Heritability Estimation and Risk Prediction in Schizophrenia

Choi Shing Wan

A thesis submitted in partial fulfillment of the
requirements for
the Degree of Doctor of Philosophy

Department of Psychiatry
University of Hong Kong
Hong Kong

November 22, 2015

Declaration

I declare that this thesis represents my own work, except where due acknowledgments is made, and that it has not been previously included in a thesis, dissertation or report submitted to this University or to any other institution for a degree, diploma or other qualification.

Signed.....

Acknowledgements

Abbreviations

CATIE Clinical Antipsychotic Trials of Intervention Effectiveness. 42

CEU Northern Europeans from Utah. 45, 49, 52, 55, 57

GCTA Genome-wide Complex Trait Analysis. 33, 50, 52–54, 56, 59–63, 65, 67–80

GWAS Genome Wide Association Study. 34, 35, 42, 46, 50–52, 56, 67

IBD identity by descent. 57

LD Linkage Disequilibrium. 34–36, 39, 43, 45–50, 52, 53, 56–59

LDSC LD SCore. 33, 47, 50, 51, 53, 54, 56, 59–63, 65–80

maf Minor Allele Frequency. 49, 50, 52–55

mb megabase. 46, 51

MSE mean squared error. 63, 67

NCP non-centrality parameter. 41

PCGC Phenotype correlation - genotype correlation regression. 33

PGC Psychiatric Genomics Consortium. 51, 57

SE standard error. 40, 41, 51, 63

SHREK SNP Heritability and Risk Estimation Kit. 57–80

SNP Single Nucleotide Polymorphism. 33–37, 39, 45–47, 49–57, 59–67, 69, 71, 73, 74, 76, 77, 79, 80

SVD Singular Value Decomposition. 43, 46

tSVD Truncated Singular Value Decomposition. 43–46

Contents

List of Figures

List of Tables

Chapter 2

Heritability Estimation

2.1 Introduction

The development of LD SCore has brought great prospect in estimating the heritability of complex disease for one can now estimate the heritability of a trait without requiring the rare genotype. However, as noted by the author of LD SCore (LDSC), when the number of causal variants were small, or when working on targeted genotype array, LDSC tends to have a larger standard error or might produce funky results(B. K. Bulik-Sullivan et al., 2015). Ideally, we would like to be able to robustly estimate the heritability for all traits, disregarding the genetic architecture (e.g. number of causal Single Nucleotide Polymorphisms (SNPs)).

On the other hand, it has been shown that there can be huge bias in the heritability estimation of Genome-wide Complex Trait Analysis (GCTA) when prevalence of a dichotomous trait is low(Golan, Eric S Lander, and Rosset, 2014). Although Golan, Eric S Lander, and Rosset (2014) developed the Phenotype correlation - genotype correlation regression (PCGC), which can provide robust estimation of heritability for traits with different prevalence, it still relies on the relationship matrix and therefore require the raw genotype of the samples.

Herein, we would like to develop an alternative algorithm to LDSC for heritability estimation using only the test statistics. We would also like to inspect whether if LDSC's heritability estimation is robust to prevalence of a trait. A number of simulations were performed to compare the performance of LDSC and our algorithm under different conditions.

The work in this chapter were done in collaboration with my colleagues who have

kindly provide their support and knowledges to make this piece of work possible. Dr Johnny Kwan, Dr Miaxin Li and Professor Sham have helped to laid the framework of this study. Dr Timothy Mak has derived the mathematical proof for our heritability estimation method. Miss Yiming Li, Dr Johnny Kwan, Dr Miaxin Li, Dr Timothy Mak and Professor Sham have helped with the derivation of the standard error of the heritability estimation. Dr Henry Leung has provided critical suggestions on the implementation of the algorithm.

2.2 Methodology

The overall aims of this study is to develop a robust algorithm for the estimation of the narrow sense heritability using only the summary statistic from a Genome Wide Association Study (GWAS). In GWAS, the test statistic of a particular SNP should be proportional to its effect size and the effect size from all the other SNPs in Linkage Disequilibrium (LD) with it. Based on this property, we may use the information from the LD matrix and the test statistic of the GWAS SNP the estimate the narrow sense heritability.

2.2.1 Heritability Estimation

Remember that the narrow-sense heritability is defined as

$$h^2 = \frac{\mathrm{Var}(X)}{\mathrm{Var}(Y)}$$

where $\mathrm{Var}(X)$ is the variance of the genotype and $\mathrm{Var}r(Y)$ is the variance of the phenotype. In a GWAS, regression were performed between the SNPs and the phenotypes, giving

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

where Y and X are the standardized phenotype and genotype respectively. ϵ is then the error term, accounting for the non-genetic elements contributing to the phenotype (e.g. Environment factors). Based on eq. (2.1), one can then have

$$\begin{aligned} \mathrm{Var}(Y) &= \mathrm{Var}(\beta X) + \mathrm{Var}(\epsilon) \\ \mathrm{Var}(Y) &= \beta^{\mathrm{Var}}(X) \\ \beta^2 \frac{\mathrm{Var}(X)}{\mathrm{Var}(Y)} &= 1 \end{aligned} \tag{2.2}$$

β^2 is then considered as the portion of phenotype variance explained by the variance of genotype, which can also be considered as the narrow-sense heritability of the phenotype.

A challenge in calculating the heritability from GWAS data is that usually only the test-statistic or p-value were provided and one will not be able to directly calculate the heritability based on eq. (2.2). In order to estimation the heritability of a trait from the GWAS test-statistic, we first observed that when both X and Y are standardized, β^2 will be equal to the coefficient of determination (r^2). Then, based on properties of the Pearson product-moment correlation coefficient:

$$r = \frac{t}{\sqrt{n-2+t^2}} \tag{2.3}$$

where t follows the student-t distribution and n is the number of samples, one can then obtain the r^2 by taking the square of eq. (2.3)

$$r^2 = \frac{t^2}{n-2+t^2} \tag{2.4}$$

It is observed that t^2 will follow the F-distribution. When n is big, t^2 will converge into χ^2 distribution.

Furthermore, when the effect size is small and n is big, r^2 will be approximately χ^2 distributed with mean ~ 1. We can then approximate eq. (2.4) as

$$r^2 = \frac{\chi^2}{n} \tag{2.5}$$

and define the *observed* effect size of each SNP to be

$$f = \frac{\chi^2 - 1}{n} \tag{2.6}$$

When there are LD between each individual SNPs, the situation will become more complicated as each SNPs' observed effect will contains effect coming from other SNPs in LD with it:

$$f_{observed} = f_{true} + f_{LD} \tag{2.7}$$

To account for the LD structure, we first assume our phenotype $\boldsymbol{Y}$ and genotype $\boldsymbol{X} = (X_1, X_2, \ldots, X_m)^t$ are standardized and that

$$\boldsymbol{Y} \sim f(0, 1)$$
$$\boldsymbol{X} \sim f(0, \boldsymbol{R})$$

Where $\boldsymbol{R}$ is the LD matrix between SNPs.

