A Bayesian Approach for Two-Phase Designs in Regional Sequencing

Zhijian Chen* Radu V. Craiu† Shelley B. Bull‡

Abstract

Following detection of signals by genome-wide association studies (GWAS), investigators may choose to sequence some or all of the members of the GWAS sample to narrow down a set of potentially causal variants. This is known as a two-phase fine-mapping design. Additional efficiencies may be achieved if phase 2 fine mapping is carried out in multiple stages, with each stage comprised of a mutually exclusive subset. We consider a Bayesian approach to two-phase sampling that allows intermediate sampling time points. At each sampling point, we assess each sequence variant within a region by a Bayes factor that compares different genetic models, e.g., additive, dominant and recessive. For variants in which no genetic model outperforms the others, we apply Bayesian model averaging to account for genetic model uncertainty. We assess the efficiency of this two-phase design in the discovery of true functional variants and investigate the impact of sample allocation and correlation between tag and the functional on the efficiency.

Key Words: Bayes factor, fine mapping, next-generation sequencing, model averaging, two-phase design

1. Introduction

In focused studies following up reasonable hits from genome-wide association studies (GWAS), investigators may choose to comprehensively sequence a whole genomic region of interest using next generation sequencing (NGS) technologies, or to selectively sequence the region using customized technology to genotype additional SNPs. The purpose of regional re-sequencing studies is to identify potential causal variants, estimate the size of genetic effect, and characterize the mode of inheritance, i.e., genetic model. Despite the reduction in the genotyping cost per base-pair, the cost of regional sequencing in large samples is still high, and considerable savings may be gained from a two-phase design which sequences only a subset of the original sample.

Two-phase designs have been used in surveys, epidemiology and clinical trials, when a target variable is difficult or expensive to measure (e.g., Breslow and Holubkov, 1997; Breslow et al., 2009). At phase 1, a sample is drawn from the population and data on response and auxiliary variables are collected. Strata are then destined using the auxiliary variables that are correlated with the target variables. At phase 2, subjects are drawn from each stratum using a simple random sampling without replacement, and measurements of the target variables are made on the sampled subjects. Recently, two-phase stratified designs have been used in fine mapping studies for complex quantitative traits and human diseases (Chen et al., 2012; Schaid et al., 2013), in which GWAS subjects are sampled from strata, defined by the genotype categories of the GWAS tag SNP that drew attention to the genomic region, and the more expensive sequence data are collected on the sampled subjects. It was shown that there was efficiency gain when the tag SNP common and rare homozygote strata were oversampled and the heterozygote stratum was undersampled, provided that the tag SNP is highly correlated with the functional seq variant whose effect followed an additive genetic model (Chen et al., 2012).

*Lunenfeld-Tanenbaum Research Institute of Mount Sinai Hospital, Toronto, ON, Canada

†Department of Statistics, University of Toronto, Toronto, ON, Canada

‡Lunenfeld-Tanenbaum Research Institute of Mount Sinai Hospital; Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada

Several authors have explored the nature of the relationship between the GWAS tag SNP, used to identify the region of interest, and a sequence (seq) variant of interest within the region, examining how their correlation affects the ability to estimate association at the seq variant and distinguish among genetic models (Spencer et al., 2011; Vukcevic et al., 2011; Faye and Bull, 2011). Characterizing the genetic model, or mode of inheritance for a putative causal variant, even approximately, is of interest in the fine-mapping process particularly for a binary disease trait and may be informative for on-going study design.

Genetic model specification in genetic analysis is a very long-standing problem (for discussion see Vukcevic et al., 2011; Strauch et al., 2003; Stephens and Balding, 2009; Wang, 2012). The additive model, though widely used in GWAS association discovery stage, may not correctly characterize the underlying genetic mechanism for the true functional seq variant. A misspecification of the genetic model may have an undesirable impact on the subsequent analysis such as leading to biased estimator of the genetic association parameter. Furthermore, using an incorrect genetic model in a two-phase stratified design may result in a sample size allocation scheme that decreases the inferential efficiency.

Multi-stage sampling designs may be adapted by large-scale studies in order to analyze data and discover variants sequentially. One strategy to deal with genetic model specification is to first localize the function seq variant and gain knowledge about the genetic model, through analyzing the first batch of sequence data, and then modify the genetic model and sample size allocation scheme at subsequent stages. Although the theory of sequential sampling was instrumental in the development of the pedigree lod score in linkage analysis (Province, 2001), sequential analysis procedures and adaptive designs have received only a modicum of attention (see also Province, 2000; Konig et al., 2001, 2003; Bull et al., 2002; Scherag et al., 2003, 2009; Yan et al., 2008). For the most part, sequential methods in genetic analysis have focused on hypothesis testing and marker detection, rather than on parameter estimation and genetic model inference.

