Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Lei Wang Date

Bayesian Functional Genome-wide Association Study using Standardized Individual-level and Summary-level GWAS Data

By

Lei Wang

Master of Science in Public Health

Biostatistics and Bioinformatics

Jingjing Yang, PhD

(Thesis Advisor)

Zhaohui "Steve" Qin, PhD

(Reader)

Bayesian Functional Genome-wide Association Study using Standardized Individual-level and Summary-level GWAS Data

By

Lei Wang

B.S.

Shanghai University

2020

Thesis Advisor: Jingjing Yang, PhD

Reader: Zhaohui "Steve" Qin, PhD

An abstract of

A thesis submitted to the Faculty of the

Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of

Master of Science in Public Health

in Biostatistics and Bioinformatics

2022

Abstract

Bayesian Functional Genome-wide Association Study using Standardized Individual-level and Summary-level GWAS Data

By Lei Wang

Background: Genome-wide association study associates specific genetic variations with the human complex traits and diseases. And Bayesian Functional Genome-wide Association Study (BFGWAS) method integrates functional annotation with GWAS data, based on a multivariate Bayesian regression model for variants in locus. The current method requires individual-level data, which limits the scope of application of the BFGWAS method to public available GWAS summary data. The main bottleneck is implementing MCMC algorithm using only GWAS summary data and reference linkage disequilibrium (LD) information. Thus, my thesis project is to adapt the BFGWAS method for standardized genotype and phenotype data so that it can be applied to summary data.

Methods and Materials: In this project, I derived the MCMC algorithm using standardized genotype and phenotype from either the individual-level or the summary-level data. A simulation study is conducted to test this novel method. I used the odds ratio from the real GWAS summary data of Age-related Macular Degeneration, and then simulated quantitative phenotype data for testing our tools with standardized individual-level and summary-level GWAS data.

Results: From the simulation results, the tools when using summary statistics can greatly improve the work efficiency comparing to the individual data. The total time cost for individual data is 2.1505 min. And this could be 0.0905 min when using summary statistics. Since the summary-level data was generated from the individual-level data, using summary-level data showed similar performance as using individual-level data. By taking variants with posterior causal probability larger than 0.1 as potential causal variants in our work, the detected potential SNPs from individual data and summary statistics are comparably consistent.

Conclusion: In this paper, I propose to extend the BFGWAS method for studying summary-level GWAS data through the MCMC algorithm based on standardized genotype and phenotype. The usefulness of summary statistics was demonstrated in a simulation. However, there are also some limitations here. The reference LD matrix may miss some values in real data, and this will cause computation error in MCMC algorithm and result in an unreliable conclusion. Thus, further real data will be considered and tested in the future work.

Bayesian Functional Genome-wide Association Study using Standardized Individual-level and Summary-level GWAS Data

By

Lei Wang

B.S.

Shanghai University

2020

Thesis Advisor: Jingjing Yang, PhD

Reader: Zhaohui "Steve" Qin, PhD

A thesis submitted to the Faculty of the

Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of

Master of Science in Public Health

in Biostatistics and Bioinformatics

2022

Acknowledgement

I really appreciate Emory University can give me an opportunity to engage in the advance research experience.

I am grateful to Dr. Yang for the guidance and encouragement she gives to me when I firstly join the lab. Her words give me a clear direction about what I want to do in the future. Even after graduation, I will still remember Dr. Yang's enthusiasm, rationality, and kindness. I will go to every challenge in the future with the same feeling. This relationship as a teacher and a friend, which has made me clearer about my roles in the public health.

To the faculty and staff of the Emory Rollins School of Public Health Department of Biostatistics and Bioinformatics, thank you for your time, lectures, assistance, and support. It is my proud to live with you all in this Emory community during this particularly difficult time of the pandemic.