We can then express eq. (2.1) in matrix form:

$$\boldsymbol{Y} = \boldsymbol{\beta}^t \boldsymbol{X} + \epsilon \tag{2.8}$$

Because the phenotype is standardized with variance of 1, the narrow sense heritability can then be expressed as

$$\begin{aligned} Heritability &= \frac{\mathrm{Var}(\boldsymbol{\beta}^t \boldsymbol{X})}{\mathrm{Var}(\boldsymbol{Y})} \\ &= \mathrm{Var}(\boldsymbol{\beta}^t \boldsymbol{X}) \end{aligned} \tag{2.9}$$

If we then assume now that $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_m)^t$ has distribution

$$\begin{aligned} \boldsymbol{\beta} &\sim f(0, H) \\ \boldsymbol{H} &= diag(\boldsymbol{h}) \\ \boldsymbol{h} &= (h_1^2, h_2^2, \ldots, h_m^2)^t \end{aligned}$$

where $\boldsymbol{H}$ is the variance of the "true" effect. It is shown that heritability can be expressed as

$$\begin{aligned} \mathrm{Var}(\boldsymbol{\beta}^t \boldsymbol{X}) &= \mathrm{E}_X \mathrm{Var}_{\beta|X}(\boldsymbol{X}^t \boldsymbol{\beta}) + \mathrm{Var}_X \mathrm{E}_{(\beta|X)}(\boldsymbol{\beta}^2 \boldsymbol{X}) \\ &= \mathrm{E}_X(\boldsymbol{X}^t \boldsymbol{\beta} \boldsymbol{\beta}^T \boldsymbol{X}) \\ &= \mathrm{E}_X(\boldsymbol{X}^t \boldsymbol{H} \boldsymbol{X}) \\ &= \mathrm{E}(\boldsymbol{X})^t \boldsymbol{H} \mathrm{E}(\boldsymbol{X}) + \mathrm{Tr}(\mathrm{Var}(\boldsymbol{X}\boldsymbol{H})) \\ &= \mathrm{Tr}(\mathrm{Var}(\boldsymbol{X}\boldsymbol{H})) \\ &= \sum_i h_i^2 \end{aligned} \tag{2.10}$$

Now if we consider the covariance between SNP i ($\boldsymbol{X_i}$) and $\boldsymbol{Y}$, we have

$$\begin{aligned} \mathrm{Cov}(\boldsymbol{X}_i, \boldsymbol{Y}) &= \mathrm{Cov}(\boldsymbol{X}_i, \boldsymbol{\beta}^t \boldsymbol{X} + \epsilon) \\ &= \mathrm{Cov}(\boldsymbol{X}_i, \boldsymbol{\beta}^t \boldsymbol{X}) \\ &= \sum_j \mathrm{Cov}(\boldsymbol{X}_i, \boldsymbol{X}_j) \boldsymbol{\beta}_j \\ &= \boldsymbol{R}_i \boldsymbol{\beta}_j \end{aligned} \tag{2.11}$$

As both $\boldsymbol{X}$ and $\boldsymbol{Y}$ are standardized, the covariance will equal to the correlation and we can define the correlation between SNP i and Y as

$$\rho_i = \boldsymbol{R}_i \boldsymbol{\beta}_j \tag{2.12}$$

In reality, the *observed* correlation usually contains error. Therefore we define the *observed* correlation between SNP_i and the phenotype($\hat{\rho}_i$) to be

$$\hat{\rho}_i = \rho_i + \frac{\epsilon_i}{\sqrt{n}} \tag{2.13}$$

for some error ϵ_i. The distribution of the correlation coefficient about the true correlation ρ is approximately

$$\hat{\rho}_i \sim f(\rho_i, \frac{(1-\rho^2)^2}{n})$$

By making the assumption that ρ_i is close to 0 for all i, we have

$$\begin{aligned} \text{E}(\epsilon_i|\rho_i) &\sim 0 \\ \text{Var}(\epsilon_i|\rho_i) &\sim 1 \end{aligned}$$

We then define our z-statistic and χ^2-statistic as

$$\begin{aligned} z_i &= \hat{\rho}_i \sqrt{n} \\ \chi^2 &= z_i^2 \\ &= \hat{\rho}_i{}^2 n \end{aligned}$$

From eq. (2.13) and eq. (2.12), χ^2 can then be expressed as

$$\begin{aligned} \chi^2 &= \hat{\rho}^2 n \\ &= n(\boldsymbol{R}_i \boldsymbol{\beta}_j + \frac{\epsilon_i}{\sqrt{n}})^2 \end{aligned}$$

The expectation of χ^2 is then

$$\begin{aligned} \text{E}(\chi^2) &= n(\boldsymbol{R}_i \boldsymbol{\beta} \boldsymbol{\beta}^t \boldsymbol{R}_i + 2\boldsymbol{R}_i \boldsymbol{\beta} \frac{\epsilon_i}{\sqrt{n}} + \frac{\epsilon_i^2}{n}) \\ &= n\boldsymbol{R}_i \boldsymbol{H} \boldsymbol{R}_i + 1 \end{aligned}$$

To derive least square estimates of h_i^2, we need to find $\hat{h_i^2}$ which minimizes

$$\begin{aligned}\sum_i(\chi_i^2 - \mathrm{E}(\chi_i^2))^2 &= \sum_i(\chi_i^2 - (n\boldsymbol{R}_i\boldsymbol{H}\boldsymbol{R}_i + 1))^2 \\ &= \sum_i(\chi_i^2 - 1 - n\boldsymbol{R}_i\boldsymbol{H}\boldsymbol{R}_i)^2\end{aligned}$$

If we define

$$f_i = \frac{\chi_i^2 - 1}{n} \tag{2.14}$$

we got

$$\begin{aligned}\sum_i(\chi_i^2 - \mathrm{E}(\chi_i^2))^2 &= \sum_i(f_i - \boldsymbol{R}_i\boldsymbol{H}\boldsymbol{R}_i)^2 \\ &= \boldsymbol{f}\boldsymbol{f}^t - 2\boldsymbol{f}^t\boldsymbol{R}_{\boldsymbol{sq}}\hat{\boldsymbol{h}} + \hat{\boldsymbol{h}}^t\boldsymbol{R}_{\boldsymbol{sq}}^t\boldsymbol{R}_{\boldsymbol{sq}}\hat{\boldsymbol{h}}\end{aligned} \tag{2.15}$$

where $\boldsymbol{R_{sq}} = \boldsymbol{R} \circ \boldsymbol{R}$. By differentiating eq. (2.15) w.r.t $\hat{h}$ and set to 0, we get

$$\begin{aligned}2\boldsymbol{R}_{\boldsymbol{sq}}^t\boldsymbol{R}_{\boldsymbol{sq}}\hat{\boldsymbol{h^2}} - 2\boldsymbol{R}_{\boldsymbol{sq}}\boldsymbol{f} &= 0 \\ \boldsymbol{R}_{\boldsymbol{sq}}\hat{\boldsymbol{h^2}} &= \boldsymbol{f}\end{aligned} \tag{2.16}$$

And the heritability is then defined as

$$\hat{Heritability} = \mathbf{1}^t\boldsymbol{R}_{\boldsymbol{sq}}^{-1}\boldsymbol{f} \tag{2.17}$$