In this paper, we consider the multi-stage phase 2 design in the fine mapping of a GWAS-identified region for a complex disease. To incorporate prior biological information or belief, we implement the seq variant analysis and the sampling strategies within the Bayesian paradigm; but to deal with model uncertainty we adapt a model averaging approach. A set of relevant non-genetic covariates can be specified according to subject-matter knowledge. The model uncertainty in single-variant analysis is associated with the choice of the genetic model rather than with covariate selection. Therefore, in the work presented here we do not include other covariate in the model.

The rest of this paper is organized as follows. In Section 2, we describe the formulation of the regression model and the data structure from a multi-stage two-phase design. In Section 3, we present the Bayesian methods for the inference of genetic association under genetic model specification as well as under model averaging. In Section 4, we conduct simulation studies to assess the validity of the two-phase sampling design and the Bayesan model averaging approach. Section 5 includes concluding remarks.

2. General Model and Data

2.1 Genetic models

Let Y denote the status of a complex disease, $Y = 1$ if diseased and $= 0$ otherwise. We assume that Y is Bernoulli distributed with probability of success $P(Y = 1) = \mu$. Let X count the number of risk/minor allele in the functional variant genotype. Without loss of generality we consider single-variant regression models for Y. The most widely used

genetic model is the additive model

$$\text{logit}(\mu) = \beta_0 + \beta_1 X,$$

where $\text{logit}(u) = \exp(u)/\{1+\exp(u)\}$ is the logit link function relating the linear predictor to the risk of disease, β_1 measures the increase or disease in the value of the trait with each additional copy of the risk allele at the functional variant.

Although the additive model is frequently used in GWAS with tag SNPs, the genetic model at the functional variants may not be truly additive. An extension of the the simple additive model is to include an extra parameter that models the deviation from the simple additivity at the heterozygote. Following Vukcevic et al. (2011), we use the general three-parameter model

$$\text{logit}(\mu) = \beta_0 + \beta_1 X + \gamma 1_{X=1},$$

where $1_{X=1}$ is an indicator function that takes value 1 for heterozygotes and 0 for homozygotes, and γ models the deviation from an additive model at the heterozygote and is referred to as the dominance parameter. In this paper we specifically consider two other commonly used models: dominant and recessive models, in additional to the additive model. All three genetic models are special cases of the general codominant model and can be recovered by setting the dominance parameter to specific values. For example, $\gamma = 0$ gives the additive model, $\gamma = \beta_1$ gives the dominant model, and $\gamma = -\beta_1$ gives the recessive model. One can see that all three genetic models involve only a single genetic effect size parameter β_1. We denote the additive, dominant, and recessive genetic models as M_1, M_2 and M_3, respectively. Let $\theta = (\beta_0, \beta_1)^T$ denote a vector of model parameters.

2.2 Phase 2 sequential sampling

Because of the low density of GWAS SNPs relative to seq variants, the GWAS hit identified in phase 1 is very likely a spurious association resulted from high LD with a true functional seq variant. Let N be the number of phase 1 subjects. Let G denote the genotype of a GWAS tag SNP, with minor allele a and reference allele A, that appears as a hit and has drawn attention to the region harbouring the functional variant. For subject i in phase 1 sample, a pair of observed values (G_i, Y_i) is available, $i = 1, \ldots, N$. The sample is then divided into three strata according to the three categories of G: common homozygote AA, heterozygote Aa, and rare homozygote aa. At phase 2, subjects are sampled randomly from each strata. Fine mapping of the interesting region is then conducted using the sequence data from the phase 2 sample.

Let K be the total number of stages (periods) specified for phase 2 sampling. For simplicity, we assume that the same number of m subjects are sampled without replacement at each sampling period. Nevertheless, the proposed method can be generalized to cases with varying sample size. Let $n = Km$ ($n < N$) be the total phase 2 sample size one can collect, which is usually predetermined by budget limits. Let $Y^{(k)}$ and $X^{(k)}$ denote the response and seq genotype data at kth sampling stage, $k = 1, \ldots, K$.

The goal of the sequential sampling is to gradually localize the function seq variant by analyzing all variants in the region and draw reliable conclusion about the genetic association. Although valid results can be obtained by analyzing available data, the variability in the inference of the association, however, is influenced by the number of minor alleles at the functional seq variant in phase 2 sample. To distinguish among genetic models, we require that the sequence data obtained in the first period contain a reasonable amount of information on all three genotypes particularly the heterozygote. This amounts to select subjects such that the first period sample contains approximately equal number of subjects carrying 0, 1 and 2 copies of the minor alleles at the function seq variant. In practice, if a

tag SNP has drawn attention to a specific region, it is reasonable to assume that the tag SNP is in high LD with the functional seq variant in the same region. Therefore, one sampling scheme is to select the same number of subjects from each of the three tag genotype-defined strata.

