol{f}_t \tag{I.18}$$

$$e_t^2 := \mathrm{Var}(y_t \mid \boldsymbol{f}_t) = \frac{1}{N}\boldsymbol{\theta}_t'(\boldsymbol{I} - \mathrm{diag}(\boldsymbol{\theta}_t))\boldsymbol{f}_t. \tag{I.19}$$

If the sample size N is sufficiently large, by central limit theorem, the distribution of $y_t \mid \boldsymbol{f}_t, \boldsymbol{\theta}_t$ can be approximated by a normal distribution,

$$y_t \sim \mathrm{N}(\boldsymbol{\theta}_t'\boldsymbol{f}_t \ , \ \frac{1}{N}\boldsymbol{\theta}_t'(\boldsymbol{I} - \mathrm{diag}(\boldsymbol{\theta}_t))\boldsymbol{f}_t). \tag{I.20}$$

In form of observation equation, this relationship can be written as

$$y_t = \boldsymbol{\theta}_t'\boldsymbol{f}_t + u_t, \quad u_t \sim \mathrm{N}(0 \ , \ e_t^2) \tag{I.21}$$

where u_t models zero-mean observational error.

I.4 Inference of Untyped SNP Frequencies

Given model parameters $\{(r_t, \boldsymbol{\alpha}_t, \boldsymbol{\theta}_t) : t = 1, ..., T\}$ and observed frequencies $\boldsymbol{Y} = (y_1, ..., y_T)$, the objective of state-space smoothing is to calculate the conditional distributions of latent cluster membership configurations given all the observations $\boldsymbol{Y}$. Analogous to HMM, the smoothing in state-space model is typically through a forward-backward procedure. Because the desired conditional distributions are Gaussian, we only need to keep track of the conditional mean and variance during the smoothing procedure. This property makes the computation highly efficient.

The widely used smoothing alogrithm for sate-space model is Kalman smoother which requires variance structures in both state and observation equations are non-singular. However, this requirement cannot be satisfied for our model. Alternatively, we apply the generalized fixed-interval state smoother proposed by De Jong (1989), this algorithm tolerates singular covariance structure in the state equation.

REFERENCES

Bravata, D. M. and Olkin, I. (2001). Simple Pooling versus Combining in Meta-Analysis. *Evaluation & the Health Professions*, **24**(2), 218–230.

Browning, S. and Browning, B. (2007). Rapid and accurate haplotype phasing and missing data inference for whole genome association studies using localized haplotype clustering. *American Journal of Human Genetics*, **81**, 1084–1097.

Butler, R. (2007). *Saddlepoint Approximations with Applications*. Cambridge University Press, 1st edition.

Butler, R. W. and Wood, A. T. A. (2002). Laplace approximations for hypergeometric functions with matrix argument. *The Annals of Statistics*, **30**(4), 1155–1177.

Cappe, O., Robert, C. P., and Ryden, T. (2003). Reversible jump, birth-and-death and more general continuous time Markov chain Monte Carlo samplers. *Journal of the Royal Statistical Society - Series B: Statistical Methodology*, **65**(3), 679–700.

Clayton, D., Chapman, J., and Cooper, J. (2004). Use of unphased multilocus genotype data in indirect association studies. *Genetic Epidemiology*, **27(4)**, 415–428.

de Bakker, P., Yelensky, R., Pe'er, I., Gabriel, S., Daly, M., and Altshuler, D. (2005). Efficiency and power in genetic association studies. *Nature Genetics*, **37(11)**, 1217–1223.

De Jong, P. (1989). Smoothing and Interpolation with the State-Space Model. *Journal of the American Statistical Association*, **84**(408), 1085–1088.

Dimas, A. S., Deutsch, S., Stranger, B. E., Montgomery, S. B., Borel, C., Attar-Cohen, H., Ingle, C., Beazley, C., Gutierrez Arcelus, M., Sekowska, M., Gagnebin, M., Nisbett, J., Deloukas, P., Dermitzakis, E. T., and Antonarakis, S. E. (2009). Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science (New York, N.Y.)*, **325**(5945), 1246–50.