2.2.2 Calculating the Standard error

From eq. (2.17), we can derive the variance of heritability H as

$$\begin{aligned}\mathrm{Var}(H) &= \mathrm{E}[H^2] - \mathrm{E}[H]^2 \\ &= \mathrm{E}[(\mathbf{1}^t\boldsymbol{R}_{\boldsymbol{sq}}^{-1}\boldsymbol{f})^2] - \mathrm{E}[\mathbf{1}^t\boldsymbol{R}_{\boldsymbol{sq}}^{-1}\boldsymbol{f}](\mathrm{E}[\mathbf{1}^t\boldsymbol{R}_{\boldsymbol{sq}}^{-1}\boldsymbol{f}])^t \\ &= \mathrm{E}[\mathbf{1}^t\boldsymbol{R}_{\boldsymbol{sq}}^{-1}\boldsymbol{f}\boldsymbol{f}^t\boldsymbol{R}_{\boldsymbol{sq}}^{-1}\mathbf{1}] - \mathrm{E}[\mathbf{1}^t\boldsymbol{R}_{\boldsymbol{sq}}^{-1}\boldsymbol{f}](\mathrm{E}[\mathbf{1}^t\boldsymbol{R}_{\boldsymbol{sq}}^{-1}\boldsymbol{f}])^t \\ &= \mathbf{1}^t\boldsymbol{R}_{\boldsymbol{sq}}^{-1}\mathrm{E}[\boldsymbol{f}\boldsymbol{f}^t]\boldsymbol{R}_{\boldsymbol{sq}}^{-1}\mathbf{1} - \mathrm{E}[\mathbf{1}^t\boldsymbol{R}_{\boldsymbol{sq}}^{-1}\boldsymbol{f}](\mathrm{E}[\mathbf{1}^t\boldsymbol{R}_{\boldsymbol{sq}}^{-1}\boldsymbol{f}])^t \\ &= \mathbf{1}^t\boldsymbol{R}_{\boldsymbol{sq}}^{-1}\mathrm{Var}(\boldsymbol{f})\boldsymbol{R}_{\boldsymbol{sq}}^{-1}\mathbf{1} + \mathrm{E}[\mathbf{1}^t\boldsymbol{R}_{\boldsymbol{sq}}^{-1}\boldsymbol{f}](\mathrm{E}[\mathbf{1}^t\boldsymbol{R}_{\boldsymbol{sq}}^{-1}\boldsymbol{f}])^t - \mathrm{E}[\mathbf{1}^t\boldsymbol{R}_{\boldsymbol{sq}}^{-1}\boldsymbol{f}](\mathrm{E}[\mathbf{1}^t\boldsymbol{R}_{\boldsymbol{sq}}^{-1}\boldsymbol{f}])^t \\ &= \mathbf{1}^t\boldsymbol{R}_{\boldsymbol{sq}}^{-1}\mathrm{Var}(\boldsymbol{f})\boldsymbol{R}_{\boldsymbol{sq}}^{-1}\mathbf{1}\end{aligned} \tag{2.18}$$

Therefore, to obtain the variance of H, we first need to calculate the variance covariance matrix of $\boldsymbol{f}$.

We first consider the standardized genotype X_i with standard normal mean z_i and

non-centrality parameter μ_i, we have

$$\begin{aligned}
\mathrm{E}[X_i] &= \mathrm{E}[z_i + \mu_i] \\
&= \mu_i \\
\mathrm{Var}(X_i) &= \mathrm{E}[(z_i + \mu_i)^2] + \mathrm{E}[(z_i + \mu_i)]^2 \\
&= \mathrm{E}[z_i^2 + \mu_i^2 + 2z_i\mu_i] + \mu_i^2 \\
&= 1 \\
\mathrm{Cov}(X_i, X_j) &= \mathrm{E}[(z_i + \mu_i)(z_j + \mu_j)] - \mathrm{E}[z_i + \mu_i]\mathrm{E}[z_j + \mu_j] \\
&= \mathrm{E}[z_i z_j + z_i\mu_j + \mu_i z_j + \mu_i\mu_j] - \mu_i\mu_j \\
&= \mathrm{E}[z_i z_j] + \mathrm{E}[z_i\mu_j] + \mathrm{E}[z_j\mu_i] + \mathrm{E}[\mu_i\mu_j] - \mu_i\mu_j \\
&= \mathrm{E}[z_i z_j]
\end{aligned}$$

As the genotypes are standardized, therefore $\mathrm{Cov}(X_i, X_j) == \mathrm{Cor}(X_i, X_j)$, we can obtain

$$\mathrm{Cov}(X_i, X_j) = \mathrm{E}[z_i z_j] = R_{ij}$$

where R_{ij} is the LD between SNP_i and SNP_j. Given these information, we can then calculate $\mathrm{Cov}(\chi_i^2, \chi_j^2)$ as:

$$\begin{aligned}
\mathrm{Cov}(X_i^2, X_j^2) =& \mathrm{E}[(z_i + \mu_i)^2(z_j + \mu_j)^2] - \mathrm{E}[z_i + \mu_i]\mathrm{E}[z_j + \mu_j] \\
=& \mathrm{E}[(z_i^2 + \mu_i^2 + 2z_i\mu_i)(z_j^2 + \mu_j^2 + 2z_j\mu_j)] \\
& - \mathrm{E}[z_i^2 + \mu_i^2 + 2z_i\mu_i]\mathrm{E}[z_j^2 + \mu_j^2 + 2z_j\mu_j] \\
=& \mathrm{E}[(z_i^2 + \mu_i^2 + 2z_i\mu_i)(z_j^2 + \mu_j^2 + 2z_j\mu_j)] \\
& - (\mathrm{E}[z_i^2] + \mathrm{E}[\mu_i^2] + 2\mathrm{E}[z_i\mu_i])(\mathrm{E}[z_j^2] + \mathrm{E}[\mu_j^2] + 2\mathrm{E}[z_j\mu_j]) \\
=& \mathrm{E}[z_i^2(z_j^2 + \mu_j^2 + 2z_j\mu_j) + \mu_i^2(z_j^2 + \mu_j^2 + 2z_j\mu_j) + 2z_i\mu_i(z_j^2 + \mu_j^2 + 2z_j\mu_j)] \\
& - (1 + \mu_i^2)(1 + \mu_j^2) \\
=& \mathrm{E}[z_i^2(z_j^2 + \mu_j^2 + 2z_j\mu_j)] + \mu_i^2\mathrm{E}[z_j^2 + \mu_j^2 + 2z_j\mu_j] \\
& + 2\mu_i\mathrm{E}[z_i(z_j^2 + \mu_j^2 + 2z_j\mu_j)] - (1 + \mu_i^2)(1 + \mu_j^2) \\
=& \mathrm{E}[z_i^2 z_j^2 + z_i^2\mu_j^2 + 2z_i^2 z_j\mu_j] + \mu_i^2 + \mu_i^2\mu_j^2 \\
& + 2\mu_i\mathrm{E}[z_i z_j^2 + z_i\mu_j^2 + 2z_i z_j\mu_j] - (1 + \mu_i^2)(1 + \mu_j^2) \\
=& \mathrm{E}[z_i^2 z_j^2] + \mu_j^2 + \mu_i^2 + \mu_i^2\mu_j^2 + 4\mu_i\mu_j\mathrm{E}[z_i z_j] - (1 + \mu_i^2 + \mu_j^2 + \mu_i\mu_j) \\
=& \mathrm{E}[z_i^2 z_j^2] + 4\mu_i\mu_j\mathrm{E}[z_i z_j] - 1
\end{aligned}$$

Remember that $\mathrm{E}[z_i z_j] = R_{ij}$, we then have

$$\mathrm{Cov}(X_i^2, X_j^2) = \mathrm{E}[z_i^2 z_j^2] + 4\mu_i\mu_j R_{ij} - 1$$