3. Bayesian Inference

3.1 Model-specific inference

Let $p(M_j)$ be the prior probability for genetic model M_j, $j = 1, 2, 3$. Let $p(\theta \mid M_j)$ be the prior distribution of the population parameters θ under model M_j. Often, independence between the priors for θ and the underlying genetic model is assumed, i.e., $p(\theta \mid M_j) = p(\theta)$. The priors specified for θ can undesirably tilt the prior distribution for disease prevalence. Under a cohort design at phase 1, the marginal probability of an event is specified to follow a uniform distribution in the interval (0,1), representing a non-informative prior belief on the prevalence of disease (see Evans and Jang, 2012; Baskurt and Evans, 2011, for additional details). For the logit link, this is equivalent to specifying an approximate standard logistic distribution for the linear predictor.

GWASs of complex diseases that have low prevalence in the population often employ case-control designs, because cohort designs may require long observation periods in order to observe enough number of cases and may be undesirable. Implementing the same Bayesian prospective analysis used for a cohort study may not be equivalent to the appropriate retrospective analysis if the same priors are specified. This is unfortunate as retrospective analyses tend to pose greater computational challenges in case-control logistic models (e.g., Craiu et al., 2011). One possible escape route has been found by Seaman and Richardson (2004) who showed that a case-control Bayesian analysis involving a prospective likelihood and a uniform prior for the log odds parameter is equivalent to an analysis that uses a retrospective likelihood along with a Dirichlet prior distribution for the exposure probabilities in the control group. To properly model the genetic association at the seq variant within a Bayesian 2 phase framework, we adopt the Seamam and Richardson approach by specifying a uniform prior for β_0 (the population log odds of disease at baseline genotype) and a normal prior for the log odds parameter β_1 associated with the genetic effect. The equivalence is valid for the inference about β_1 only but not for β_0.

The selection of phase 2 subjects involves a sample allocation scheme that determines the probability of each phase 1 subject being included. Note that the inclusion probabilities or sampling weights can be ignored in the Bayesian analysis. Specifically, the tag SNP genotype, the factor used for stratifying the phase 1 sample, is conditionally independent of the response given the functional variant despite that the tag SNP is correlated with the functional variant. Let R_i be the phase 2 sampling indicator for subject i. Then, the observed genetic model-specific likelihood contributed by subject i is

$$p(Y_i, X_i, R_i \mid G_i, \theta, M_j) \;\; = \;\; p(Y_i \mid X_i, G_i, \theta, M_j)\; p(X_i \mid G_i, \theta, M_j)\; p(R_i \mid G_i),$$

where $p(R_i \mid G_i)$ is the inclusion probability which is free of θ and the underlying genetic model. Disease status is independent of the tag SNP given that data on the seq variant is observed. Thus, the above equation becomes $p(Y_i, X_i, R_i \mid G_i, \theta, M_j) = p(R_i \mid G_i)\; p(Y_i \mid X_i, \theta, M_j)\; p(X_i \mid G_i)$. Because both $p(R_i \mid G_i)$ and $p(X_i \mid G_i)$ are free of θ and M_j and do not contribute to the calculation of the posterior for θ, the analysis is simplified by ignoring sampling weights and treating the observed data as from a cross-sectional study.

At each phase 2 sampling period, cumulated data up to and including current period are analyzed and the posterior for θ and the posterior weights of genetic models are updated.

To avoid excessive notation induced by sequential sampling design, we focus on describing the computation using phase 2 data from the first sampling period. The analysis can be generalized to subsequent sampling stages.

The posterior for θ given data $(Y^{(1)}, X^{(1)})$ at the first period and a specific genetic model, say, M_j, is

$$p(\theta \mid Y^{(1)}, X^{(1)}, M_j) = \frac{p(Y^{(1)} \mid X^{(1)}, \theta, M_j)\; p(\theta \mid M_j)\; p(M_j)}{p(Y^{(1)} \mid X^{(1)}, M_j)},$$

where $p(Y^{(1)} \mid X^{(1)}, M_j) = \int p(Y^{(1)} \mid X^{(1)}, \theta, M_j)\; p(\theta \mid M_j)\; p(M_j)\; d\theta$ is the normalizing constant of the posterior distribution. Let $\mu_j^{(1)} = E(\theta \mid Y^{(1)}, X^{(1)}, M_j)$ be the posterior mean of θ, which is given by $\mu_j^{(1)} = \int \theta\; p(\theta \mid Y^{(1)}, X^{(1)}, M_j)\; d\theta$. The posterior variance is

$$\mathrm{var}(\theta \mid Y^{(1)}, X^{(1)}, M_j) = \int \{\theta - \mu_j^{(1)}\}\; \{\theta - \mu_j^{(1)}\}^T\; p(\theta \mid Y^{(1)}, X^{(1)}, M_j)\; d\theta,$$

respectively.

