Master’s Thesis of Data Science

Meta-Analysis of the Association Between Rare Genetic Variants and Imbalanced Binary Phenotypes

불균형한 이진 표현형과 희귀 유전자 연관성의 메타분석

August 2023

Graduate School of Data Science
Seoul National University
Data Science Major

Park Eun Jae

Meta-Analysis of the Association Between Rare Genetic Variants and Imbalanced Binary Phenotypes

Adviser Lee Seunggeun

Submitting a master's thesis of
Data Science

July 2023

Graduate School of Data Science
Seoul National University
Data Science Major

Park Eun Jae

Confirming the master's thesis written by
Park Eun Jae
July 2023

Chair ________ Hyung-Sin Kim ________ (Seal)

Vice Chair ________ Seunggeun Lee ________ (Seal)

Examiner ________ Jay-Yoon Lee ________ (Seal)

Abstract

While rare genetic variants significantly contribute to diverse phenotypes, their low frequency poses challenges for association tests. Meta-analysis is a practical approach for identifying such variants by combining summary statistics from multiple studies. However, current methods for rare variant meta-analysis exhibit limitations, particularly when analyzing low-prevalence dichotomous traits.

In this paper, we introduce Meta-SAIGE, a novel approach for rare variant meta-analysis. Meta-SAIGE is designed to reduce type 1 error inflation through precise estimation of the distribution of test statistics. By allowing to reuse the LD-matrix for different phenotypes, Meta-SAIGE enhance the computational efficiency and enables phenom-wide analysis.

We evaluated the performance of Meta-SAIGE using Whole Exome Sequencing data from the UKBiobank. Simulated null phenotypes were used to assess the type 1 error rate, and real UK-Biobank case-control phenotypes showed the consistency of the meta-analysis results with SAIGE-GENE+, a joint analysis of individual level data.

Keyword: Meta-Analysis, Rare-variant, Gene-based association test

Student Number: 2021-27059

Table of Contents

Abstract i

Table of Contents ii

1 INTRODUCTION 1

1.1 GWAS 1

1.2 Importance of Rare Variants 2

1.3 Gene-based Association Test 2

1.4 Meta-analysis of Rare Variants 3

2 Methods 5

2.1 Workflow of Meta-SAIGE 5

2.1.1 Generating Summary Statistics From Each Study Cohorts 5

2.1.2 Combining Summary Statistics Into a Single Super-Set 7

2.1.3 Running Gene-Based Test With the Super-Set 8

2.2 Further Adjustment with GC-based Meta-Analysis Approach 9

2.3 Cohort-Specific Collapsing 10

2.4 Type 1 Error Evaluation 10

2.5 Real Data Evaluation 11

3 Results 13

3.1 Type 1 Error Evaluation 13

3.2 Real Data Evaluation 13

4 Discussions 18

Bibliography 20

Appendix 22

Abstract in Korean 24

Chapter 1

INTRODUCTION

1.1 GWAS

The Genome-Wide Association Studies (GWAS) are one of the important methods that changed the way to explore the genetic basis of complex phenotypes. Before GWAS, genetic association tests were performed on a small number of candidate genes that were assumed to be related to the traits of interest. The genetic association tests were often carried out using linkage analysis(Ott et al. [2015]), which was very limited in the sample sizes and statistical power. However, with the completion of the Human Genome (Consortium et al. [2001]) and advancements of high-throughput sequencing technologies(Reuter et al. [2015]), researchers can genotype millions of single nucleotide polymorphisms (SNPs), which enabled large-scale association studies.

The first GWAS was published in 2005, which identified the single-marker association to the age-related macular degeneration(Klein et al. [2005]). This study demonstrated that GWAS could be a useful tool to detect novel genetic associations for complex diseases. Since then, the number of GWAS has grown exponentially. Numerous studies are published to date, uncovering associations of not only continuous phenotypes, but binary phenotypes such as diagnosis codes for diseases (Visscher et al. [2017]).

The establishment of large biobanks and consortia have further facilitated the growth of GWAS by providing researchers with access to extensive datasets and resources. These collaborative efforts have enabled the discovery of numerous genetic associations, providing valuable insights into the genetic architecture of complex traits and diseases.

1.2 Importance of Rare Variants

Although GWAS has proven to be a valuable tool in identifying novel genetic associations, they have a significant limitation. Specifically, the heritability from GWAS variants are considerably lower than those obtained through sibling recurrence risk or residual phenotypic variance measurements (Manolio et al. [2009]). This phenomenon, known as ”missing heritability,” has been observed in a wide range of diseases, including age-related macular degeneration, Crohn’s disease, and type 2 diabetes, among others. Various hypotheses have been proposed to explain this discrepancy. One of the possible reason is that GWAS primarily captures the effects of common genetic variants, while failing to account for the contributions of rare variants.

1.3 Gene-based Association Test

Due to low minor allele frequencies, single variant tests in GWAS have limited efficacy for identifying phenotype associated rare variants. As an alternative, gene-based association tests have been developed, capable of detecting the collective impact of multiple rare variants. Methods such as the Burden test and SKAT have been proposed to the combine the effects of variants in each gene. SKAT-O method, which integrated both the Burden test and SKAT, has gained widespread adoption for gene-based association tests (Lee et al. [2012]).

While this effect combining method is useful, statistical power can be further improved by collapsing ultra-rare variants (Zhou et al. [2022]). The ultra-rare variant collapsing also has been shown to effectively control both type 1 error rates. The new SAIGE-GENE+ method combines all the listed methods and can also incorporate functional annotations and different minor allele frequency (MAF) cutoffs. It performs multiple gene-based test with different functional annotations and MAF cutoffs, and combines significance using the Cauchy combination(Liu and Xie [2020]).