Table of Contents

Introduction

Genome-wide association studies (GWAS) is very useful to identify genes that may related to human complex traits and diseases.[1] Currently, such studies, and the large number of valid results obtained, have greatly changed the way and efficiency of research on complex traits. Through the analysis of thousands of single-nucleotide polymorphisms (SNPs), underlying important variants of a disease can be detected. For example, the GWAS study about age-related macular degeneration (AMD) shows that Y402H variant in CFH, the rs10490924 single-nucleotide polymorphisms are the potential risk factors for AMD.[2] Hundreds of associated loci are in linkage disequilibrium (LD). And most of them have unknown functions or located outside the protein-coding regions. And a flexible Bayesian selection model can dramatically increase the computational efficiency and power for detecting potential variants according to the linkage disequilibrium (LD). [3]

However, the current Genome-wide association studies have a limitation of computational cost. When using Markov chain Monte Carlo (MCMC) algorithm to generate the result from complete genotype data, a long CPU hour is expected. [3] Typically, using summary statistics will be a good option to reduce the working time. Using the estimated linkage disequilibrium (LD) from a reference panel with individual-level genotype data, a meta-analysis of genome-wide association studies (GWAS) can compute fast. And another benefit is that the LD between the unknown causal variants at the locus can explain the total variation more comprehensively. [4] The reason is that a single genotyped SNP may cannot account for the variation. Thus, applying summary

statistics to current Bayesian functional Genome-wide Association Studies will increase both the working efficiency and the power to detect unknown casual variants.

Besides, the importance of using public GWAS data set with large sample size is become an advanced topic nowadays. [5] However, when only handling individual data with original approach, the computational time is usually very time consuming. This is a challenge for us to apply public GWAS data with summary level data in an easier way.

Intuitively, if a method only uses summary statistics of Genome-wide Association Studies (GWAS) and an external LD reference panel, this method will give a more efficient computation. According to the recent polygenic prediction method, PRS-CS indicates that standardized genotypes, which have each column centered and have unit variance), will give us an option to apply the summary statistics by using LD information from an external reference panel without individual-level data. [6] This method is very revealing. Because on the conceptual framework, PRS-CS and Bayesian Functional Genome-wide Association studies (GWAS) have very similar structures. If standardized genotypes are used, the external LD reference panel and summary statistics can be directly used. This will greatly improve the current operating efficiency, reduce the variance between different distribution and results less errors, and even improve the detection ability of unknown potential causal SNPs. Thus, I plan to extend current BFGWAS tool for public data with the help of summary statistics and standardization.

Thus, in this project, I combine genome-wide association studies (GWAS) with Bayesian functional methods and standardization together, using only summary statistics to detect the

potential variants through linkage disequilibrium (LD) with the simulation data. As expected, this method will increase the working efficiency and the power to detect unknown variants.

Material and Method

Bayesian Variable Selection Model:

In the genome-wide association studies, millions of genetic variants are measured. The aim of identifying these SNPs in BFGWAS can be treated as a variable selection problem. [7] The standard Bayesian variable selection regression (BVSR) model is used as the basic framework construction for our process. With the help of this basic model, I can build a simple way to apply the summary statistics.

$$Y_{n\times 1} = Z_{n\times p}\beta_{p\times 1} + \epsilon_{n\times 1}, \epsilon_{n\times 1} \sim MVN(0, I_{n\times n})$$

In this linear regression model, $Y_{n\times 1}$ is the standardized phenotype vector with n individuals, $Z_{n\times p}$ is the standardized genotype matrix with p genetic variants. Different from the BVSR model with centered phenotype vector and centered genotype matrix, the standardized version has $\epsilon_{n\times 1}$ follows a multivariable normal distribution with 0 mean and unit variance. Thus, I can give a simpler posterior for summary statistics. The $\beta_{i,q}$ is the vector of the effect sizes of each specific annotation category q, and the $\beta_{i,q}$ follows a "spike-and-slab" [8] prior:

$$\beta_{i,q} \sim \pi_q N\left(0, \frac{1}{n}\sigma_q^2\right) + (1-\pi_q)\delta_0(\beta_{i,q}), q \in \{1,2,\dots,Q\}$$