Posterior weights of genetic models can be computed by comparing between genetic models. The Bayes factor between genetic models $M_{j'}$ and M_j is defined by

$$BF_{j'j}^{(1)} = \frac{p(Y^{(1)} \mid X^{(1)}, M_{j'})\; p(M_{j'})}{p(Y^{(1)} \mid X^{(1)}, M_j)\; p(M_j)},$$

where M_j is treated as the reference. Whenever $p(M_{j'}) = p(M_j)$, i.e., in the case of each model being equally likely *a priori*, the Bayes factor is simply the ratio of the two normalizing constants of the posteriors under M_j and $M_{j'}$. Let $\pi_j^{(1)} = p(M_j \mid Y^{(1)}, X^{(1)})$ be the posterior weight for M_j. It can be readily shown that

$$\pi_j^{(1)} \propto \frac{1}{1 + \sum_{j' \neq j} BF_{j'j}^{(1)}}.$$

There is no closed form for $\pi_j^{(1)}$, however, due to the absence of conjugate priors. Markov chain Monte Carlo sampling techniques are typically required for the calculation of Bayes factor, as detailed in the Section 3.3.

For $k > 1$, the cumulated data up to and including the kth period of phase 2 sampling can be analyzed using our Bayesian approach as if all data were obtained in a single period. For instance, analysis using only data obtained in the second period but with an updated new priors, derived from the first period, is equivalent to the analysis that uses data from both periods with the initial priors. That is, information about θ obtained from previous periods can be conserved through updating the prior, a feature exactly what Bayesian inference is well known for. It allows us to make inference about the genetic association parameter and the genetic model in a simpler way.

3.2 Bayesian model averaging

In cases where the genetic effect at the seq variant is believed to follow one particular genetic model, one may choose the genetic model corresponding to the highest posterior weight. However, posterior weights of the genetic models are greatly influenced by prior weights in general. Even for equal prior specification, the underlying genetic model for the seq variant may not be identified if data support the three genetic models equally. Making decisions without accounting for model uncertainty may be risky, and making inference

about genetic effect size based on a single best genetic model may result in underestimation of the uncertainty in genetic association (e.g., Raftery et al., 1997), which leads to inflated type 1 error.

Instead of choosing a best genetic model, we employ a Bayesian model averaging (BMA) approach that allows uncertainty in the genetic model for the seq variant. With model averaging, the posterior for θ given the phenotype and genotype data observed in the first sampling period is $p(\theta \mid Y^{(1)}, X^{(1)}, M_j) = \sum_{j=1}^{3} p(\theta \mid Y^{(1)}, X^{(1)}, M_j)\pi_j^{(1)}$. Let $\mu_{\text{BMA}}^{(1)} = E(\theta \mid Y^{(1)}, X^{(1)})$. It is shown that $\mu_{\text{BMA}}^{(1)} = \sum_{j=1}^{3} \mu_j^{(1)}\pi_j^{(1)}$ (e.g., Hoeting et al., 1999). The BMA posterior variance is given by

$$\begin{aligned}\text{var}(\theta \mid Y^{(1)}, X^{(1)}) &= \sum_{j=1}^{3} \left(\pi_j \left[\mu_j^{(1)}\{\mu_j^{(1)}\}^T + \text{var}(\theta \mid Y^{(1)}, X^{(1)}, M_j)\right]\right) \\ &\quad + \mu_{\text{BMA}}^{(1)}\{\mu_{\text{BMA}}^{(1)}\}^T.\end{aligned}$$

One of the attractive features with this approach is that only one hypothesis test for association is needed. Unlike when separate tests are conducted with each genetic model, there is no need for multiplicity adjustment (e.g., Biswas and Papachristou, 2010).

Because the BMA approach aggregates results from analyses under the specification of individual genetic models, the BMA posterior mean can be seen to be close to the the one with the specification of model corresponding to the highest posterior weight. However, by taking uncertainty into account the BMA posterior variance can be seen to be larger than the posterior variances under individual models. Therefore, the BMA approach typically reduces association signals at all seq variants in the region.

3.3 MCMC algorithms

The posterior mean and variance of the genetic association parameter β_1 are intractable because the normalizing constant for the posterior cannot be computed exactly. We use a Markov chain Monte Carlo algorithm to obtain draws from the posterior distribution of θ. Specifically, we use the Data Augmentation (DA) algorithm (e.g. Albert and Chib, 1993; van Dyk and Meng, 2001) to sample from the posterior distributions corresponding to each genetic model. At the first sampling period, the logistic model is augmented with the auxiliary variables $Z^{(1)} = (Z_1, \ldots, Z_m)^T$ such that $Z_i \sim \text{Logistic}(\beta_0 + X_i\beta_1, 1)$, and $Y_i = 1$ if $Z_i > 0$ and $Y_i = 0$ otherwise, for all $i = 1, \ldots, m$.