1.4 Meta-analysis of Rare Variants

Meta-analysis is a widely used statistical method that enables the identification of genetic associations that may not be detectable in individual studies, but become significant when data from multiple studies are combined. This is particularly relevant in the context of rare variant association tests, where the low frequencies of the variants can limit their detection in individual studies. As biobanks continue to grow in size and scope, researchers are increasingly motivated to collaborate in large-scale meta-analyses of rare variants. To facilitate such efforts, international consortia such as the Biobank Rare Variant Analysis (BRaVa) has been established. These consortia aim to harmonize data collection, annotation, and analysis across multiple studies, thereby increasing statistical power and enabling the identification of rare variants that may be associated with complex diseases. By leveraging the strengths of multiple studies, meta-analysis of rare variants has the potential to provide a more comprehensive understanding of the genetic basis of disease and to inform the development of personalized diagnostic and therapeutic strategies.

Several meta-analysis methods have been developed to address the challenges of detecting rare genetic variants, including RareMetal (Feng et al. [2014]), MetaSKAT (Lee et al. [2014]), and MetaSTAAR (Li et al. [2023]). While these methods have made significant contributions to the field, they have limitations in terms of type I error control, scalability and statistical power. For example, RareMetal and MetaSKAT have been shown to have limited scalability for large whole-genome sequencing studies.

MetaSTAAR is a recently developed method targeted for large whole-genome sequencing studies. It incorporates multiple variant functional annotations into the analysis process and accounts for sample relatedness for both continuous and dichotomous traits. However, for low-prevalence dichotomous traits, MetaSTAAR can produce false positives. It is a critical limitation as many diseases have low prevalence. Additionally, MetaSTAAR requires generating covariance matrices of individual variant score statistics for each phenotype separately, which hampers the analysis of many phenotypes.

In this study, I propose a new rare variant meta-analysis method called Meta-SAIGE. Meta-SAIGE aims to 1) reduce the type 1 error inflation by accurately estimating the variance of score statistics through GC-based saddlepoint approximation method(Dey et al. [2019]), 2) be computationally more efficient by sharing one sparse LD matrix only for the covariance matrix of score statistics for all phenotypes, 3) propose a novel method that complies to biobanks' privacy policies regarding ultra-rare variant handling.

Chapter 2

Methods

2.1 Workflow of Meta-SAIGE

Meta-SAIGE is a three-step process that involves: (1) preparing single variant level association summaries for each cohort, (2) combining summary statistics from all the studies into a single super-set, and (3) running gene-based tests, as illustrated in Figure 2.1.

2.1.1 Generating Summary Statistics From Each Study Cohorts

Initially, the first step can be performed using SAIGE, which generates score statistics for each variant (S), its variance (V), and sparse LD-matrix (Ω). For binary phenotypes, the following logistic regression model is employed for association tests (Dey et al. [2017]).

$$logit[Pr(Y_i = 1|X_i, G_i)] = X^T\beta + G_i\gamma$$

where for the i^{th} sample, Y_i denotes the phenotype, X_i denotes the non-genetic covariates, and G_i denote the genotype.

Assuming there are K studies, the score statistic for each variant j in the k^{th} study is computed as follows, where n_k is the sample size, $\tilde{G}_{i,j}$ is the covariate adjusted