By definition,

$$z_i | z_j \sim N(\mu_i + R_{ij}(z_j - \mu_j), 1 - R_{ij}^2)$$

We can then calculate $\mathrm{E}[z_i^2 z_j^2]$ as

$$\begin{aligned}
\mathrm{E}[z_i^2 z_j^2] &= \mathrm{Var}[z_i z_j] + \mathrm{E}[z_i z_j]^2 \\
&= \mathrm{E}[\mathrm{Var}(z_i z_j | z_i)] + \mathrm{Var}[\mathrm{E}[z_i z_j | z_i]] + R_{ij}^2 \\
&= \mathrm{E}[z_j^2 \mathrm{Var}(z_i | z_j)] + \mathrm{Var}[z_j \mathrm{E}[z_i | z_j]] + R_{ij}^2 \\
&= (1 - R_{ij}^2)\mathrm{E}[z_j^2] + \mathrm{Var}(z_j(\mu_i + R_{ij}(z_j - \mu_j))) + R_{ij}^2 \\
&= (1 - R_{ij}^2) + \mathrm{Var}(z_j\mu_i + R_{ij} z_j^2 - \mu_j z_j R_{ij}) + R_{ij}^2 \\
&= 1 + \mu_i^2 \mathrm{Var}(z_j) + R_{ij}^2 \mathrm{Var}(z_j^2) - \mu_j^2 R_{ij}^2 \mathrm{Var}(z_j) \\
&= 1 + 2R_{ij}^2
\end{aligned}$$

As a result, the variance covariance matrix of the χ^2 variances represented as

$$\mathrm{Cov}(X_i^2, X_j^2) = 2R_{ij}^2 + 4R_{ij}\mu_i\mu_j \tag{2.19}$$

As we only have the *observed* expectation, we should re-define eq. (2.19) as

$$\mathrm{Cov}(X_i^2, X_j^2) = \frac{2R_{ij}^2 + 4R_{ij}\mu_i\mu_j}{n^2} \tag{2.20}$$

where n is the sample size.

By substituting eq. (2.20) into eq. (2.18), we will get

$$\mathrm{Var}(H) = \mathbf{1}^t \boldsymbol{R}_{\boldsymbol{sq}}^{-1} \frac{2\boldsymbol{R}_{\boldsymbol{sq}} + 4\boldsymbol{R} \circ \boldsymbol{z}\boldsymbol{z}^t}{n^2} \boldsymbol{R}_{\boldsymbol{sq}}^{-1} \mathbf{1} \tag{2.21}$$

where $\boldsymbol{z} = \sqrt{\boldsymbol{\chi}^{\mathbf{2}}}$ from eq. (2.14), with the direction of effect as its sign and $\circ$ is the element-wise product (Hadamard product).

The problem with eq. (2.21) is that it requires the direction of effect. Without the direction of effect, the estimation of standard error (SE) will be inaccurate. If we consider that $\boldsymbol{f}$ is approximately χ^2 distributed, we might view eq. (2.16) as a decomposition of a vector of χ^2 distributions with degree of freedom of 1. Replacing the vector $\boldsymbol{f}$ with a vector of 1, we can perform the decomposition of the degree of freedom, getting the "effective

number"(e) of the association(M.-X. X. Li et al., 2011). Substituting e into the variance equation of non-central χ^2 distribution will yield

$$\mathrm{Var}(H) = \frac{2(e + 2H)}{n^2} \tag{2.22}$$

eq. (2.22) should in theory gives us an heuristic estimation of the SE. Moreover, the direction of effect was not required for eq. (2.22), reducing the number of input required from the user.

2.2.3 Case Control Studies

When dealing with case control data, as the phenotype were usually discontinuous, we cannot directly use eq. (2.17) to estimate the heritability. Instead, we will need to employ the concept of liability threshold model from section 1.6.

Based on the derivation of Jian Yang, Naomi R. Wray, and Peter M. Visscher (2010), the approximate ratio between the non-centrality parameter (NCP) obtained from case control studies (NPC_{CC}) and quantitative trait studies(NCP_{QT}) were

$$\frac{NCP_{CC}}{NCP_{QT}} = \frac{i^2 v(1-v)N_{CC}}{(1-K)^2 N_{QT}} \tag{2.23}$$

where

$$\begin{aligned} K &= \text{Population Prevalence} \\ v &= \text{Proportion of Cases} \\ N &= \text{Total Number of Samples} \\ i &= \frac{z}{K} \\ z &= \text{height of standard normal curve at truncation pretained to K} \end{aligned}$$

Using this approximation deviated by Jian Yang, Naomi R. Wray, and Peter M. Visscher (2010), we can directly transform the NCP between the case control studies and quantitative trait studies. As we were transforming the NCP of a single study, the N_{CC} and N_{QT} will be the same, therefore eq. (2.23) became

$$NCP_{QT} = \frac{NCP_{CC}(1-K)^2}{i^2 v(1-v)} \tag{2.24}$$

By combining eq. (2.24) and eq. (2.14), we can then have

$$f = \frac{(\chi^2_{CC} - 1)(1-K)^2}{ni^2v(1-v)} \tag{2.25}$$

where χ^2_{CC} is the test statistic from the case control association test. Finally, the heritability estimation of case control studies can be simplified to

$$\hat{Heritability} = \frac{(1-K)^2}{i^2v(1-v)}\mathbf{1}^t\boldsymbol{R}_{\boldsymbol{sq}}^{-1}\boldsymbol{f} \tag{2.26}$$

2.2.4 Extreme Phenotype Selections

Although the development of GWAS now provide unprecedented power to perform hypothesis free association throughout the whole genome, studies of complex traits still require a large amount of samples, which sometimes are difficult to obtain. For example, in the studies of antipsychotic treatment response, the largest GWAS performed by the Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) project only contain 738 subjects. To assist the identification of causal variants with a small effect size, a larger power is required. A common technique is to perform extreme phenotype selection in the detection stage of the study. The extreme phenotype selection will inflate the frequency distortion between samples from the two extreme end of phenotype and thus increase the statistical power (Guey et al., 2011). It was estimated that for a 0.5% variant with a fivefold effect in the general population, a discovery studies using extreme phenotype selection requires four times less samples in the replication to achieve 80% power when compared to studies using random samples (Guey et al., 2011). This allows studies to be conducted using a smaller amount of samples with the same degree of power which is vital for studies where it is difficult to obtain a large sample size.

A problem of extreme phenotype selection was that the variance of the selected phenotype will not be representative of that in the population. The effect size are generally overestimated (Guey et al., 2011). Thus, to adjust for this bias, one can multiple the estimated heritability $\hat{h^2}$ by the ratio between the variance before V_P and after $V_{P'}$ the selection process (Pak C Sham and S. M. Purcell, 2014):

$$\hat{Heritability} = \frac{V_{P'}}{V_P}\mathbf{1}^t\boldsymbol{R}_{\boldsymbol{sq}}^{-1}\boldsymbol{f} \tag{2.27}$$

2.2.5 Inverse of the Linkage Disequilibrium matrix

In order to obtain the heritability estimation, we will require to solve eq. (2.17). If $\boldsymbol{R_{sq}}$ is of full rank and positive semi-definite, it will be straight-forward to solve the matrix equation. However, more often than not, the LD matrix are rank-deficient and suffer from multicollinearity, making it ill-conditioned, therefore highly sensitive to changes or errors in the input. To be exact, we can view eq. (2.17) as calculating the sum of $\hat{\boldsymbol{h}}^{\boldsymbol{2}}$ from eq. (2.16). This will involve solving for

$$\hat{\boldsymbol{h}}^{\boldsymbol{2}} = \boldsymbol{R}_{\boldsymbol{sq}}^{-1}\boldsymbol{f} \tag{2.28}$$

where an inverse of $\boldsymbol{R_{sq}}$ is observed.