The DA algorithm requires alternate sampling from the conditional distribution of the auxiliary variables given the parameter and the observed data, and from the conditional distribution of the parameter vector given the observed and augmented data. The full conditional distribution for Z_i is

$$Z_i \mid \theta, Y^{(1)}, X^{(1)}, M_j \sim \text{Truncated Logistic}(\beta_0 + X_i\beta_1, 1),$$

which can be sampled directly. The full conditional for θ is

$$p(\theta \mid Z^{(1)}, Y^{(1)}, X^{(1)}, M_j) \propto p(\theta) \prod_{i=1}^{m} p(Z_i \mid X_i, \theta),$$

which is not of standard form.

We use the random-walk Metropolis algorithm and specify the proposal density to be a bivariate normal density $\text{N}(0, \tau^2 \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right))$. The initial value for the chain is the ML estimate of θ obtained from fitting a standard logistic regression model. At each iteration, we first

generate θ^* from the proposal density and calculate, given the current draw $\theta^{(t)}$ and the simulated vector $\{Z^{(1)}\}^{(t+1)}$, the acceptance probability

$$\alpha = \min\left\{1, \frac{p(\theta^* \mid \{Z^{(1)}\}^{(t+1)}, X^{(1)}, M_j)}{p(\theta^{(t)} \mid \{Z^{(1)}\}^{(t+1)}, X^{(1)}, M_j)}\right\}.$$

Then, with probability α we accept the proposed value and set $\theta^{(t+1)} = \theta^*$ and otherwise we set $\theta^{(t+1)} = \theta^{(t)}$. Note that the unknown normalizing constants for the conditional distributions used to compute α are in fact equal and cancel out.

The first B generated samples are discarded as the burn-in samples. Remaining samples are then used to compute the posterior mean and variance of model parameters under each of the three genetic models. Because there are no analytic forms for the posterior distributions, we cannot directly obtain the posterior weight for each genetic model. Nevertheless, the required ratios of normalizing constants can be computed using the bridge sampling estimator (e.g., Meng and Wong, 1996; Gelman and Meng, 1998), and posterior weights for all three genetic models can be estimated based on the MCMC samples. Since the bridge sampling estimator is derived under the independence assumption, we suggest subsampling the MCMC output at a reasonable lag to reduce the serial correlation between draws. For instance, in our simulations we used every 20th sample produced by the algorithm.

4. Simulation Studies

4.1 Design of simulation

We conducted simulation studies to evaluate the performance of the multi-stage sampling strategy for the phase 2 fine mapping. We assumed that the phase 1 sample consisted of 2500 cases and 2500 controls. The MAF for the tag and the seq variants were specified by $P_d = P_a = 0.4$. Based on the suggestion by Vukcevic et al. (2011) that the value of r^2 will be quite high when the tag SNP has been identified in a GWAS of the same sample but may be smaller if the tagSNP was identified in a previous independent study, we specified the correlation between the two SNPs by $r^2 = 0.01, 0.80$, corresponding to an uninformative and a highly informative tag, respectively.

After generating the tag-seq haplotypes for all phase 1 subjects using a specific tag-seq r^2, we simulated phenotypes under the additive, dominant and recessive genetic models. Then, a small subset of the phase 1 sample was selected and sequence data in the region suggested by the tag SNP are collected on these phase 2 subjects. We assumed $K = 5$ sequential sampling periods for phase 2, each resulting in a subsample of size 200, for a total sample of 1000. We used equal prior weight for each genetic model, i.e., $p(M_j) = 1/3$, $j = 1, 2, 3$.

Regression parameters in the logistic model were specified by $\beta_0 = -3$ and $\beta_1 = 0.25$. The log-odds parameter β_0 gave a disease prevalence of 4.7% in the baseline genotype group. The positive value of the association parameter β_1 implied deleterious effect of the minor allele of the functional variant. Based on the arguments in Seaman and Richardson (2004) for case-control Bayesian analysis, we specified a uniform prior and a normal prior with mean 0 and standard deviation 1 for β_0 and β_1, respectively.

We generated independent candidates for β_0 and β_1 using the random walk Metropolis algorithm with a normal proposal density and standard deviation $0.25\sqrt{K/k}$. That is, the proposal's standard deviation decreased as the accumulated phase 2 sample size increased, yielding a relatively stable acceptance rate of about 27%. We specified the number of MCMC iterations to be 205,000, and discarded the first 5000 draws as burn-in. For the

calculation of the ratio of normalizing constants via bridge sampling, we retained every 20th draw so that the correlations within the Monte Carlo sample were less than 0.1.

To ensure good mixing properties of the MCMC sampling for a binary trait, in our simulations we drew samples from the posterior distribution using nine parallel chains that started at different places in the parameter space. We performed diagnostics on the samples using the method proposed by Gelman and Rubin (1992), which relies on parallel chains to test whether they all converge to the same posterior distribution.