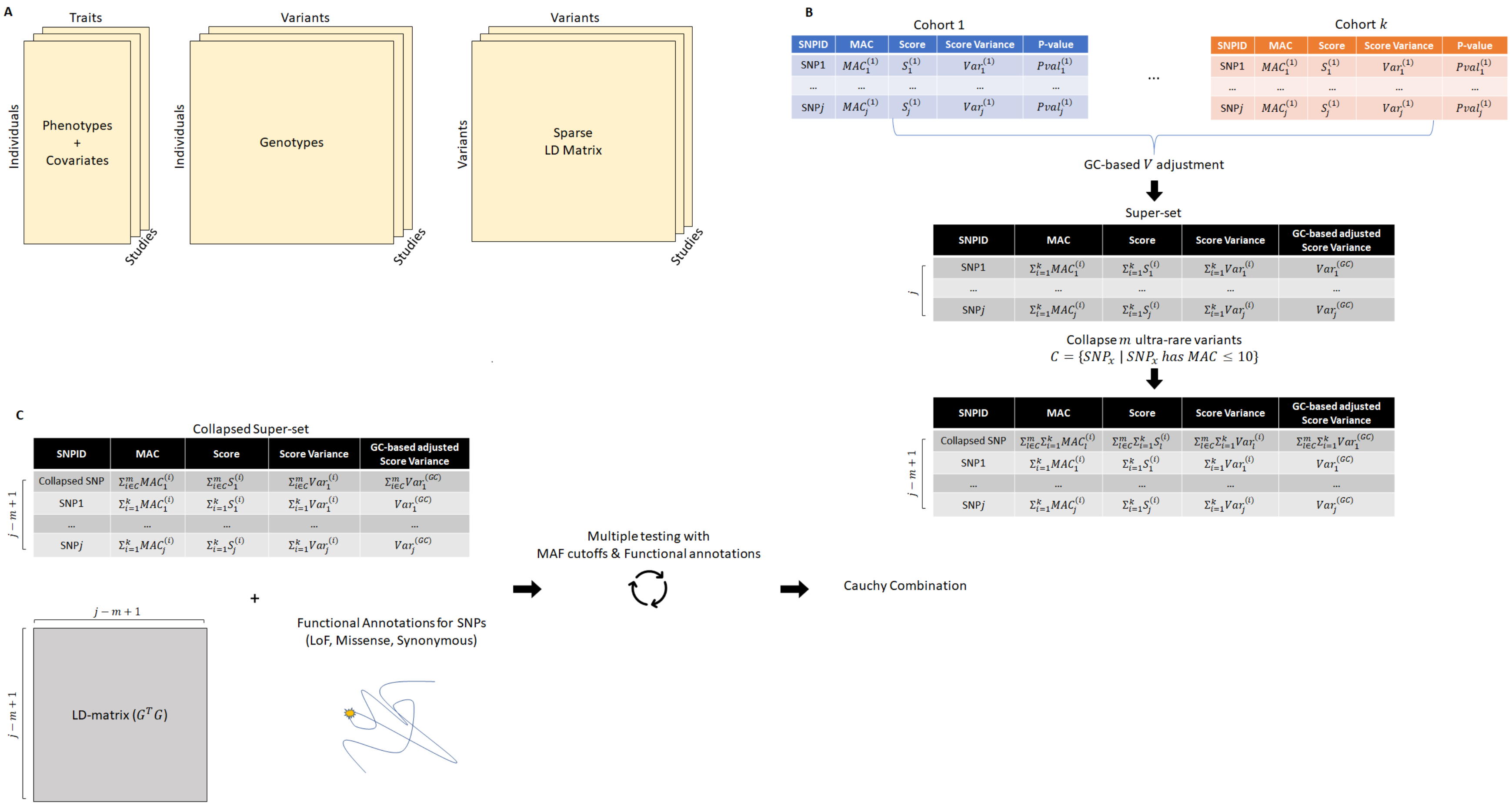

A
Traits
Individuals
Phenotypes + Covariates
Studies
Variants
Individuals
Genotypes
Studies
Variants
Variants
Sparse LD Matrix
Studies
B
Cohort 1
SNPID | MAC | Score | Score Variance | P-value
SNP1 | $MAC_1^{(1)}$ | $S_1^{(1)}$ | $Var_1^{(1)}$ | $Pval_1^{(1)}$
SNPj | $MAC_j^{(1)}$ | $S_j^{(1)}$ | $Var_j^{(1)}$ | $Pval_j^{(1)}$
Cohort k
SNPID | MAC | Score | Score Variance | P-value
SNP1 | $MAC_1^{(1)}$ | $S_1^{(1)}$ | $Var_1^{(1)}$ | $Pval_1^{(1)}$
SNPj | $MAC_j^{(1)}$ | $S_j^{(1)}$ | $Var_j^{(1)}$ | $Pval_j^{(1)}$
GC-based V adjustment
Super-set
SNPID | MAC | Score | Score Variance | GC-based adjusted Score Variance
j
SNP1 | $\Sigma_{i=1}^{k} MAC_1^{(i)}$ | $\Sigma_{i=1}^{k} S_1^{(i)}$ | $\Sigma_{i=1}^{k} Var_1^{(i)}$ | $Var_1^{(GC)}$
SNPj | $\Sigma_{i=1}^{k} MAC_j^{(i)}$ | $\Sigma_{i=1}^{k} S_j^{(i)}$ | $\Sigma_{i=1}^{k} Var_j^{(i)}$ | $Var_j^{(GC)}$
Collapse m ultra-rare variants
$C = \{SNP_x \mid SNP_x \; has \; MAC \leq 10\}$
$j - m + 1$
Collapsed SNP | $\Sigma_{i \in C}^{m} \Sigma_{i=1}^{k} MAC_i^{(i)}$ | $\Sigma_{i \in C}^{m} \Sigma_{i=1}^{k} S_i^{(i)}$ | $\Sigma_{i \in C}^{m} \Sigma_{i=1}^{k} Var_i^{(i)}$ | $\Sigma_{i \in C}^{m} \Sigma_{i=1}^{k} Var_1^{(GC)}$
C
Collapsed Super-set
Collapsed SNP | $\Sigma_{i \in C}^{m} MAC_1^{(i)}$ | $\Sigma_{i \in C}^{m} S_1^{(i)}$ | $\Sigma_{i \in C}^{m} Var_1^{(i)}$ | $\Sigma_{i \in C}^{m} Var_1^{(GC)}$
+
$j - m + 1$
LD-matrix ($G^T G$)
Functional Annotations for SNPs (LoF, Missense, Synonymous)
Multiple testing with MAF cutoffs & Functional annotations
Cauchy Combination

Figure 2.1: **Key steps of Meta-SAIGE** - (A) Generation of single variant level association summary using SAIGE. (B) Share of single variant summary statistics (score statistics S and its variance V) and a sparse LD matrix. The variance V is adjusted by SPA. Multiple summaries are merged into a super-set and the variance is additionally adjusted by the GC-based method. (C) Gene-based test using the generated super-set is performed multiple times with different functional annotations and MAF cut-offs, and p-values are combined with Cauchy-combination. As like SAIGE-GENE+, Ultra-rare collapsing is also performed for $MAC \leq 10$.

genotype, and $\hat{\mu}_i$ is the estimate of $Pr(Y_i = 1|X_i)$ under H_0.

$$S_{j,k} = \sum_{i=1}^{n_k} \tilde{G}_{i,j,k}(Y_{i,k} - \hat{\mu}_{i,k})$$

$V_{j,k}$, the variance of $S_{j,k}$, can be calculated as follows, where $\tilde{G}_{j,k}$ is a covariate adjusted genotype vector, and W_k is a diagonal matrix with $(1 - \hat{\mu}_{i,k})\hat{\mu}_{i,k}$ as the i^{th} diagonal element.