The individual data was separated into Q non-overlap segments. The β_i in each annotation category q, it has π_q probability to be a normal distribution of $N\left(0, \frac{1}{n}\sigma_q^2\right)$, and $(1-\pi_q)$ probability to choose a point mass function $\delta_0(\beta_{i,q})$ at 0. Thus, π_q is the unknown probability

for variants in the q^{th} category and σ_q^2 is the corresponding effect-size variance. Assuming that this (π_q, σ_q^2) summary statistics for category q:

$$\sigma_q^2 \sim Inverse\ Gamma(K_1, K_2), \pi_q \sim Beta(a, b)$$

The priors [9] are used for (π_q, σ_q^2). For example, I assumed $\sigma_q^2 \sim Inverse\ Gamma(K_1, K_2)$ with $K_1 = K_2 = 0$, and $\pi_q \sim Beta(a, b)$ with a mean 10^{-6}. The prior distribution will let our Bayesian model to estimate mainly depend on the association from real data in each category. I can give an inference for categorical specific parameters (π_q, σ_q^2), which can represent the most importance information in each category. Our goal is to estimate this pair parameters more easily, I introduce an indicator latent variable [10] instead:

$$\gamma_{i,q} \sim Bernulli(\pi_q)$$

Assuming a new indicator latent variable γ_i for each category that follows a Bernoulli distribution with a probability π_q. And I already have $Y_{n\times 1}$ and $Z_{n\times p}$ standardized. Thus, I have $Y^T Y = n$, and $Z^T Z = nD$, and n is sample size. Matrix D can directly derive from the external reference LD matrix. According to these, I can derive the posterior distribution for each estimator according to the simplified transformation.

Posterior Distribution of Estimators:

Through the Bayesian method, the joint posterior probability for each category is given:

$$P(\sigma^2, \pi, \beta, \gamma | Y, Z) \sim P(Y|Z, \beta, \gamma) P(\beta|\gamma, \sigma^2) P(\gamma|\pi) P(\pi) P(\sigma^2)$$

The joint posterior distribution is the product of each conditional posterior of estimators and their own prior distributions. The Bayesian method gives us a way to provide estimators from the individual data and apply them into the Markov Chain Monte Carlo (MCMC) algorithm.

The Conditional Posterior Distribution:

According to the joint posterior distribution, I can derive conditional posterior distribution for each estimator. For category q, I should have conditional posterior for β:

$$P(\beta_{|\gamma|}|Y, Z, \gamma, \sigma^2, \pi) \propto P(Y|Z, \beta_{|\gamma|}, \gamma) P(\beta|\gamma, \sigma^2)$$

$$P(\beta_{|\gamma|}|Y, Z, \gamma, \sigma^2, \pi) \propto \exp\left\{-\frac{1}{2}\left(Y - Z\beta_{|\gamma|}\right)^T \left(Y - Z\beta_{|\gamma|}\right)\right\} \cdot \exp\left\{-\frac{1}{2}\beta_{|\gamma|}^T \left(n \cdot V_{|\gamma|}^{-1}\right) \cdot \beta_{|\gamma|}\right\}$$

$$P(\beta_{|\gamma|}|Y, Z, \gamma, \sigma^2, \pi) \propto \exp\left\{-\frac{1}{2}\left[\beta_{|\gamma|}^T \left(Z^T Z + nV_{|\gamma|}^{-1}\right)\beta_{|\gamma|} - 2\beta_{|\gamma|}^T Z^T Y\right]\right\}$$

The conditional posterior distribution of $\beta_{|\gamma|}$ is as proportion to the product of conditional distribution of Y and the distribution of β given the pair estimated parameter γ and σ^2. After a transformation, the conditional posterior distribution of $\beta_{|\gamma|}$ actually follows a multivariate normal distribution:

$$P\left(\beta_{|\gamma|}\middle|Y,Z,\gamma,\sigma^2,\pi\right)\sim MVN\left(\mu_{\beta_{|\gamma|}},\Sigma_{\beta_{|\gamma|}}\right)$$
$$\Sigma_{\beta_{|\gamma|}}=\left(Z^TZ+nV_{|\gamma|}^{-1}\right)^{-1}=\frac{1}{n}(D+V_{|\gamma|}^{-1})$$
$$\mu_{\beta_{|\gamma|}}=\Sigma_{\beta_{|\gamma|}}\cdot Z^TY=n\cdot\Sigma_{\beta_{|\gamma|}}\cdot\hat{\beta}$$

The reason that using summary statistics to generate the posterior distribution is there is a connection between our estimators and the summary statistics after standardization. GWAS summary statistics are given:

$$\hat{\beta}=\frac{Z^TY}{n};\ Z^TY=n\cdot\hat{\beta}$$
$$D=\frac{Z^TZ}{n};\ Z^TZ=n\cdot D$$
$$Y^TY=n$$