In normal circumstances (e.g. when $\boldsymbol{R_{sq}}$ is full rank and positive semi-definite), one can easily solve eq. (2.28) using the QR decomposition or LU decomposition. However, when $\boldsymbol{R_{sq}}$ is ill-conditioned, the traditional decomposition method will fail. Even if the decomposition is successfully performed, the result tends to be a meaningless approximation to the true $\hat{\boldsymbol{h}}^{\boldsymbol{2}}$.

Therefore, to obtain a meaningful solution, regularization techniques such as the Tikhonov Regularization (also known as Ridge Regression) and Truncated Singular Value Decomposition (tSVD) has to be performed(Neumaier, 1998). There are a large variety of regularization techniques, yet the discussion of which is beyond the scope of this study. In this study, we will focus on the use of tSVD in the regularization of the LD matrix. This is because the Singular Value Decomposition (SVD) routine has been implemented in the EIGEN C++ library (Guennebaud and Jacob, 2010), allowing us to implement the tSVD method without much concern with regard to the detail of the algorithm.

To understand the problem of the ill-conditioned matrix and regularization method, we consider the matrix equation $\boldsymbol{Ax} = \boldsymbol{B}$ where $\boldsymbol{A}$ is ill-conditioned or singular with $n \times n$ dimension. The SVD of $\boldsymbol{A}$ can be expressed as

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^t \tag{2.29}$$

where $\boldsymbol{U}$ and $\boldsymbol{V}$ are both orthogonal matrix and $\boldsymbol{\Sigma} = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n)$ is the diagonal matrix of the *singular values*(σ_i) of matrix $\boldsymbol{A}$. Based on eq. (2.29), we can get the inverse of $\boldsymbol{A}$ as

$$\boldsymbol{A}^{-1} = \boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^t \tag{2.30}$$

Where $\boldsymbol{\Sigma}^{-1} = \text{diag}(\frac{1}{\sigma_1}, \frac{1}{\sigma_2}, \ldots, \frac{1}{\sigma_n})$. Now if we consider there to be error within $\boldsymbol{B}$ such that

$$\hat{\boldsymbol{B}}_{\boldsymbol{i}} = \boldsymbol{B}_{\boldsymbol{i}} + \epsilon_i \tag{2.31}$$

we can then represent $\boldsymbol{Ax} = \boldsymbol{B}$ as

$$\begin{aligned} \boldsymbol{Ax} &= \hat{\boldsymbol{B}} \\ \boldsymbol{U\Sigma V}^t\boldsymbol{x} &= \hat{\boldsymbol{B}} \\ \boldsymbol{x} &= \boldsymbol{V\Sigma}^{-1}\boldsymbol{U}^t\hat{\boldsymbol{B}} \end{aligned} \tag{2.32}$$

A matrix $\boldsymbol{A}$ is considered as ill-condition when its condition number $\kappa(\boldsymbol{A})$ is large or singular when its condition number is infinite. One can represent the condition number as $\kappa(\boldsymbol{A}) = \frac{\sigma_1}{\sigma_n}$. Therefore it can be observed that when σ_n is tiny, $\boldsymbol{A}$ is likely to be ill-conditioned and when $\sigma_n = 0$, $\boldsymbol{A}$ will be singular.

One can also observe from eq. (2.32) that when the singular value σ_i is small, the error ϵ_i in eq. (2.31) will be drastically magnified by a factor of $\frac{1}{\sigma_i}$. Making the system of equation highly sensitive to errors in the input.

To obtain a meaningful solution from this ill-conditioned/singular matrix $\boldsymbol{A}$, we may perform the tSVD method to obtain a pseudo inverse of $\boldsymbol{A}$. Similar to eq. (2.29), the tSVD of $\boldsymbol{A}$ can be represented as

$$\boldsymbol{A}^+ = \boldsymbol{U\Sigma}_k\boldsymbol{V}^t \qquad \text{and} \qquad \boldsymbol{\Sigma}_k = \text{diag}(\sigma_1, \ldots, \sigma_k, 0, \ldots, 0) \tag{2.33}$$

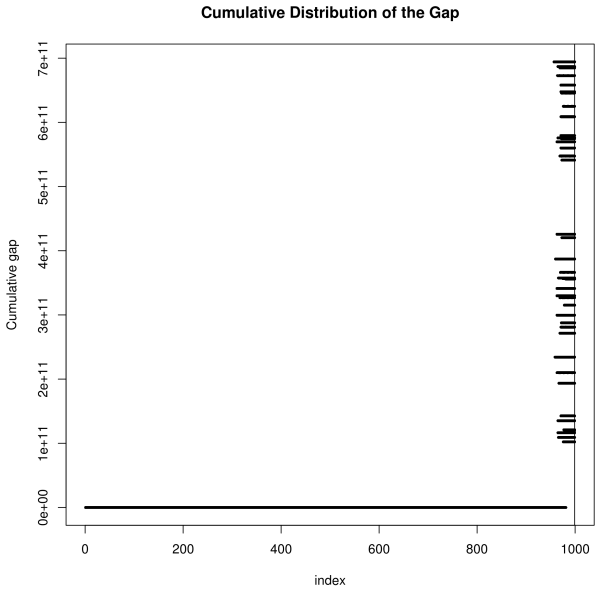
where $\boldsymbol{\Sigma}_k$ equals to replacing the smallest $n-k$ singular value replaced by 0 (Hansen, 1987). Alternatively, we can define

$$\sigma_i = \begin{cases} \sigma_i & \text{for} \quad \sigma_i \geq t \\ 0 & \text{for} \quad \sigma_i < t \end{cases} \tag{2.34}$$

where t is the tolerance threshold. Any singular value σ_i less than the threshold will be replaced by 0.

By selecting an appropriate t, tSVD can effectively regularize the ill-conditioned matrix and help to find a reasonable approximation to x. A problem with tSVD however is that it only work when matrix $\boldsymbol{A}$ has a well determined numeric rank(Hansen, 1987). That is, tSVD work best when there is a large gap between σ_k and σ_{k+1}. If a matrix has ill-conditioned rank, then $\sigma_k - \sigma_{k+1}$ will be small. For any threshold t, a small error can change whether if σ_{k+1} and subsequent singular values should be truncated, leading to unstable

Figure 2.1: Cumulative Distribution of "gap" of the LD matrix, the vertical line indicate the full rank. It can be observed that there is a huge increase in "gap" before full rank is achieved. Suggesting that the rank of the LD matrix is well defined



results.