$$V_{j,k} = \tilde{G}_{j,k}^T W_k \tilde{G}_{j,k}$$

However, utilizing this $V_{j,k}$ in the score test can lead to an inflation of type 1 error. This issue arises because the score distribution may not follow a Gaussian distribution, but instead exhibit skewness, which is commonly observed for phenotypes with unbalanced case-control ratios. To address this problem, SAIGE incorporates the saddlepoint approximation (SPA) method, as introduced by Daniels [1954], which utilizes the cumulant generating function(CGF). SPA is a moment-based method that provides a more accurate approximation for skewed distributions, thus reducing the occurrence of type 1 errors. The CGF, $K(t)$, can be defined as below, and its second derivative can be used to approximate adjusted $V_{j,k}$.

$$K_k(t) = \sum_{i=1}^{n_k} log(1 - \hat{\mu}_{i,k} + \hat{\mu}_{i,k} e^{\tilde{G}_{i,k}t}) - \sum_{i=1}^{n_k} \tilde{G}_{i,k}\hat{\mu}_{i,k}$$

The sparse LD matrix (Ω) can also be generated by SAIGE as below.

$$\Omega_k = G_k^T G_k$$

2.1.2 Combining Summary Statistics Into a Single Super-Set

In Meta-SAIGE, summary statistics from multiple studies are consolidated into a single table, namely, super-set. Subsequently, the super-set is employed to conduct gene-based tests. In this step, S_j and V_j from each study are combined as

$$S_j = S_{j,1} + ... + S_{j,k}$$

$$V_j = V_{j,1} + ... + V_{j,k}$$

Suppose that researchers want to test a region with the first m variants ($j = 1, ..., m$). To calculate the covariance matrix of $S = (S_1, \ldots, S_m)^T$, we first calculate the correlation matrix of S, Cor, using sparse LD matrix Ω_k and the MAF of each variant obtained from each respective study. Where MAC represents the vector of minor allele counts, Cor matrix of a certain region can be expressed as:

$$\Omega = \sum_{k=1}^{K} \Omega_k, \quad F = \frac{\sum_{k=1}^{K} MAC_k}{\sum_{k=1}^{K} n_k}$$

$$Cov = \Omega/n - 4FF^T$$

$$Cor = (Cov^T * \frac{1}{\sqrt{\text{diag}(Cov)}})^T * \frac{1}{\sqrt{\text{diag}(Cov)}}$$

Then, S follows a multivariate normal distribution, given that Cor is the correlation among genetic markers

$$S \sim MVN(0, \tilde{V}^{\frac{1}{2}} Cor \tilde{V}^{\frac{1}{2}})$$

where $\tilde{V}$ is a diagonal matrix with diagonal element being V.

2.1.3 Running Gene-Based Test With the Super-Set

Prior to running the gene-based test, ultra-rare variant collapsing, a key concept that contributed to the reduction of type 1 and 2 error rates in SAIGE-GENE+, was employed. Ultra-rare variants (minor allele count (MAC) $\leq$ 10) were collapsed as a pseudo-variant and treated as a single variant. Summary statistics including S_j, V_j, and MAC were simply added up for the collapsing variants.

Subsequently, the super-set undergoes a gene-based test performed using SKAT-O, which is an optimal unified association test that combines Burden test and SKAT test (Lee et al. [2014]). The Burden method, which collapses rare variants within a specific gene, can be potent when the majority of variants in a region are causal and the effects are uni-directional. Conversely, the SKAT method, a kernel-based test method, can be more powerful when a substantial fraction of the variants in a region are noncausal or the effects of causal variants are different. The test statistics Q for Burden and SKAT can be defined as follows:

$$Q_{Burden} = (\sum_{j=1}^{m} w_j S_j)^2$$

$$Q_{SKAT} = \sum_{j=1}^{m} w_j^2 S_j^2$$

Here, w represents a weighting vector, which is generated by employing MAF and a flexible beta density function with $Beta(1, 25)$. This weighting method can be used under the assumption that rarer variants are more likely to be causal variants with larger effect sizes. The SKAT-O test is an improved method that maximizes power by leveraging the strengths of both the Burden and SKAT methods. It optimally combines the Burden and SKAT methods by striking a balance between them. The final test statistic can be derived as follows:

$$Q_{SKAT-O} = (1 - \rho)Q_{SKAT} + \rho Q_{Burden}$$

Finally, for each gene, multiple tests were conducted with various MAF cutoffs and functional annotations. P-values from multiple testing results were collectively combined using the Cauchy combination method (Liu and Xie [2020]). In this study, MAF cutoffs of 1%, 0.1%, and 0.01% were used, and for functional annotations, LoF (Loss of Function), LoF+Missense, and LoF+Missense+Synonymous were used.

2.2 Further Adjustment with GC-based Meta-Analysis Approach

In conventional meta-analysis, the aggregation of V_j across various studies is typically performed as outlined earlier. However, when dealing with rare variant association tests for highly imbalanced binary phenotypes, additional adjustments to V_j are necessary. This is due to the discrete distributions of study-specific statistics, rendering a simple summation inadequate. Hence, in the study Dey et al. [2019], a proposed solution is the utilization of genotype-count based (GC) meta-analysis to redefine the cumulative generating function (CGF) $K(t)$. Through simulation studies, GC-based meta-analysis was shown to substantially reduce the type 1 errors by more accurately estimating the score distribution.

2.3 Cohort-Specific Collapsing

To enhance the robustness of the test, Meta-SAIGE can optionally employ cohort-specific collapsing to generate a super-set. Each cohort underwent ultra-rare collapsing with a MAC of ≤ 5 before being merged into the super-set. Importantly, if an ultra-rare variant was collapsed in any of the considered cohorts due to its MAC being ≤ 5, it was treated as an ultra-rare variant across all other cohorts, regardless of whether its MAC exceeded 5 in other cohorts. This method provides a substantial benefit in terms of clinical data privacy.