These summary statistics can give a simple way to generate estimated $\hat{\beta}$. And D can be approximated by a reference panel so that I can do the estimation easier and more efficient [11]. Also, with the help of these summary statistics generated from the individual data, I can only use summary statistics to do the association analysis. Since the conditional posterior distribution consider of β has given, we can also have the latent indicator variable γ in each category from the integration [12] of β :

$$P(\gamma|Y,Z,\sigma^2,\pi) \propto \int_{\beta} P(Y|Z,\beta,\gamma)P(\beta|\gamma,\sigma^2)P(\gamma|\pi)\,d\beta$$

$$\propto \frac{1}{\sqrt{(2\pi)^n}\cdot\sqrt{(2\pi)^m}\cdot\sqrt{\left|\frac{1}{n}V_{|\gamma|}\right|}}$$

$$\cdot\int_{\beta} \exp\left(-\frac{1}{2}[\beta_{|\gamma|}^T\left(Z^TZ+nV_{|\gamma|}^{-1}\right)\beta_{|\gamma|}-2\beta_{|\gamma|}^T(Z^TY)+Y^TY]\right)\cdot P(\gamma|\pi)d\beta$$

$$\propto P(\gamma|\pi)\cdot\frac{\sqrt{|\Sigma_\beta|}}{\sqrt{\left|\frac{1}{n}V_{|\gamma|}\right|}}\cdot\exp\left\{-\frac{n}{2}\left(1-n\cdot\hat{\beta}^T\Sigma_\beta\hat{\beta}\right)\right\}$$

After integration, the conditional posterior distribution can be estimated. And here, $|\gamma|$ is the number of SNPs with $\gamma_{iq}=1$. Also, there are some key computations. These computations will give us an overview about how standardization can simplify the calculation:

$$\Sigma_{\beta_{|\gamma|}}=\frac{1}{n}(D+V_{|\gamma|}^{-1})^{-1}$$

With the help of this key computational equation, summary statistics are applied to estimate both β and γ by using Markov chain Monte Carlo (MCMC) algorithm.

Markov Chain Monte Carlo Algorithm:

The challenges of standard MCMC algorithm are the memory usage and convergence rate. In this process, there are K non-overlap genome blocks to improve the efficiency. Firstly, the initial category specific parameters are set up [13]. An category specific parameters pair (π_q, σ_q^2) was assumed. The $\sigma_q^2 \sim Inverse\ Gamma(K_1, K_2)$ with $K_1 = K_2 = 0$, and $\pi_q \sim Beta(a, b)$ with a mean 10^{-6}. The summary statistics $Z^T Y$ and $Z^T Z$ can improve the working efficiency here since the correlation from the external reference panel can be provided.

EM Update:

In each genome block, an estimation step and a maximum step are applied [14]. The E-step estimates the variant specific parameters according to the most recent (π_q, σ_q^2). And the M-step maximized the expected log-posterior-likelihood functions based on the variant specific parameters in the E-step.

Simulation Study Results

Test Data:

The 1KG reference data is used as the individual genotype data. I used the genotype data of 1KG samples of the AMD associated CFH locus from one genome block. This genome block was from base position 196179832 to 197268053 in Chromosome 1 with total 31330 SNPs and 3689 SNPs are analyzed. The true casual SNPs contains 15 SNPs for 2504 observations. Also, the phenotypes are simulated based on a logistic regression model [15]. Thus, we can generate the cases and controls by using the odd ratio directly. Phenotype is taken as 1's for cases and 0's for controls in our Bayesian regression model.

Individual Data with Standardization:

After simulating the test data, the summary statistics is firstly generated from the individual data. The summary statistics contains two parts, one is the z-score and another one is the LD inference. The setting for the generating process includes that 0.005 is used as the minor allele frequency threshold [16]. After the MCMC algorithm, 9 potential SNPs are introduced with causal probability higher than 0.1 which means it has a relatively high probability to be a potential true causal SNP. And there are 3689 total SNPs was analyzed. There are 9 potential causal SNPs were selected from the result of the individual data. The table 1 show the ID, reference allele and minor allele and other useful information.