According to Hansen (1987), matrix where its rank has meaning will have well defined rank. As LD matrix is the correlation matrix between each individual SNPs, the rank of the LD matrix is the maximum number of linear independent SNPs in the region, therefore likely to have a well-defined rank. The easiest way to test whether if the threshold t and whether if the matrix $\boldsymbol{A}$ has well-defined rank is to calculate the "gap" in the singular value:

$$gap = \sigma_k / \sigma_{k+1} \tag{2.35}$$

a large gap usually indicate a well-defined gap. In this study, we adopt the threshold as defined in MATLAB, NumPy and GNU Octave: $t = \epsilon \times \max(m, n) \times \max(\boldsymbol{\Sigma})$ where ϵ is the machine epsilon (the smallest number a machine can define as non-zero). And we perfomed a simulation study to investigate the performance of tSVD under the selected threshold. Ideally, if the "gap" is large under the selected threshold, then tSVD will provide a good regularization to the equation.

1,000 samples were randomly simulated from the HapMap(Altshuler et al., 2010) CEU population with 1,000 SNPs randomly select from chromosome 22. The LD matrix and its corresponding singular value were calculated. The whole process were repeated 50 times and the cumulative distribution of the "gap" of singular values were plotted (fig. 2.1). It is clearly show that the LD matrix has a well-defined rank with a mean of maximum "gap" of

466,198,939,298. Therefore the choice of tSVD for the regularization is appropriate.

By employing the tSVD as a method for regularization, we were able to solve the ill-posed eq. (2.16), and obtain the estimated heritability.

2.2.6 Implementation

Our algorithm was implemented using C++ programming languages and the matrix algebra was performed using the EIGEN C++ header library (Guennebaud and Jacob, 2010). Although the Armadillo library (Sanderson, 2010) is much faster in the calculation of SVD when compared to EIGEN (Ho, 2011), it is dependent on additional libraries such as OpenBLAS. The use of EIGEN therefore simplify the programme installation, making it more user friendly.

Although tSVD allow one to solve the ill-posed eq. (2.16), it is an $\mathrm{O}(n^3)$ algorithm, making the computation run time prohibitive when the number of SNPs is large. Unfortunately, the number of SNPs in a GWAS is generally large, making it impossible for one to calculate the tSVD of the whole genome at once.

If we consider eq. (2.29), the matrix $\boldsymbol{U}$ and $\boldsymbol{V}$ are the eigenvectors of $\boldsymbol{A}\boldsymbol{A}^t$ and $\boldsymbol{A}^t\boldsymbol{A}$ respectively. So for any symmetric matrix such as that of the LD matrix, $\boldsymbol{U}$ and $\boldsymbol{V}$ should be the same except for their signs. Thus eq. (2.29) reduce into the problem of eigenvalue decomposition where the singular values are the magnitude of the eigenvalues. Although the eigenvalue decomposition is still an $\mathrm{O}(n^3)$ algorithm, it has a smaller constant, therefore has a faster run time when compared to the computation of SVD.

However, even with the use of eigenvalue decomposition in place of SVD, the large matrix size is still making the computation of eq. (2.16) impossible. Given that it is unlikely inter chromosomal LD or for SNPs to be in LD if they are more than 1 megabase (mb) apart, one can safely assume SNPs more than 1mb apart are independent of each other. We therefore separate SNPs into 1mb bins where start of each bin are at least 1mb away from each other. Three bins are then combined to form one window, and we perform the decomposition on each windows using eq. (2.16) and only update the $\hat{\boldsymbol{h^2}}$ for the bin forming the center of the window. We then transverse the genome with step size of 1 bin until $\hat{\boldsymbol{h^2}}$ for all bins were computed. By breaking down the genome into windows, we were able to reduce the matrix dimension which makes the analysis plausible. Users can also choose distance other than 1mb as the distance between bins, allowing for a more flexible usage of

the algorithm.

2.2.7 Comparing with LD SCore

Conceptually, the fundamental hypothesis of LDSC and our algorithm were quite different. LDSC were based on the "global" inflation of test statistic and its relationship to the LD pattern. LDSC hypothesize that the larger the LD score, the more likely will the SNP be able to "tag" the causal SNP and the heritability can then be estimated through the regression between the LD score and the test statistic.

On the other hand, our algorithm focuses more on the per-SNP level. Our main idea was that the individual test statistic of each SNPs is a combination of its own effect and effect from SNPs in LD with it. Thus, based on this concept, our algorithm aimed to "remove" the inflation of test statistic introduced through the LD between SNPs and the heritability can be calculated by adding the test statistic of all SNPs after "removing" the inflation.

Mathematically, the calculation of LDSC and our algorithm were also very different. LDSC take the sum of all R^2 within a 1cM region as the LD score and regress it against the test statistic to obtain the slope and intercept which represent the heritability and amount of confounding factors respectively. In their model, LDSC assume that each SNPs will explain the same portion of heritability

$$\begin{aligned} \mathrm{Var}(\beta) &= \frac{h^2}{M}\boldsymbol{I} \\ M &= \text{number of SNPs} \\ \beta &= \text{vector containing per normalized genotype effect sizes} \\ I &= \text{identity matrix} \\ h^2 &= \text{heritability} \end{aligned} \tag{2.36}$$

As for our algorithm, the whole LD matrix were used and inverted to decompose the LD from the test statistic. There were no assumption of the amount of heritability explained by each SNPs. However, our algorithm does assumed that the null should be 1 and therefore cannot detect the amount of confounding factors.

2.3 Comparing Different LD correction Algorithms

Another important consideration in our algorithm is the bias in LD. In reality, one does not have the population LD matrix, instead we have to estimate he LD based on various reference panels such as those from the 1000 genome project(Project et al., 2012) or the HapMap project(Altshuler et al., 2010). These reference panels were a subsamples from the whole population and therefore LD estimated from the reference panels usually contains sampling bias. Under normal circumstances, because the symmetric nature of sampling error, one would expect there to be little to no bias in the estimated LD. However, in our algorithm , the R^2 is required for the estimation of heritability (eq. (2.17)). Because we were using the squared LD, the sampling error will also be squared, generating a positive bias.

On average, there were around 500 samples for each super population from the 1000 genome project reference panel. Given the relatively small sample size, the sampling bias might be high, therefore lead to systematic bias in the heritability estimation in our algorithm.

To correct for the bias, we would like to apply a LD correction algorithm to correct for the bias in the sample LD. Different authors (Weir and W G Hill, 1980; Wang and Thompson, 2007) have proposed methods for the correction of sample R^2 and can be applied for the correction of sample bias in LD. Therefore we considered the following R^2 correction algorithms:

$$\text{Ezekiel} : \tilde{R^2} = 1 - \frac{n-1}{n-2}(1-\hat{R^2}) \tag{2.37}$$

$$\text{Olkin-Pratt} : \tilde{R^2} = 1 - \frac{(n-3)(1-\hat{R^2})}{n-2}(1+\frac{2(1-\hat{R^2})}{n}) \tag{2.38}$$

$$\text{Pratt} : \tilde{R^2} = 1 - \frac{(n-3)(1-\hat{R^2})}{n-2}(1+\frac{2(1-\hat{R^2})}{n-3.3}) \tag{2.39}$$

$$\text{Smith} : \tilde{R^2} = 1 - \frac{n}{n-1}(1-\hat{R^2}) \tag{2.40}$$

$$\text{Weir} : \tilde{R^2} = \hat{R^2} - \frac{1}{2n} \tag{2.41}$$

where n is the number of samples used to calculate the R^2, $\hat{R^2}$ is the sample R^2 and $\tilde{R^2}$ is the corrected $R2$.