As ultra-rare variants typically exhibit very low MACs, they could potentially be used to identify specific samples, which would breach the biobank's privacy policy. As a result, it is highly probable that future GWAS will conceal ultra-rare variants by collapsing, even in single-variant summaries. Evaluations on this method is provided in **Appendix**.

2.4 Type 1 Error Evaluation

The Type 1 error rate was assessed using the UK Biobank (UKB) whole-exome sequencing (WES) data of 160,000 white British samples. The null phenotypes were generated following the same procedure described in Zhou et al. [2022]. The logistic regression model used for generating the null phenotypes is presented below, where α_0 represents the intercept term determined by the phenotype prevalence (5% and 1%), Xs are the simulated covariates, and $L = 30,000$ linkage disequilibrium (LD) pruned markers from alternating chromosomes. This implies that only even chromosome markers L were used for null phenotype generation if the analysis was conducted on odd chromosomes and vice versa. Null phenotype generation was replicated 40 times to yield approximately one million tests.

$$logit(\pi_{i0}) = \alpha_0 + X_{i1} + X_{i2} + \sum_{j=1}^{L} \hat{G_{ij}}\beta$$

Subsequently, the population was divided into three cohorts with varying sample sizes (Table 2.1), ensuring that related samples were assigned to the same cohort ($relatedness \geq 0.05$). Meta-SAIGE was then applied to these cases, and the Type 1 error rates were evaluated.

Table 2.1: Sample sizes of study cohorts for type 1 error evaluation

Case	Sample size ratio	1^{st} cohort	2^{nd} cohort	3^{rd} cohort
1	1 : 1 : 1	55655	55654	50652
2	4 : 3 : 2	74205	55654	37102

2.5 Real Data Evaluation

We further evaluated Meta-SAIGE using real binary phenotypes with low prevalences, as indicated in 2.2. Type 2 diabetes (T2D) was selected to represent a phenotype with a prevalence of 5%, while glaucoma and colorectal cancer (ColCa) were chosen to represent phenotypes with a prevalence of 1%. A cohort comprising 160,000 white British samples was selected from the UK Biobank Whole Exome Sequencing (WES) dataset, and subsequently divided into two subgroups with respective sizes of $81,657$ and $85,304$. Subsequently, Meta-SAIGE was employed to perform the analysis on these cohorts. The meta-analysis results were compared to the SAIGE-GENE+ results, which was performed on the whole 160,000 white British population.

Table 2.2: List of tested binary phenotypes.

PHENOTYPES	T2D	Glaucoma	ColCa
PHECODE	250.2	365	153
CASE-CONTROL RATIO	1:22	1:91	1:89

Chapter 3

Results

3.1 Type 1 Error Evaluation

The results of the type 1 error simulation on a phenotype with a prevalence of 1% are shown in Figure 3.1. Without any adjustment for V (A), the type 1 error rate is exceedingly high. However, a substantial reduction in the type 1 error rate is observed when the SPA adjustment is applied. The GC-based adjustment further reduces the type 1 error inflation. Similar trends were discerned for a phenotype with a prevalence of 5% (refer to Figure 3.2). Table 3.1 computed and summarized the type I error rate with a significance level of $2.50 * 10^{-6}$. Despite the GC-based method proving to be the most effective in curtailing the number of type 1 errors and the associated inflation, there exist slight inflation of type 1 error. The cohort-specific method also reduced the type 1 error rate and inflation (refer to Appendix Figure 1.), albeit not to the same extent as the GC-based method.

3.2 Real Data Evaluation

Figures 3.3, 3.4, and 3.5 show scatter plots that compare the results from Meta-SAIGE and SAIGE-GENE+ for type 2 diabetes, glaucoma, and colorectal cancer. As anticipated, the meta-analysis without adjustment exhibited many false positives (A). However, they were eliminated with either the SPA (B) or the GC-based method (C).

In general, SPA and GC-based adjustment successfully preserved the true signals as indicated by SAIGE-GENE+ and exhibited high correlations (Table 3.2), without presenting any genes that were not significant in SAIGE-GENE+ results, but significant in Meta-SAIGE. It is noteworthy that the cohort-specific collapsing method also effectively maintained the signals from SAIGE-GENE+, albeit with one instance of a false positive observed in the case of glaucoma (refer to Appendix Figure 2. (B)).

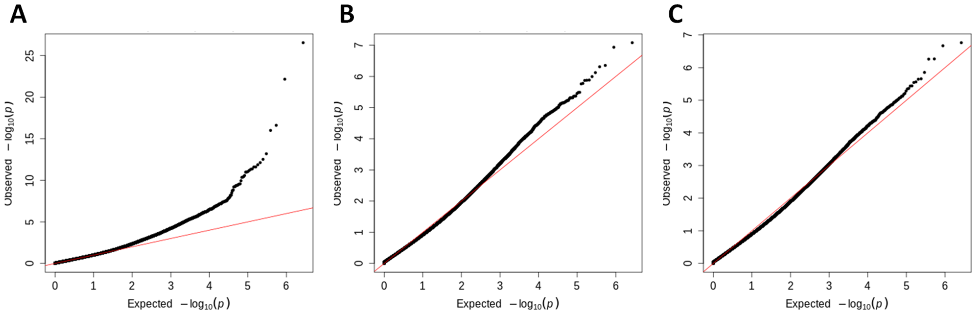


Figure 3.1: **qq-plot for prevalence 1% null phenotype** - A) qq-plot without any V adjustment. B) qq-plot with SPA adjusted V. C) qq-plot with GC-based method adjusted V.