In the estimator part, we can see rs78217329 has a relatively large p-value. And the rank is 727 which is the lowest rank in these 9 SNPs. Typically, in the analysis part, this kind of SNP will be removed since the result is not such significant and the rank is not such reasonable.

Use Summary Statistics with Standardization:

With the MAF 0.005 threshold and the casual probability with 0.1, the basic information when using summary statistics with standardization shows a similar ability to detect the potential SNPs with the individual one. And the casual probability is relatively high and reasonable. However, the rs61818915 is not in the individual result and rs3766405 is not in the potential SNPs when using summary statistics with standardization.

Efficiency Comparison:

Table 1 Basic Information for Individual Data with Standardization

ID	Reference	Alternative	MAF	Probability	BETA	P-value	Rank
rs10801558	T	G	0.5076	0.1880	-0.3755	5.97E-32	1
rs10922108	A	T	0.5080	0.3180	-0.3761	6.00E-32	3
rs10922109	C	A	0.5076	0.2050	-0.3749	5.97E-32	2
rs1410996	G	A	0.5050	0.1010	-0.3729	6.09E-32	5
rs148072867	A	C	0.2835	1.0000	-0.2607	1.02E-11	32
rs3766405	C	T	0.5050	0.1880	-0.3732	6.09E-32	4
rs68045083	CAT	C	0.3712	1.0000	0.2426	3.50E-12	29
rs74861068	G	A	0.0795	0.8500	0.1714	1.15E-10	39
rs78217329	G	GT	0.4904	1.0000	0.3820	0.01064	727

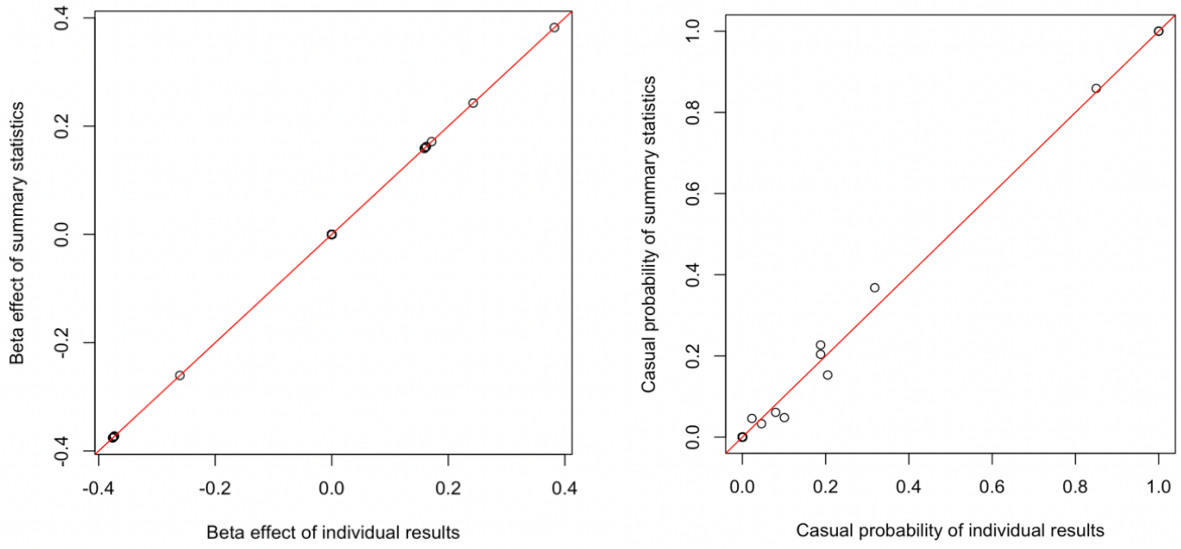
Table 2 Basic Information of SNPs when using Summary Statistics with Standardization

ID	Reference	Alternative	MAF	Probability	BETA	P-value	Rank
rs10801558	T	G	0.5076	0.2686	-0.3720	5.97E-32	1
rs10922108	A	T	0.5080	0.3100	-0.3728	6.00E-32	3
rs10922109	C	A	0.5076	0.2076	-0.3713	5.97E-32	2
rs1410996	G	A	0.5050	0.1140	-0.3696	6.09E-32	5
rs148072867	A	C	0.2835	0.4468	-0.2609	1.02E-11	32
rs61818915	C	A	0.2542	0.5532	-0.2414	4.48E-12	30
rs68045083	CAT	C	0.3712	0.9779	0.2492	3.50E-12	29
rs74861068	G	A	0.0795	0.9204	0.1710	1.15E-10	39
rs78217329	G	GT	0.4904	1.0000	0.3731	0.01064	727

Besides, considering the efficiency of results between individual data and summary statistics, we can see the casual probability are all relatively reasonable in table 2. Although some of the potential casual SNPs in the table are not as same as the true causal SNPs, they have a reasonable high correlation between each other.