4.2 Result

Here, we report results based on 100 independent data sets. For a random data set, the trace and autocorrelation plots show good mixing properties for the MCMC sampling from the posterior distribution of β_1 (Figure 1). The Gelman-Rubin diagnostic plots. The autocorrelation function does show a reduction of the effective sample size, but we compensate by letting the chain run longer. After about 1000 iterations in the chain, the potential scale reduction factor converged to 1, and both the estimate of the shrinkage factor and the upper limit of its 95% confidence interval were essentially 1 (Figure 2).

The posterior mean of β_1 under the true genetic model gradually approaches the nominal value but was biased when obtained under incorrect models (Figure 3). The dominant and recessive models quickly achieved the highest posterior probabilities when they were specified as the true genetic models (data not shown). When the additive model was the generating model, however, it appeared to be close to dominant and recessive models and did not emerge as the best genetic model until late sampling periods.

The proportion of simulations in which the true genetic models were identified as the best at the final sampling point is about 40% for the additive model and 75% for the dominant and recessive models. One explanation is that the phase 2 sample size and the association parameter value at the seq variant are relatively small, leading to low power to detect the difference in the disease proportions among the three seq genotype categories. As a result, the dominant or recessive model appeared to fit the data better than the additive model, even though the genetic model for the seq variant was truly additive.

Finally, we note that the main contribution of the tag-seq LD correlation to inference about the magnitude of the association is obtained through a more informative phase 2 sample that benefits the true genetic model. When correlation is high, e.g., $r^2 = 0.8$, the minor allele at the functional seq variant could be enriched in phase 2 sample through oversampling from the tag rare homozygote stratum. When correlation is low, e.g., $r^2 = 0.01$, the tag SNP was not informative for the seq variant, and the tag-based sampling would perform no better than a simple random sampling. The advantages of using a tag-based informative sampling included the reduction of variability in the posterior of genetic association, which was confirmed by a comparison between the standard deviations of β_1 obtained under high and low r^2 values. In general, standard deviations were smaller for $r^2 = 0.8$ than for $r^2 = 0.01$, regardless of the true underlying genetic model and the genetic model that was specified in the analysis (Table 1). The result for BMA also showed reduction of the variability in the association when an informative tag was used.

5. Discussion

In this paper we consider the fine mapping of a region identified by case-control GWAS by sequencing the whole region in a subsample of GWAS subjects. We consider the case in which the two-phase stratified sampling design is conducted sequentially using information

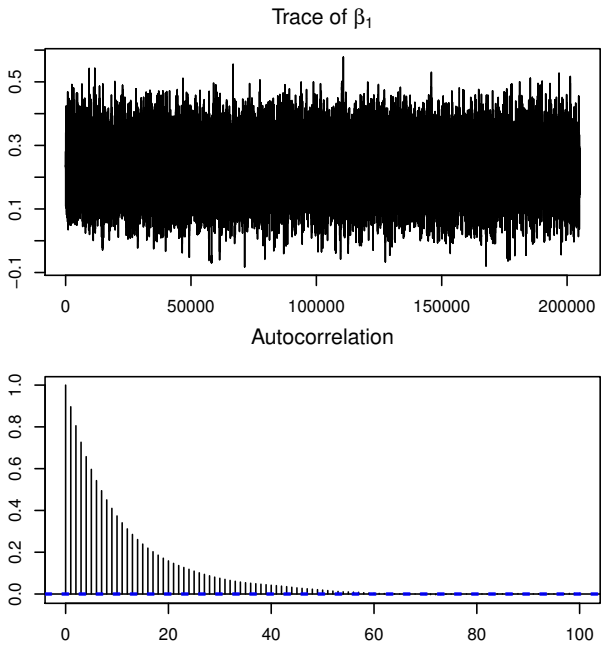

Figure 1: Trace and autocorrelation plots of the MCMC sample from the posterior distribution of β_1. (The number of MCMC iteration is 205000.)

from GWAS tag SNPs. We adapt a Bayesian approach to analyzing the seq variants and propose to use model averaging that deals with genetic model uncertainty.

In the absence of knowledge of the genetic model for a true functional variant, the effect of a candidate genetic variant is often coded additively, leading to underestimation of the association parameter when this assumption is incorrect. If a true underlying genetic model exists, adaptive sample size allocation strategy may be employed to obtain a more informative phase 2 sample with the aim of improving the localization of the functional seq variant and the inference of the genetic association parameter by accumulating evidence in favour of the true genetic model. Investigation of stopping criteria in the sequential sampling procedures appears warranted, for example, early stopping of the sampling process when sufficient evidence for association at a seq variant has accumulated or a genetic model with high posterior probability has emerged.