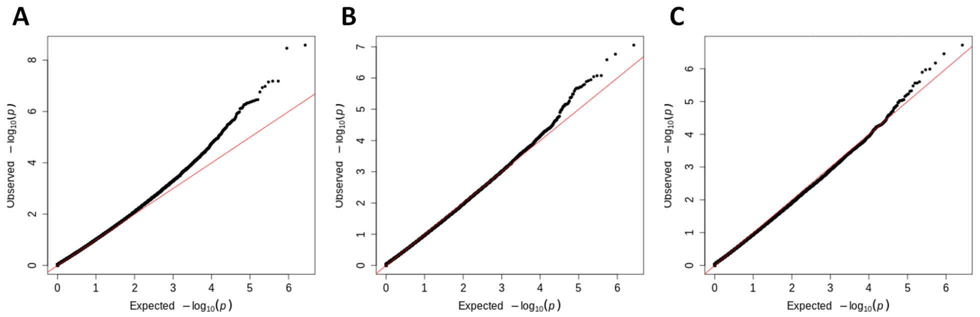


Figure 3.2: **qq-plot for prevalence 5% null phenotype** - A) qq-plot without any V adjustment. B) qq-plot with SPA adjusted V. C) qq-plot with GC-based method adjusted V.

Table 3.1: Number of false positives with significance level at $2.5 * 10^{-6}$ for null phenotypes.

METHODS	Prevalence	Number of FP	FPR
V NOT ADJUSTED	1%	311	$2.13 * 10^{-4}$
V NOT ADJUSTED	5%	41	$2.80 * 10^{-4}$
V ADJUSTED WITH SPA	1%	11	$7.53 * 10^{-6}$
V ADJUSTED WITH SPA	5%	17	$1.16 * 10^{-5}$
V ADJUSTED WITH GC-BASED METHOD	1%	7	$4.80 * 10^{-7}$
V ADJUSTED WITH GC-BASED METHOD	5%	6	$4.10 * 10^{-7}$

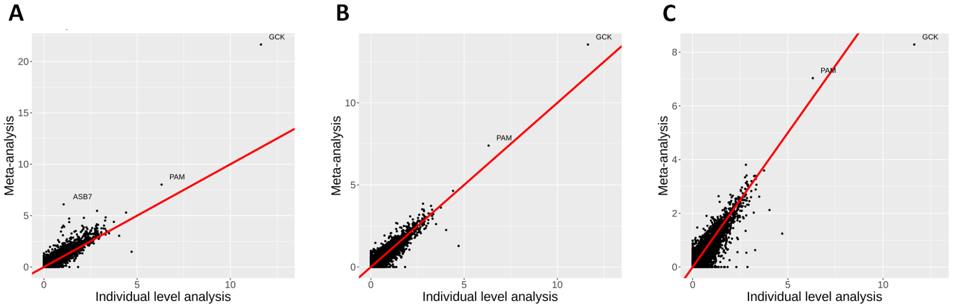


Figure 3.3: **Real data analysis on type 2 diabetes (prevalence approximately 5%)** - X-axis represents the $-log10(pval)$ of SAIGE-GENE+ and the Y-axis represents the $-log10(pval)$ of the meta-analysis. Significant genes are annotated in the plot. A) scatter plot without any V adjustment. B) scatter plot with SPA adjusted V. C) scatter plot with GC-based method adjusted V.

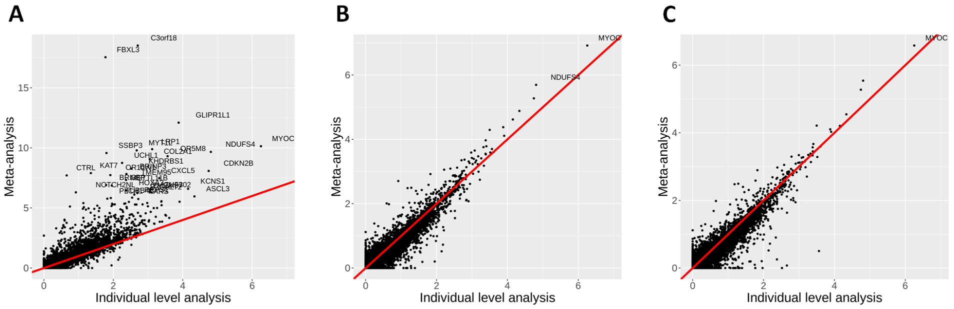


Figure 3.4: **Real data analysis on glaucoma (prevalence approximately 1%)** - X-axis represents the $-log10(pval)$ of SAIGE-GENE+ and the Y-axis represents the $-log10(pval)$ of the meta-analysis. Significant genes are annotated in the plot. A) scatter plot without any V adjustment. B) scatter plot with SPA adjusted V. C) scatter plot with GC-based method adjusted V.

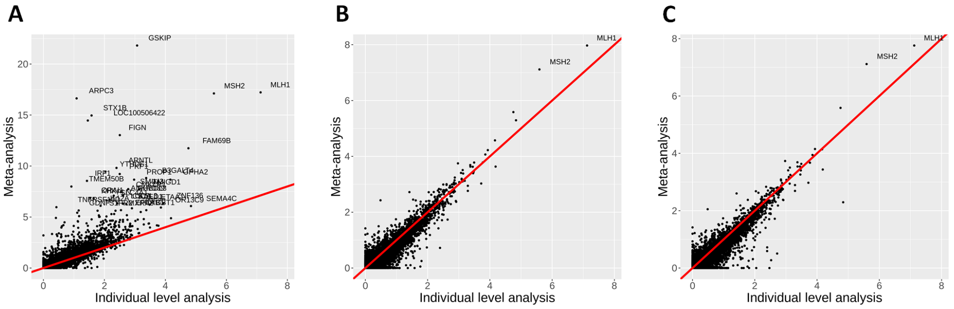


Figure 3.5: **Real data analysis on colorectal cancer (prevalence approximately 1%)** - X-axis represents the $-log10(pval)$ of SAIGE-GENE+ and the Y-axis represents the $-log10(pval)$ of the meta-analysis. Significant genes are annotated in the plot. A) scatter plot without any V adjustment. B) scatter plot with SPA adjusted V. C) scatter plot with GC-based method adjusted V.