Also, the effect of potential casual SNPs between the individual and summary statistics results has a high consistency. According to the figure 1, we can also see that the casual probability of between two models have a similar result. This indicates a reasonable power and accuracy of using summary statistics in detecting the potential casual SNPs. And at the same time, the sum of casual probability of summary statistics' result is 5, which is same as the result with the individual data. Thus, a consistency result can be provided.

Figure 1 The Consistency of the Results between the Summary Statistics and Individual Data



However, there is a potential problem. Although the result of summary statistics and individual data have a consistent similarity, these detected potential causal potential SNPs may not as same as the true SNPs. One of the reasons is that we use the same data for generating the summary statistics and then apply these summary statistics into the original data. This could lead to an unreliable result. And most points are located near 0 since a large number of SNPs in 3689 total analyzed SNPs are not possible to be a true casual SNP.

Another objective is that hoping standardization and the using of summary statistics will improve the work efficiency. In other words, our computational time costs will have a great decrease. When conduct the simulation study with the same setting for MCMC and LD windows. The results show a quite clear improvement on using the summary statistics.

Table 3 The Work Efficiency for Different Method

	Individual data with standardization	Summary statistics with standardization
Time on MCMC	0.0062 min	0.0040 min
Time on Proposal	0.0043 min	0.0022 min
Time on Posterior	4294.3300 min	176.03700 min
Total computation time	2.1505 min	0.0905 min

Discussion

This project has a main aim that to make BFGWAS applicable for public summary-level GWAS data. In the simulation study, the computational time can be reduced greatly by using the summary statistics. And the computational errors can be reduced as well. However, when only using the simulation data, the result of potential SNPs is not such reliable. To some extent, the non-differential measurement error in phenotypes can lead to unreliable causal inference. [17]

Typically, standardization will reduce some potential error in MCMC algorithm. However, the improvement of the standardization may not be such reliable in the real data. One of the reasons is that we the LD matrix in posterior computation is not such ideal. Sometimes, standardization will cause the LD matrix to become hard to interpretate when having missing values in LD matrix. To be specific, the covariance will be scaled and lead to some very similar correlations between different SNPs pairs. [18] But this problem is not such matter since we can consider the correlation between these SNPs with the true SNPs to give a relative convincing result when we apply this tool in the real data. Difference between the cohort of real data and LD reference will also cause some potential errors.

We simulated the test data from the odds ratio of the true causal SNPs for a 1KG reference data. And then we use this simulated data to generate the summary statistics. This can avoid many potential out range problems (like the likelihood ratio). But this leads to an unreliable result for detecting the potential SNPs. This can be exacerbated by the extensive correlation between genetic variants caused by LD. [19] In brief, as a simulation for tool testing, it is enough. But for detecting potential SNPs, more data are needed to get a reasonable result. In the real data, the LD

matrix can have many unexpected problems like missing values. This will greatly affect the standardization process and results some unreliable conclusion. We need more tests in the real data in the future so that we can fix this tool more.

And for the work efficiency, the summary statistics really can improve the work efficiency. However, considering standardization part, there could be a potential problem about the LD matrix when applying summary statistics in the real data. Since this is a simulation study, which indicates that the result could be similar since we use the summary statistics and simulation dataset from the same true causal SNPs data. Thus, more results from the real data to detect the potential SNPs are needed.

In the future, I plan to apply this tool with summary statistics and standardization to the real data of Alzheimer's disease, which will be a public GWAS data. Since the limitation of LD matrix in the real data may cause computation error and lead to MCMC fail, the real data test is necessary in the future. At the same time, the reference LD matrix may have some missing values and lead to an inaccuracy result. Thus, some improvements are needed as well. These improvement and real data analysis will be conducted in my future work. I need to address these potential problems when I apply this method to the real data, which is Alzheimer's disease. And then a more comprehensive tool will be provided.