Durbin, R. M., Altshuler, D. L., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Collins, F. S., De La Vega, F. M., Donnelly, P., Egholm, M., Flicek, P., Gabriel, S. B., Gibbs, R. A., Knoppers, B. M., Lander, E. S., Lehrach, H., Mardis, E. R., McVean, G. A., Nickerson, D. A., Peltonen, L., Schafer, A. J., Sherry, S. T., Wang, J., Wilson, R. K., Deiros, D., Metzker, M., Muzny, D., Reid, J., Wheeler, D., Li, J., Jian, M., Li, G., Li, R., Liang, H., Tian, G., Wang, B., Wang, J., Wang, W., Yang, H., Zhang, X., Zheng, H., Ambrogio, L., Bloom, T., Cibulskis, K., Fennell, T. J., Jaffe, D. B., Shefler, E., Sougnez, C. L., Gormley, N., Humphray, S., Kingsbury, Z., Koko-Gonzales, P., Stone, J., McKernan, K. J., Costa, G. L., Ichikawa, J. K., Lee, C. C., Sudbrak, R., Borodina, T. A., Dahl, A., Davydov, A. N.,

Marquardt, P., Mertes, F., Nietfeld, W., Rosenstiel, P., Schreiber, S., Soldatov, A. V., Timmermann, B., Tolzmann, M., Affourtit, J., Ashworth, D., Attiya, S., Bachorski, M., Buglione, E., Burke, A., Caprio, A., Celone, C., Clark, S., Conners, D., Desany, B., Gu, L., Guccione, L., Kao, K., Kebbel, A., Knowlton, J., Labrecque, M., McDade, L., Mealmaker, C., Minderman, M., Nawrocki, A., Niazi, F., Pareja, K., Ramenani, R., Riches, D., Song, W., Turcotte, C., Wang, S., Dooling, D., Fulton, L., Fulton, R., Weinstock, G., Burton, J., Carter, D. M., Churcher, C., Coffey, A., Cox, A., Palotie, A., Quail, M., Skelly, T., Stalker, J., Swerdlow, H. P., Turner, D., De Witte, A., Giles, S., Bainbridge, M., Challis, D., Sabo, A., Yu, F., Yu, J., Fang, X., Guo, X., Li, Y., Luo, R., Tai, S., Wu, H., Zheng, H., Zheng, X., Zhou, Y., Marth, G. T., Garrison, E. P., Huang, W., Indap, A., Kural, D., Lee, W.-P., Fung Leong, W., Quinlan, A. R., Stewart, C., Stromberg, M. P., Ward, A. N., Wu, J., Lee, C., Mills, R. E., Shi, X., Daly, M. J., DePristo, M. A., Ball, A. D., Banks, E., Browning, B. L., Garimella, K. V., Grossman, S. R., Handsaker, R. E., Hanna, M., Hartl, C., Kernytsky, A. M., Korn, J. M., Li, H., Maguire, J. R., McCarroll, S. A., McKenna, A., Nemesh, J. C., Philippakis, A. A., Poplin, R. E., Price, A., Rivas, M. A., Sabeti, P. C., Schaffner, S. F., Shlyakhter, I. A., Cooper, D. N., Ball, E. V., Mort, M., Phillips, A. D., Stenson, P. D., Sebat, J., Makarov, V., Ye, K., Yoon, S. C., Bustamante, C. D., Boyko, A., Degenhardt, J., Gravel, S., Gutenkunst, R. N., Kaganovich, M., Keinan, A., Lacroute, P., Ma, X., Reynolds, A., Clarke, L., Cunningham, F., Herrero, J., Keenen, S., Kulesha, E., Leinonen, R., McLaren, W. M., Radhakrishnan, R., Smith, R. E., Zalunin, V., Zheng-Bradley, X., Korbel, J. O., Stütz, A. M., Bauer, M., Keira Cheetham, R., Cox, T., Eberle, M., James, T., Kahn, S., Murray, L., Ye, K., Fu, Y., Hyland, F. C. L., Manning, J. M., McLaughlin, S. F., Peckham, H. E., Sakarya, O., Sun, Y. A., Tsung, E. F., Batzer, M. A., Konkel, M. K., Walker, J. A., Albrecht, M. W., Amstislavskiy, V. S., Herwig, R., Parkhomchuk, D. V., Agarwala, R., Khouri, H. M., Morgulis, A. O., Paschall, J. E., Phan, L. D., Rotmistrovsky, K. E., Sanders, R. D., Shumway, M. F., Xiao, C., Auton, A., Iqbal, Z., Lunter, G., Marchini, J. L., Moutsianas, L., Myers, S., Tumian, A., Knight, J., Winer, R., Craig, D. W., Beckstrom-Sternberg, S. M., Christoforides, A., Kurdoglu, A. A., Pearson, J. V., Sinari, S. A., Tembe, W. D., Haussler, D., Hinrichs, A. S., Katzman, S. J., Kern, A., Kuhn, R. M., Przeworski, M., Hernandez, R. D., Howie, B., Kelley, J. L., Cord Melton, S., Li, Y., Anderson, P., Blackwell, T., Chen, W., Cookson, W. O., Ding, J., Min Kang, H., Lathrop, M., Liang, L., Moffatt, M. F., Scheet, P., Sidore, C., Snyder, M., Zhan, X., Zöllner, S., Awadalla, P., Casals, F., Idaghdour, Y., Keebler, J., Stone, E. A., Zilversmit, M., Jorde, L., Xing, J., Eichler, E. E., Aksay, G., Alkan, C., Hajirasouliha, I., Hormozdiari, F., Kidd, J. M., Cenk Sahinalp, S., Sudmant, P. H., Chen, K., Chinwalla, A., Ding, L., Koboldt, D. C., McLellan, M. D., Wallis, J. W., Wendl, M. C., Zhang, Q., Albers, C. A., Ayub, Q., Balasubramaniam, S., Barrett, J. C., Chen, Y., Conrad, D. F., Danecek, P., Dermitzakis, E. T., Hu, M., Huang, N., Hurles, M. E., Jin, H., Jostins, L., Keane, T. M., Quang Le, S., Lindsay, S., Long,