Table 3.2: R^2 correlation between Meta-SAIGE results and SAIGE-GENE+ results of real binary phenotypes.

METHODS	T2D	Glaucoma	ColCa
V NOT ADJUSTED	0.92	0.86	0.84
V ADJUSTED WITH SPA	0.94	0.95	0.95
V ADJUSTED WITH GC-BASED METHOD	0.91	0.94	0.94

Chapter 4

Discussions

This study introduces Meta-SAIGE, a novel approach for meta-analyzing rare variants. By employing SPA and GC-based adjustment methods, Meta-SAIGE is able to more accurately estimate the distribution of score statistics, thereby reducing type 1 error rates while preserving true signals. As demonstrated in the null phenotype simulation results, substantial reduction in type 1 error rates were observed when compared to scenarios where V was not adjusted.

Examining real phenotypes demonstrated the high degree of consistency of Meta-SAIGE and SAIGE-GENE+. The findings from the Meta-SAIGE analysis not only exhibited a strong correlation with those of SAIGE-GENE+, but also showed no instances of false positives or false negatives. This highlights the potential of Meta-SAIGE as a reliable and robust method for meta-analyzing rare variants in imbalanced binary phenotypes.

As like SAIGE-GENE+, Meta-SAIGE can incorporate multiple variant sets, constructed with different functional annotations and MAF cutoffs, by testing selected variants only and combining the results using the Cauchy combination. This enables researchers to curate the effects of particular rare variants such as deleterious missense, loss of function and etc.

Meta-SAIGE offers improved computational efficiency, a significant advantage over other methods. For instance, Meta-STAAR computes the PC accounted null model for each phenotype, a process that is notably time-consuming. In contrast, Meta-SAIGE simply aggregates G^TG from each cohort. This can dramatically decrease the computational cost associated with running a meta-analysis on numerous different phenotypes, thereby facilitating a more cost-effective phenome-wide analysis.

Despite the aforementioned advantages, Meta-SAIGE is not without its limitations. A primary constraint is the degree of inflation observed when dealing with an extremely unbalanced binary phenotype. While this degree of inflation was neglectable for phenotypes with a 5% prevalence, it became larger for those with a 1% prevalence. Potential solutions could include posterior adjustment methods such as Fisher's method or supplementing more cases to relieve the case-control imbalance.

Bibliography

I. H. G. S. Consortium et al. International human genome sequencing consortium. *Nature*, 409:860–921, 2001.

H. E. Daniels. Saddlepoint approximations in statistics. *The Annals of Mathematical Statistics*, pages 631–650, 1954.

R. Dey, E. M. Schmidt, G. R. Abecasis, and S. Lee. A fast and accurate algorithm to test for binary phenotypes and its application to phewas. *The American Journal of Human Genetics*, 101(1):37–49, 2017.

R. Dey, J. B. Nielsen, L. G. Fritsche, W. Zhou, H. Zhu, C. J. Willer, and S. Lee. Robust meta-analysis of biobank-based genome-wide association studies with unbalanced binary phenotypes. *Genetic epidemiology*, 43(5):462–476, 2019.

S. Feng, D. Liu, X. Zhan, M. K. Wing, and G. R. Abecasis. Raremetal: fast and powerful meta-analysis for rare variants. *Bioinformatics*, 30(19):2828–2829, 2014.

R. J. Klein, C. Zeiss, E. Y. Chew, J.-Y. Tsai, R. S. Sackler, C. Haynes, A. K. Henning, J. P. SanGiovanni, S. M. Mane, S. T. Mayne, et al. Complement factor h polymorphism in age-related macular degeneration. *Science*, 308(5720):385–389, 2005.

S. Lee, M. J. Emond, M. J. Bamshad, K. C. Barnes, M. J. Rieder, D. A. Nickerson, D. C. Christiani, M. M. Wurfel, X. Lin, N. G. E. S. Project, et al. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *The American Journal of Human Genetics*, 91(2):224–237, 2012.

S. Lee, G. R. Abecasis, M. Boehnke, and X. Lin. Rare-variant association analysis: study designs and statistical tests. *The American Journal of Human Genetics*, 95(1): 5–23, 2014.

X. Li, C. Quick, H. Zhou, S. M. Gaynor, Y. Liu, H. Chen, M. S. Selvaraj, R. Sun, R. Dey, D. K. Arnett, et al. Powerful, scalable and resource-efficient meta-analysis of rare variant associations in large whole genome sequencing studies. *Nature genetics*, 55(1):154–164, 2023.

Y. Liu and J. Xie. Cauchy combination test: a powerful test with analytic p-value calculation under arbitrary dependency structures. *Journal of the American Statistical Association*, 115(529):393–402, 2020.

T. A. Manolio, F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, et al. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, 2009.

J. Ott, J. Wang, and S. M. Leal. Genetic linkage analysis in the age of whole-genome sequencing. *Nature Reviews Genetics*, 16(5):275–284, 2015.