In order to assess the performance of each individual correction methods, we perform simulations to compare the performance of our algorithm using different LD bias correction algorithms. Most importantly, we would like to assess the performance of different

algorithms not only under one specific LD range, but also under the complex LD structure observed in real life scenarios. First, 5,000 SNPs with Minor Allele Frequency (maf) ≥ 0.1 were randomly selected from chromosome 22 from the 1000 genome Northern Europeans from Utah (CEU) haplotypes and were used as an input to HAPGEN2 (Su, Marchini, and Donnelly, 2011) to simulate 1,000 individuals. HAPGEN2 is a simulation tools which simulates new haplotypes as an imperfect mosaic of haplotpyes from a reference panel and the haplotypes that have already been simulated using the *Li and Stephens* (LS) model of LD (N. Li and Stephens, 2003). This allow us to simulate genotypes with LD structures comparable to those observed in CEU population. Of those 5,000 SNPs, 100 of them were randomly selected as the causal variant. Orr (1998) suggested that the exponential distribution can be used to approximate the genetic architecture of adaptation. As a result of that, we used the exponential distribution with $\lambda = 1$ as an approximation to the effect size distribution:

$$\begin{aligned} \theta &= \exp(\lambda = 1) \\ \beta &= \pm\sqrt{\frac{\theta \times h^2}{\sum \theta}} \end{aligned} \tag{2.42}$$

with a random direction of effect. The simulated effects were then randomly distributed to each causal SNPs.

Using the normalized genotype of the causal SNPs of each individual ($\boldsymbol{X}$), the vector of effect size ($\boldsymbol{\beta}$) we can simulate a phenotype with target heritability of h^2 as

$$\begin{aligned} \epsilon_i &\sim N(0, \sqrt{\mathrm{Var}(\boldsymbol{X}\boldsymbol{\beta})\frac{1-h^2}{h^2}}) \\ \boldsymbol{\epsilon} &= (\epsilon_1, \epsilon_2, ..., \epsilon_n)^t \\ \boldsymbol{y} &= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \end{aligned} \tag{2.43}$$

To simulate the whole spectrum of heritability, we varies the target h^2 from 0 to 0.9 with increment of 0.1.

The test statistics of association between the genotype and phenotype were then calculated using PLINK (Shaun Purcell et al., 2007). Resulting test statistic were then input to our algorithm to estimate the heritability, using different LD correction algorithms. An independent 500 samples, a size roughly correpsond to the average sample size of each super population form the 1,000 genome project, were simulated as a reference panel for the calculation of LD matrix. This is because in reality, one usually doesn't have assess to the sample genotype and has to rely on an independent reference panel for the calculation of LD

matrix. Thus this simulation procedure should provide a realistic representation of how the algorithm was commonly used in real life scenario.

The whole process will be repeated 50 times such that a distribution of the estimate can be obtained. In summary, we simulate a large population of samples (e.g. $50 \times 1,000 + 500 = 50,500$) where 500 samples were randomly selected as a reference panel. In the subsequent iteration of simulation, 1,000 samples were randomly selected from the population *without replacement* and estimation were performed.

1. Randomly select 5,000 SNPs with maf$>$ 0.1 from chromosome 22
2. Simulate 500 samples using HAPGEN2 and used as a reference panel
3. Randomly generate 100 effect size with following eq. (2.42)
4. Randomly assign the effect size to 100 SNPs with heritability from 0 to 0.9 (increment of 0.1)
5. Simulate 1,000 samples using HAPGEN2 and calculate their phenotype according to eq. (2.43)
6. Perform heritability estimation using our algorithm with different ways of LD correction
7. Repeat step 5-6 50 times

2.4 Comparison with Other Algorithms

After identifying the optimal LD correction algorithm, we would like to compare our algorithm to existing methods for the performance in estimating the narrow sense heritability. It is important for us to consider most if not all conditions in our simulation. Therefore, we would like to simulate quantitative traits and case control studies with different number of causal SNPs; quantitative traits with extreme effect sizes; and last but not least, quantitative traits with extreme phenotype selection.

Currently, the only other algorithm that is capable to estimate the narrow sense heritability using only test statistic is the LDSC (B. K. Bulik-Sullivan et al., 2015). On the other hand, GCTA (J Yang et al., 2011) is commonly used for heritability estimation in GWAS data. Therefore, we choose to compare the performance of our algorithm to that of LDSC and GCTA. It is important to note that as we are assessing the performance of

the algorithms through controlled simulation, there should be little confounding factors. For LDSC, the default intercept estimation function allows it to estimate and correct for confounding factors with an increase in SE. The simulation will therefore be unfair to LDSC with intercept estimation, as the SE is increased yet there are little confounding factors for it to correct. Thus, we also simulate LDSC with a fixed intercept (--no-intercept) parameters to avoid bias against LDSC.

2.4.1 Sample Size

One important consideration in our simulation was the number of sample simulated. The sample size was the most important parameter in determining the standard error of the heritability estimation. As sample size increases, study will be more representative of the true population. The increased number of information also means a better estimation of parameters, therefore a smaller standard error (SE). Based on information from GWAS catalog(Welter et al., 2014), we calculate the sample size distribution using simple text mining and exclude studies with conflicting sample size information in multiple entries. The average sample size for all GWAS recorded on the GWAS catalog was 7,874, with a median count of 2,506 and a lower quartile at 940 (fig. 2.2). We argue that if the algorithm works for studies with a small sample size (e.g lower quartile sample size), then it should perform even better when the sample size is larger. Thus, we only simulate 1,000 samples in our simulation, which roughly represent the lower quartile sample size range.

2.4.2 Number of SNPs in Simulation

Another consideration in the simulation was the number of SNPs included. In a typical GWAS study, there are usually a larger number of SNPs when compared to the sample size. For example, in the Psychiatric Genomics Consortium (PGC) schizophrenia GWAS, more than 9 million SNPs were included, with around 700,000 SNPs on chromosome 1. In reality, the estimation of heritability based on 700,000 SNPs can be done quickly. However, in our simulation, we will repeat the calculation 50(iteration) $\times$ 10(number of heritability) = 500 times for *each* condition tested. The time required to finish all the simulation quickly becomes infeasible given the large amount of SNPs. To compromise, we simulate a total of 50,000 SNPs from chromosome 1 as a balance between run time of simulation and the total SNPs simulated. With 50,000 SNPs, there are roughly 200 SNPs within a 1 mb region.

2.4.3 Genetic Architecture

Of all simulation parameter, the genetic architecture was the most complicated and important parameter. The LD pattern, the number of causal SNPs, the effect size of the causal SNPs and the heritability of the trait were all important factors contribute to the genetic architecture of a trait.

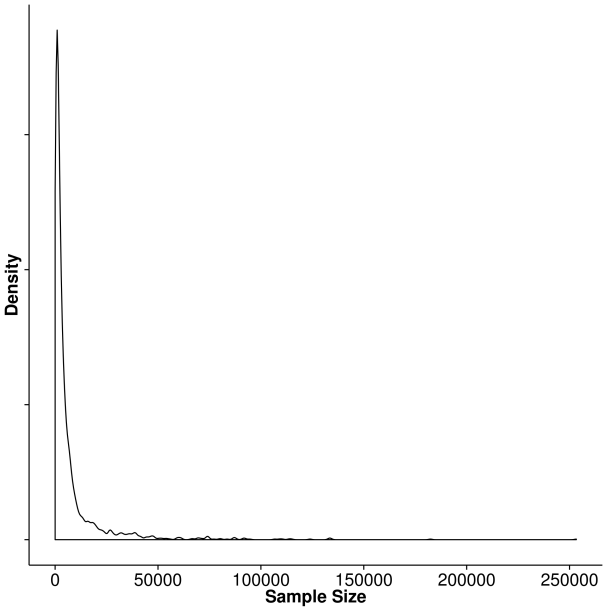


Figure 2.2: GWAS sample size distribution.