Q., MacArthur, D. G., Montgomery, S. B., Parts, L., Tyler-Smith, C., Walter, K., Zhang, Y., Gerstein, M. B., Snyder, M., Abyzov, A., Balasubramanian, S., Bjornson, R., Du, J., Grubert, F., Habegger, L., Haraksingh, R., Jee, J., Khurana, E., Lam, H. Y. K., Leng, J., Jasmine Mu, X., Urban, A. E., Zhang, Z., Coafra, C., Dinh, H., Kovar, C., Lee, S., Nazareth, L., Wilkinson, J., Coffey, A., Scott, C., Gharani, N., Kaye, J. S., Kent, A., Li, T., McGuire, A. L., Ossorio, P. N., Rotimi, C. N., Su, Y., Toji, L. H., Brooks, L. D., Felsenfeld, A. L., McEwen, J. E., Abdallah, A., Juenger, C. R., Clemm, N. C., Duncanson, A., Green, E. D., Guyer, M. S., and Peterson, J. L. (2010). A map of human genome variation from population-scale sequencing. *Nature*, **467**(7319), 1061–1073.

Guan, Y. and Stephens, M. (2008). Practical issues in imputation-based association mapping. *PLoS Genetics*, **4(12)**, e1000279.

Guan, Y. and Stephens, M. (2011). Bayesian Variable Selection Regression for Genome-wide Association Studies, and other Large-Scale Problems. *Annals of Applied Statistics*, page in press.

Homer, N., Tembe, W., Szelinger, S., Redman, M., Stephan, D., Pearson, J., Nelson, D., and Craig, D. (2008a). Multimarker analysis and imputation of multiple platform pooling-based genome-wide association studies. *Bioinformatics*, **24(17)**, 1896–1902.

Homer, N., Szelinger, S., Redman, M., Duggan, D., Tembe, W., Muehling, J., Pearson, J., Stephan, D., Nelson, S., and Craig, D. (2008b). Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays. *PLoS Genetics*, **4(8)**, e1000167.

Howie, B., Donnelly, P., and Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genetics*, **5(6)**, e1000529.

Huang, L., Li, Y., Singleton, A., Hardy, J., Abecasis, G., Rosenberg, N., and Scheet, P. (2009). Genotype-imputation accuracy across worldwide human populations. *American Journal of Human Genetics*, **84(2)**, 235–250.

Hudson, R. (2001). Two-locus sampling distributions and their application. *Genetics*, **159(4)**, 1805–1817.

Hunter, D. J. (2005). Gene-environment interactions in human diseases. *Nature reviews. Genetics*, **6**(4), 287–98.

Johnson, V. E. (2005). Bayes factors based on test statistics. *Journal of the Royal Statistical Society - Series B: Statistical Methodology*, **67**(5), 689–701.

Johnson, V. E. (2008). Properties of Bayes Factors Based on Test Statistics. *Scandinavian Journal of Statistics*, **35**(2), 354–368.