References:

[1] Manolio, T. A. (2010). Genomewide association studies and assessment of the risk of disease. *New England journal of medicine*, *363*(2), 166-176.

[2] Yang, J., Li, Y., Chan, L., Tsai, Y. T., Wu, W. H., Nguyen, H. V., ... & Tsang, S. H. (2014). Validation of genome-wide association study (GWAS)-identified disease risk alleles with patient-specific stem cell lines. *Human molecular genetics*, *23*(13), 3445-3455.

[3] Yang, J., Fritsche, L. G., Zhou, X., Abecasis, G., & International Age-Related Macular Degeneration Genomics Consortium. (2017). A scalable Bayesian method for integrating functional information in genome-wide association studies. *The American Journal of Human Genetics*, *101*(3), 404-416.

[4] Yang, J., Ferreira, T., Morris, A. P., Medland, S. E., Madden, P. A., Heath, A. C., ... & Visscher, P. M. (2012). Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nature genetics*, *44*(4), 369-375.

[5] Zheng, J., Erzurumluoglu, A. M., Elsworth, B. L., Kemp, J. P., Howe, L., Haycock, P. C., ... & Neale, B. M. (2017). LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. *Bioinformatics*, *33*(2), 272-279.

[6] Ge, T., Chen, C. Y., Ni, Y., Feng, Y. C. A., & Smoller, J. W. (2019). Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nature communications*, *10*(1), 1-10.

[7] Guan, Y., & Stephens, M. (2011). Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *The annals of applied statistics*, *5*(3), 1780-1815.

[8] Ishwaran, H., & Rao, J. S. (2005). Spike and slab variable selection: frequentist and Bayesian strategies. *The Annals of Statistics*, *33*(2), 730-773.

[9] Ramos, P. L., Mota, A. L., Ferreira, P. H., Ramos, E., Tomazella, V. L., & Louzada, F. (2021). Bayesian analysis of the inverse generalized gamma distribution using objective priors. *Journal of Statistical Computation and Simulation*, *91*(4), 786-816.

[10] Titsias, M., & Lawrence, N. D. (2010, March). Bayesian Gaussian process latent variable model. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics* (pp. 844-851). JMLR Workshop and Conference Proceedings.

[11] Zhu, X., & Stephens, M. (2017). Bayesian large-scale multiple regression with summary statistics from genome-wide association studies. *The annals of applied statistics*, *11*(3), 1561.

[12] Müller, P. (1991). *A generic approach to posterior integration and Gibbs sampling*. Purdue University, Department of Statistics.

[13] Brooks, S. (1998). Markov chain Monte Carlo method and its application. *Journal of the royal statistical society: series D (the Statistician)*, *47*(1), 69-100.

[14] Li, Z., Zhang, Z., He, Z., Tang, W., Li, T., Zeng, Z., ... & Shi, Y. (2009). A partition-ligation-combination-subdivision EM algorithm for haplotype inference with multiallelic markers: update of the SHEsis (http://analysis. bio-x. cn). *Cell research*, *19*(4), 519-523.

[15] Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., & Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics*, *89*(1), 82-93.

[16] Tabangin, M. E., Woo, J. G., & Martin, L. J. (2009, December). The effect of minor allele frequency on the likelihood of obtaining false positives. In *BMC proceedings* (Vol. 3, No. 7, pp. 1-4). BioMed Central.

[17] Hu, Y., Lu, Q., Powles, R., Yao, X., Yang, C., Fang, F., ... & Zhao, H. (2017). Leveraging functional annotations in genetic risk prediction for human complex diseases. *PLoS computational biology*, *13*(6), e1005589.

[18] Hemani, G., Tilling, K., & Davey Smith, G. (2017). Orienting the causal relationship between imprecisely measured traits using GWAS summary data. *PLoS genetics*, *13*(11), e1007081.

[19] Asimit, J. L., Rainbow, D. B., Fortune, M. D., Grinberg, N. F., Wicker, L. S., & Wallace, C. (2019). Stochastic search and joint fine-mapping increases accuracy and identifies previously unreported associations in immune-mediated diseases. *Nature communications*, *10*(1), 1-15.