Acknowledgments

This research was supported by funding from the Canadian Institutes of Health Research: CIHR Operating Grant MOP-84287 (R.C., S.B.B.), CIHR Training Grant GET-101831 (Z.C). Z.C. is a CIHR Fellow in Genetic Epidemiology and Statistical Genetics with CIHR STAGE (Strategic Training for Advanced Genetic Epidemiology) - CIHR Training Grant in Genetic Epidemiology and Statistical Genetics. Computations were performed on the GPC supercomputer at the SciNet HPC Consortium. SciNet is funded by: the Canada Foundation for Innovation under the auspices of Compute Canada; the Government of Ontario; Ontario Research Fund - Research Excellence; and the University of Toronto.

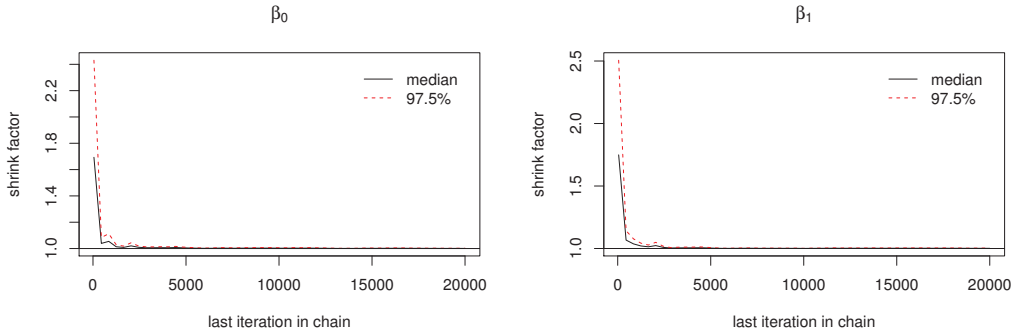

Figure 2: An illustration of the Gelman-Rubin diagnostic plots. The true genetic model is additive. Nine MCMC chains were used with each chain consisting of 25000 draws. The first 5000 draws in each chain were discarded.

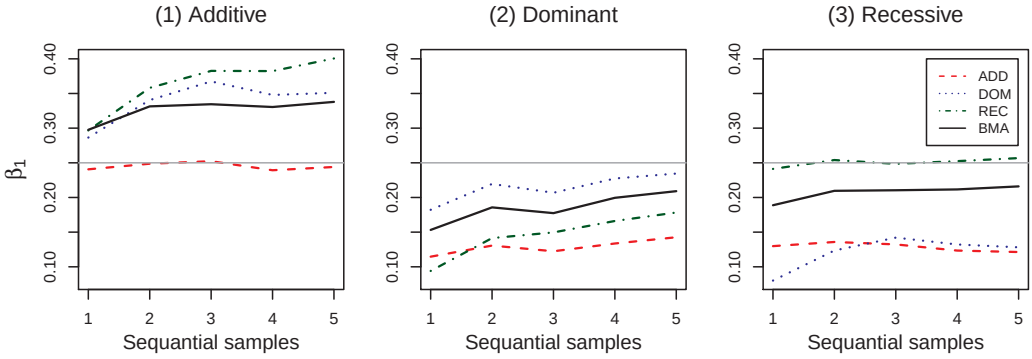

Figure 3: Sequential posterior mean of β_1. From left to right: The effect of the seq variant followed the additive, dominant and recessive genetic models, respectively. Within each panel, the dashed line, dotted line and dot-dashed line correspond to analysis under additive, dominant and recessive genetic model specifications, respectively, and the black solid line corresponds to BMA.

References

Albert, J. and Chib, S. (1993), "Bayesian analysis of binary and polychotomous response data," *J Am Stat Assoc*, 88, 669–679.

Baskurt, Z. and Evans, M. (2011), "Inequalities for Bayes Factors and Relative Belief Ratios," Tech. Rep. 1105, University of Toronto.

Biswas, S. and Papachristou, C. (2010), "Accounting for disease model uncertainty in mapping heterogeneous traits - Bayesian model averaging approach," *Hum Hered*, 69, 242–253.

Breslow, N. and Holubkov, R. (1997), "Maximum likelihood estimation of logistic regression parameters under two-phase, outcome-dependent sampling," *J R Stat Soc, B*, 59, 447–461.

Table 1: Posterior standard deviation of β_1 from the analysis of accumulated phase 2 data. Three tag-seq correlations were considered: 0.01 and 0.80. Equal prior weight is specified for the three genetic models. The first three rows correspond to analyses with additive, dominant and recessive genetic models specified for the effect at the seq variant, and the fourth row corresponds to BMA.