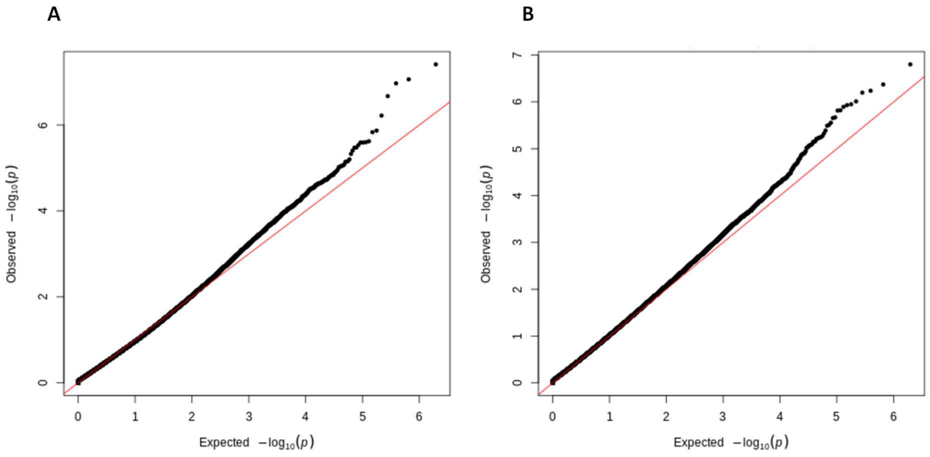
J. A. Reuter, D. V. Spacek, and M. P. Snyder. High-throughput sequencing technologies. *Molecular cell*, 58(4):586–597, 2015.

P. M. Visscher, N. R. Wray, Q. Zhang, P. Sklar, M. I. McCarthy, M. A. Brown, and J. Yang. 10 years of gwas discovery: biology, function, and translation. *The American Journal of Human Genetics*, 101(1):5–22, 2017.

W. Zhou, W. Bi, Z. Zhao, K. K. Dey, K. A. Jagadeesh, K. J. Karczewski, M. J. Daly, B. M. Neale, and S. Lee. Saige-gene+ improves the efficiency and accuracy of set-based rare variant association tests. *Nature genetics*, 54(10):1466–1469, 2022.

Appendix

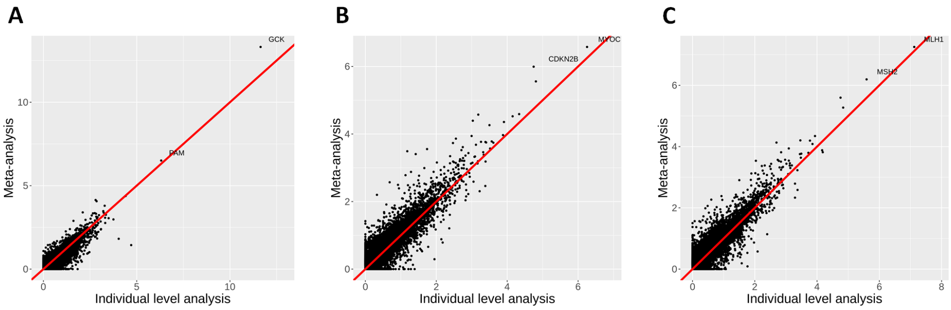
Results for Cohort-Specific Collapsing method



Appendix Figure 1: Type 1 error evaluation on cohort-specific collapsing method using simulated null phenotype - A) qq-plot for prevalence 1% phenotype. B) qq-plot for prevalence 5% phenotype.

Appendix Table 1: Number of false positives with significance level at $2.5 * 10^{-6}$ for null phenotypes for cohort-specific collapsing method.

METHODS	Prevalence	Number of FP	FPR
V ADJUSTED WITH GC-BASED METHOD + COHORT-SPECIFIC COLLAPSING	1%	8	$7.53 * 10^{-6}$
V ADJUSTED WITH GC-BASED METHOD + COHORT-SPECIFIC COLLAPSING	5%	12	$1.12 * 10^{-5}$



Appendix Figure 2: Real data analysis on cohort-specific collapsing method - X-axis represents the $-log10(pval)$ of SAIGE-GENE+ and the Y-axis represents the $-log10(pval)$ of the meta-analysis. Significant genes are annotated in the plot. A) scatter plot for T2D ($R^2 = 0.92$). B) scatter plot for glaucoma ($R^2 = 0.92$). C) scatter plot for ColCa ($R^2 = 0.91$).

초 록

희귀 유전 변이들은 다양한 표현형 발현에 중요한 인자로 여겨진다. 하지만 희귀성은 일반적인 변이 단위의 연관성 분석을 어렵게 하며, 이에따라 유전자 기반의 메타분석이 그 해결방법으로 제시되었다. 많은 희귀 유전 변이들의 메타분석 방법들이 개발되었으나 이는 불균형한 이진표현형을 분석할 때 제 1종 오류 인플레이션이 심한 것으로 나타났고, 높은 계산 비용이 발생한다는 한계가 있다.

이러한 한계를 해결하기 위해, 본 논문에서는 Meta-SAIGE를 소개한다. Meta-SAIGE는 점수 분포의 정확한 추정을 통해 제 1종 오류 인플레이션을 줄이고, 인구계층화를 고려한 영 모델을 계산하지 않아 효율성을 향상시킨다. 또한, 바이오뱅크의 개인 정보 보호 정책과 일치하는 초희귀 변이 처리 방법을 제안한다.

이 연구에서는 UKBiobank의 Whole Exome Sequencing 데이터를 사용하여 Meta-SAIGE의 성능을 평가했다. 시뮬레이션된 표현형을 사용하여 제 1종 오류율을 평가하고, 유병율이 불균형한 실제 질병들을 사용하여 메타 분석 결과가 일반적인 유전자 기반 희귀 변이 연관성 검사(SAIGE-GENE+)와 일치하는지 확인했다.

주요어: Meta-Analysis, Rare-variant, Gene-based association test

학번: 2021-27059