Kong, A., Thorleifsson, G., Stefansson, H., Masson, G., Helgason, A., Gudbjartsson, D. F., Jonsdottir, G. M., Gudjonsson, S. A., Sverrisson, S., Thorlacius, T., Jonasdottir, A., Hardarson, G. A., Palsson, S. T., Frigge, M. L., Gulcher, J. R., Thorsteinsdottir, U., and Stefansson, K. (2008). Sequence variants in the RNF212 gene associate with genome-wide recombination rate. *Science (New York, N.Y.)*, **319**(5868), 1398–401.

Kudaravalli, S., Veyrieras, J.-B., Stranger, B. E., Dermitzakis, E. T., and Pritchard, J. K. (2009). Gene expression levels are a target of recent natural selection in the human genome. *Molecular biology and evolution*, **26**(3), 649–58.

Lebrec, J. J., Stijnen, T., and Houwelingen, H. C. V. (2010). Statistical Applications in Genetics and Molecular Biology Dealing with Heterogeneity between Cohorts in Genomewide SNP Association Studies Dealing with Heterogeneity between Cohorts in Genomewide SNP Association Studies. *Statistical Applications in Genetics and Molecular Biology*, **9**(1).

Li, N. and Stephens, M. (2003). Modelling linkage disequilibrium and identifying recombination hotspots using snp data. *Genetics*, **165**, 2213–2233.

Li, Y., Ding, J., and Abecasis, G. (2006). Mach 1.0: Rapid haplotype reconstruction and missing genotype inference. *American Journal of Human Genetics*, **79**, S2290.

Marchini, J., Howie, B., Myers, S., McVean, G., and Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genetics*, **39(7)**, 906–913.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*. Chapman and Hall, London.

McVean, G., Awadalla, P., and Fearnhead, P. (2002). A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics*, **160(3)**, 1231–1241.

Meaburn, E., Butcher, L., Schalkwyk, L., and Plomin, R. (2006). Genotyping pooled DNA using 100k snp microarrays: a step towards genomewide association scans. *Nucleic Acids Research*, **34(4)**, e28.

Meng, X. and Rubin, D. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, **80(2)**, 267278.

Montgomery, S. B. and Dermitzakis, E. T. (2011). From expression QTLs to personalized transcriptomics. *Nature Reviews Genetics*, **12**(4), 277–282.

Munafò, M. R. and Flint, J. (2004). Meta-analysis of genetic association studies. *Trends in genetics : TIG*, **20**(9), 439–44.

Nica, A. C., Parts, L., Glass, D., Nisbet, J., Barrett, A., Sekowska, M., Travers, M., Potter, S., Grundberg, E., Small, K., Hedman, A. s. K., Bataille, V., Tzenova Bell, J., Surdulescu, G., Dimas, A. S., Ingle, C., Nestle, F. O., di Meglio, P., Min, J. L., Wilk, A., Hammond, C. J., Hassanali, N., Yang, T.-P., Montgomery, S. B., O'Rahilly, S., Lindgren, C. M., Zondervan, K. T., Soranzo, N., Barroso, I., Durbin, R., Ahmadi, K., Deloukas, P., McCarthy, M. I., Dermitzakis, E. T., and Spector, T. D. (2011). The Architecture of Gene Regulatory Variation across Multiple Human Tissues: The MuTHER Study. *PLoS Genetics*, **7**(2), e1002003.

Nicolae, D. (2006a). Quantifying the amount of missing information in genetic association studies. *Genetic Epidemiology*, **30(8)**, 703–717.

Nicolae, D. (2006b). Testing untyped alleles (tuna)-applications to genome-wide association studies. *Genetic Epidemiology*, **30(8)**, 718–727.

Ober, C., Loisel, D. A., and Gilad, Y. (2008). Sex-specific genetic architecture of human disease. *Nature reviews. Genetics*, **9**(12), 911–22.

Owen, A. B. (2009). Karl Pearsons meta-analysis revisited. *The Annals of Statistics*, **37**(6B), 3867–3892.

Pickrell, J. K., Marioni, J. C., Pai, A. A., Degner, J. F., Engelhardt, B. E., Nkadori, E., Veyrieras, J.-B., Stephens, M., Gilad, Y., and Pritchard, J. K. (2010). Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, **464**(7289), 768–72.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M., Bender, D., Maller, J., Sklar, P., de Bakker, P., Daly, M., and Sham, P. (2007). Plink: a toolset for whole-genome association and population-based linkage analysis. *American Journal of Human Genetics*, **81(3)**, 559–575.

Sankararaman, S., Obozinski, G., Jordan, M., and Halperin, E. (2009). Genomic privacy and limits of individual detection in a pool. *Nature Genetics*, **Epub**.