	Additive		Dominant		Recessive	
	tag-seq r^2		tag-seq r^2		tag-seq r^2	
Analysis	0.01	0.8	0.01	0.8	0.01	0.8
ADD	0.090	0.073	0.091	0.073	0.089	0.073
DOM	0.132	0.127	0.131	0.126	0.130	0.127
REC	0.157	0.129	0.162	0.129	0.157	0.128
BMA	0.139	0.135	0.155	0.132	0.160	0.131

Breslow, N., Lumley, T., Ballantyne, C., Chambless, L., and Kulich, M. (2009), "Improved Horvitz-Thompson estimation of model parameters from two-phase stratified samples: Applications in epidemiology," *Stat Biosci*, 1, 32–49.

Bull, S., Greenwood, C., Mirea, L., and Morgan, K. (2002), "Regression models for allele sharing: Analysis of accumulating data in affected sib pair studies," *Stat Med*, 21, 431–444.

Chen, Z., Craiu, R., and Bull, S. (2012), "Two-phase stratified sampling designs for regional sequencing," *Genet Epidemiol*, 36, 320–332.

Craiu, R., Duchesne, T., Fortin, D., and Baillargeon, S. (2011), "Conditional logistic regression with longitudinal follow up and individual-level random coefficients: A stable and efficient two-step estimation method," *J Comput Graph Stat*, 20, 767–784.

Evans, M. and Jang, G. (2012), "Weak informativity and the information in one prior relative to another," *Stat Sci*, 26, 423–439.

Faye, L. and Bull, S. (2011), "Two-stage study designs combining genome-wide association studies, tag single-nucleotide polymorphisms, and exome sequencing: accuracy of genetic effect estimates," *BMC Proc*, 5, 1–7.

Gelman, A. and Meng, X. (1998), "Simulating normalizing constants: From importance sampling to bridge sampling to path sampling," *Stat Sci*, 13, 163–185.

Gelman, A. and Rubin, D. (1992), "Inference from iterative simulation using multiple sequences," *Stat Sci*, 7, 457–511.

Hoeting, J., Madigan, D., Raftery, A., and Volinsky, C. (1999), "Bayesian model averaging: a tutorial," *Stat Sci*, 14, 382–417.

Konig, I., Schafer, H., Muller, H., and Ziegler, A. (2001), "Optimized group sequential study designs for tests of genetic linkage and association in complex diseases," *Am J Hum Genet*, 69, 590–600.

Konig, I., Schafer, H., Ziegler, A., and Muller, H. (2003), "Reducing sample sizes in genome scans: group sequential study designs with futility stops," *Genet Epidemiol*, 25, 339–349.

Meng, X. and Wong, W. (1996), "Simulating ratios of normalizing constants via a simple identity: A theoretical exploration," *Stat Sinica*, 6, 831–860.

Province, M. (2000), "A single, sequential, genome-wide test to identify simultaneously all promising areas in a linkage scan," *Genet Epidemiol*, 19, 301–322.

— (2001), "Sequential methods of analysis for genome scans," *Adv Genet*, 42, 499–514.

Raftery, A., Madigan, D., and Hoeting, J. (1997), "Bayesian model averaging for linear regression models," *J Am Stat Assoc*, 92, 179–191.

Schaid, D., Jenkins, G., Ingle, J., and Weinshilboum, R. (2013), "Two-phase designs to follow-up genome-wide association signals with DNA resequencing studies," *Genet Epidemiol*, published online, 24 January 2013.

Scherag, A., Hebebrand, J., Schafer, H., and Muller, H. (2009), "Flexible designs for genomewide association studies," *Biometrics*, 65, 815–821.

Scherag, A., Muller, H., Dempfle, A., Hebebrand, J., and Schafer, H. (2003), "Data adaptive interim modification of sample sizes for candidate-gene association studies," *Hum Hered*, 56, 56–62.

Seaman, S. and Richardson, S. (2004), "Equivalence of prospective and retrospective models in the Bayesian analysis of case-control studies," *Biometrika*, 91, 15–25.

Spencer, C., Hechter, E., Vukcevic, D., and Donnelly, P. (2011), "Quantifying the Underestimation of Relative Risks from Genome-Wide Association Studies," *PLoS Genet*, 7, e1001337.

Stephens, M. and Balding, D. (2009), "Bayesian statistical methods for genetic association studies," *Nat Rev Genet*, 10, 681–690.

Strauch, K., Fimmers, R., Baur, M., and Wienker, T. (2003), "How to model a complex trait, 1. General considerations and suggestions," *Hum Hered*, 55, 202–210.

van Dyk, D. and Meng, X. (2001), "The art of data augmentation," *J Comput Graph Stat*, 10, 1–50.

Vukcevic, D., Hechter, E., Spencer, C., and Donnelly, P. (2011), "Disease Model Distortion in Association Studies," *Genet Epidemiol*, 35, 278–290.

Wang, K. (2012), "Statistical tests of genetic association for case-control study designs," *Biostat*, 13, 724–733.

Yan, L., Zheng, G., and Li, Z. (2008), "Two-stage group sequential robust tests in family-based association studies: controlling type I error," *Ann Hum Genet*, 72, 557–565.