Scheet, P. and Stephens, M. (2005). A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotype phase. *American Journal of Human Genetics*, **78**, 629–644.

Servin, B. and Stephens, M. (2007). Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS genetics*, **3**(7), e114.

Servin, B. and Stephens, M. (2008). Imputation-based analysis of association studies: Candidate regions and quantitative traits. *PLoS Genetics*, **3(7)**, e114.

Stephens, M. and Balding, D. J. (2009). Bayesian statistical methods for genetic association studies. *Nature reviews. Genetics*, **10**(10), 681–90.

Stephens, M. and Scheet, P. (2005). Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *American Journal of Human Genetics*, **76**, 449–462.

Stephens, M., Smith, N., and Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics*, **68(4)**, 978–989.

Stranger, B. E., Nica, A. C., Forrest, M. S., Dimas, A., Bird, C. P., Beazley, C., Ingle, C. E., Dunning, M., Flicek, P., Koller, D., Montgomery, S., Tavaré, S., Deloukas, P., and Dermitzakis, E. T. (2007). Population genomics of human gene expression. *Nature genetics*, **39**(10), 1217–24.

Teslovich, T. M., Musunuru, K., Smith, A. V., Edmondson, A. C., Stylianou, I. M., Koseki, M., Pirruccello, J. P., Ripatti, S., Chasman, D. I., Willer, C. J., Johansen, C. T., Fouchier, S. W., Isaacs, A., Peloso, G. M., Barbalic, M., Ricketts, S. L., Bis, J. C., Aulchenko, Y. S., Thorleifsson, G., Feitosa, M. F., Chambers, J., Orho-Melander, M., Melander, O., Johnson, T., Li, X., Guo, X., Li, M., Shin Cho, Y., Jin Go, M., Jin Kim, Y., Lee, J.-Y., Park, T., Kim, K., Sim, X., Twee-Hee Ong, R., Croteau-Chonka, D. C., Lange, L. a., Smith, J. D., Song, K., Hua Zhao, J., Yuan, X., Luan, J., Lamina, C., Ziegler, A., Zhang, W., Zee, R. Y. L., Wright, A. F., Witteman, J. C. M., Wilson, J. F., Willemsen, G., Wichmann, H.-E., Whitfield, J. B., Waterworth, D. M., Wareham, N. J., Waeber, G., Vollenweider, P., Voight, B. F., Vitart, V., Uitterlinden, A. G., Uda, M., Tuomilehto, J., Thompson, J. R., Tanaka, T., Surakka, I., Stringham, H. M., Spector, T. D., Soranzo, N., Smit, J. H., Sinisalo, J., Silander, K., Sijbrands, E. J. G., Scuteri, A., Scott, J., Schlessinger, D., Sanna, S., Salomaa, V., Saharinen, J., Sabatti, C., Ruokonen, A., Rudan, I., Rose, L. M., Roberts, R., Rieder, M., Psaty, B. M., Pramstaller, P. P., Pichler, I., Perola, M., Penninx, B. W. J. H., Pedersen, N. L., Pattaro, C., Parker, A. N., Pare, G., Oostra, B. a., ODonnell, C. J., Nieminen, M. S., Nickerson, D. a., Montgomery, G. W., Meitinger, T., McPherson, R., McCarthy, M. I., McArdle, W., Masson, D., Martin, N. G., Marroni, F., Mangino, M., Magnusson, P. K. E., Lucas, G., Luben, R., Loos, R. J. F., Lokki, M.-L., Lettre, G., Langenberg, C., Launer, L. J., Lakatta, E. G., Laaksonen, R., Kyvik, K. O., Kronenberg, F., König, I. R., Khaw, K.-T., Kaprio, J., Kaplan, L. M., Johansson, A. s., Jarvelin, M.-R., Cecile J. W. Janssens, a., Ingelsson, E., Igl, W., Kees Hovingh, G., Hottenga, J.-J., Hofman, A., Hicks, A. a., Hengstenberg, C., Heid, I. M., Hayward, C., Havulinna, A. S., Hastie, N. D., Harris, T. B., Haritunians, T., Hall, A. S., Gyllensten, U., Guiducci, C., Groop, L. C., Gonzalez, E., Gieger, C., Freimer, N. B., Ferrucci, L., Erdmann, J., Elliott, P., Ejebe, K. G., Döring, A., Dominiczak, A. F., Demissie, S., Deloukas, P., de Geus, E. J. C., de Faire, U., Crawford, G., Collins, F. S., Chen, Y.-d. I.,

Caulfield, M. J., Campbell, H., Burtt, N. P., Bonnycastle, L. L., Boomsma, D. I., Boekholdt, S. M., Bergman, R. N., Barroso, I., Bandinelli, S., Ballantyne, C. M., Assimes, T. L., Quertermous, T., Altshuler, D., Seielstad, M., Wong, T. Y., Tai, E.-S., Feranil, A. B., Kuzawa, C. W., Adair, L. S., Taylor Jr, H. a., Borecki, I. B., Gabriel, S. B., Wilson, J. G., Holm, H., Thorsteinsdottir, U., Gudnason, V., Krauss, R. M., Mohlke, K. L., Ordovas, J. M., Munroe, P. B., Kooner, J. S., Tall, A. R., Hegele, R. a., Kastelein, J. J., Schadt, E. E., Rotter, J. I., Boerwinkle, E., Strachan, D. P., Mooser, V., Stefansson, K., Reilly, M. P., Samani, N. J., Schunkert, H., Cupples, L. A., Sandhu, M. S., Ridker, P. M., Rader, D. J., van Duijn, C. M., Peltonen, L., Abecasis, G. R., Boehnke, M., and Kathiresan, S. (2010). Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*, **466**(7307), 707–713.

The International HapMap Consortium (2005). A haplotype map of the human genome. *Nature*, **437**, 1299–1320.

Veyrieras, J.-B., Kudaravalli, S., Kim, S. Y., Dermitzakis, E. T., Gilad, Y., Stephens, M., and Pritchard, J. K. (2008). High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLoS genetics*, **4**(10), e1000214.

Wakefield, J. (2009). Bayes factors for genome-wide association studies: comparison with P-values. *Genetic epidemiology*, **33**(1), 79–86.

Weir, B. (1979). Inferences about linkage disequilibrium. *Biometrics*, **35(1)**, 235–254.

Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447(7145)**, 661–678.

West, M. and Harrison, J. (1997). *Bayesian Forecasting and Dynamic Models*. Springer-Verlag, New York, 2nd edition.

Willer, C. J., Li, Y., and Abecasis, G. R. (2010). METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics (Oxford, England)*, **26**(17), 2190–1.

Yu, Z. and Schaid, D. (2007). Methods to impute missing genotypes for population data. *Human Genetics*, **122**, 495–504.

Zeggini, E., Scott, L. J., Saxena, R., Voight, B. F., Marchini, J. L., Hu, T., de Bakker, P. I., Abecasis, G. R., Almgren, P., Andersen, G., Ardlie, K., Boström, K. B., Bergman, R. N., Bonnycastle, L. L., Borch-Johnsen, K., Burtt, N. P., Chen, H., Chines, P. S., Daly, M. J., Deodhar, P., Ding, C.-J., Doney, A. S. F., Duren, W. L., Elliott, K. S., Erdos, M. R., Frayling, T. M., Freathy, R. M., Gianniny, L., Grallert, H., Grarup, N., Groves, C. J., Guiducci, C., Hansen, T., Herder, C., Hitman, G. A.,

Hughes, T. E., Isomaa, B., Jackson, A. U., Jø rgensen, T., Kong, A., Kubalanza, K., Kuruvilla, F. G., Kuusisto, J., Langenberg, C., Lango, H., Lauritzen, T., Li, Y., Lindgren, C. M., Lyssenko, V., Marvelle, A. F., Meisinger, C., Midthjell, K., Mohlke, K. L., Morken, M. A., Morris, A. D., Narisu, N., Nilsson, P., Owen, K. R., Palmer, C. N., Payne, F., Perry, J. R. B., Pettersen, E., Platou, C., Prokopenko, I., Qi, L., Qin, L., Rayner, N. W., Rees, M., Roix, J. J., Sandbæ k, A., Shields, B., Sjögren, M., Steinthorsdottir, V., Stringham, H. M., Swift, A. J., Thorleifsson, G., Thorsteinsdottir, U., Timpson, N. J., Tuomi, T., Tuomilehto, J., Walker, M., Watanabe, R. M., Weedon, M. N., Willer, C. J., Illig, T., Hveem, K., Hu, F. B., Laakso, M., Stefansson, K., Pedersen, O., Wareham, N. J., Barroso, I., Hattersley, A. T., Collins, F. S., Groop, L., McCarthy, M. I., Boehnke, M., and Altshuler, D. (2008). Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nature Genetics*, **40**(5), 638–645.