First and foremost, because the aim of the algorithm was to estimating the heritability of the trait, it is important that the algorithm works for traits from different heritability spectrum. We therefore simulate traits with heritability ranging from 0 to 0.9, with increment of 0.1.

Secondly, in real life scenario, the "causal" variant might not be readily included on the GWAS chip and were only "tagged" by SNPs included on the GWAS chip. However, to simplify our simulation, all "causal" variants were included in our simulation (e.g. perfectly "tagged")

Thirdly, to obtain a realistic LD pattern, we simulate the genotypes using the HAPGEN2 programme(Su, Marchini, and Donnelly, 2011), using the 1000 genome CEU haplotypes as an input. In a typical GWAS , one usually only have power in detecting "common variants", defined as variants with maf ≥ 0.05. We therefore only consider scenario with "common" variants and only use SNPs with maf ≥ 0.05 in the CEU haplotypes as an input to HAPGEN2 to simulate 1,000 samples.

Finally, we would like to simulate traits with different inheritance model such as oligogenic traits and polygenic traits. We therefore varies the number of causal SNPs (k) with $k \in \{5, 10, 50, 100, 250, 500\}$. The effect size were then simulated using eq. (2.42) and the phenotype were simulated using eq. (2.43).

For GCTA, the sample genotypes were provided to calculate the genetic relationship matrix and the sample phenotype were used in combination with the genetic relationship matrix to estimate the heritability.

On the other hand, for LDSC and our algorithm, an independent 500 samples were simulated as the reference panel for the calculation of LD scores and LDmatrix, mimicking real life scenario where an independent reference panel were used. The genotype association test statistics calculated from PLINK and the LD score / LD matrix were then used for the estimation of heritability for LDSC and our algorithm respectively.

The whole process will be repeated 50 times such that a distribution of the estimate can be obtained. 10 independent population were simulated and the whole processed were repeated. In summary, the simulation follows the following procedures:

1. Randomly select 50,000 SNPs with maf$>$ 0.1 from chromosome 1
2. Simulate 500 samples using HAPGEN2 and used as a reference panel
3. Randomly generate k effect size with $k \in \{5, 10, 50, 100, 250, 500\}$ following eq. (2.42), with heritability ranging from 0 to 0.9 (increment of 0.1)
4. Randomly assign the effect size to k SNPs
5. Simulate 1,000 samples using HAPGEN2 and calculate their phenotype according to eq. (2.43)
6. Perform heritability estimation using our algorithm, GCTA, LDSC with fixed intercept and LDSC with intercept estimation.
7. Repeat step 5-6 50 times
8. Repeat step 1-7 10 times

2.4.4 Extreme Effect Size

On top of the original quantitative trait simulation, another condition we were interested in was the performance of the algorithms when there is a small amount of SNPs with a much larger effect size. This can be observed in disease such as Hirschsprung's disease. The Hirschsprung's disease is a congenital disorder where deleterious mutations on *RET* account for ≈50% of the familial cases yet there is still missing heritability, suggesting that there might be more variants with small effects that have not been identified (Gui et al., 2013).

To simulate extreme effect size, we consider scenarios where m SNPs accounts 50% of all the effect size with $m \in \{1, 5, 10\}$. The effect size was then calculated as

$$\begin{aligned} \beta_{eL} &= \pm\sqrt{\frac{0.5h^2}{m}} \\ \beta_{eS} &= \pm\sqrt{\frac{0.5h^2}{100-m}} \\ \beta &= \{\beta_{eL}, \beta_{eS}\} \end{aligned} \tag{2.44}$$

The effect size were then randomly assigned to 100 causal SNPs and phenotype will be calculated as in eq. (2.43). The simulation procedure then becomes

1. Randomly select 50,000 SNPs with maf> 0.1 from chromosome 1
2. Simulate 500 samples using HAPGEN2 and used as a reference panel
3. Randomly generate 100 effect size where m has extreme effect, following eq. (2.44), with $m \in \{1, 5, 10\}$
4. Randomly assign the effect size to 100 SNPs
5. Simulate 1,000 samples using HAPGEN2 and calculate their phenotype according to eq. (2.43)
6. Perform heritability estimation using our algorithm, LDSC with fixed intercept, LDSC with intercept estimation and GCTA
7. Repeat step 5-6 50 times
8. Repeat step 1-7 10 times

2.4.5 Case Control Studies

The simulation of case control studies was similar to the simulation of quantitative trait. However, there were two additional parameters to consider: the population prevalence and the observed prevalence. These parameters were required to simulate the samples under a liability model for case control studies.

Although there were only two additional parameter, it is significantly more challenging for to simulate when compared to the simulation of quantitative traits. It is mainly

because of the number of samples required to simulate adequate samples under the liability threshold model. Take for example, if one like to simulate a trait with population prevalence of p and observed prevalence of q and would like to have n cases in total, one will have to simulate $\min(\frac{n}{p}, \frac{n}{q})$ samples. Considering the scenario where the observed prevalence is 50%, the population prevalence is 1%, if we want to simulate 1,000 cases, a minimum of 100,000 samples will be required.

Given limited computer resources, it will be infeasible for us to simulate 1,000 cases with 50,000 SNPs when the population prevalence is small. To simplify the simulation and reduce the burden of computation, we limited the observed prevalence to 50% and varies the population prevalence p such that $p \in \{0.5, 0.1, 0.05, 0.01\}$. Most importantly, we reduce the number of SNPs simulated to 5,000 on chromosome 22 instead of 50,000 SNPs on chromosome 1. The change from chromosome 1 to chromosome 22 allow us to reduce the number of SNPs without changing much of the SNP density. We acknowledged that the current simulation was relatively brief, however, it should serves as a prove of concept simulation to study the performance of the algorithms under the case control scenario.

In the case control simulation, we randomly select 5,000 SNPs from chromosome 22 with maf ≥ 0.1 in the CEU haplotypes as an input to HAPGEN2. We then randomly select k SNPs where $k \in \{10, 50, 100, 500\}$, each with effect size simulated based on eq. (2.42). In order to simulate a case control samples with 1,000 cases, we then simulate $\frac{1{,}000}{p}$ samples and calculate their phenotype using eq. (2.43). The phenotype was then standardized and cases were defined as sample with phenotype passing the liability threshold with respect to p. An equal amount of samples were then randomly selected from samples with phenotype lower than the liability threshold and defined as controls.

Finally, the case control simulation were performed as:

1. Randomly select 5,000 SNPs with maf> 0.1 from chromosome 22
2. Simulate 500 samples using HAPGEN2 and used as a reference panel
3. Randomly generate k effect size following eq. (2.42) where $k \in \{10, 50, 100, 500\}$
4. Randomly assign the effect size to k SNPs
5. Simulate $\frac{1{,}000}{p}$ samples using HAPGEN2 and calculate their phenotype according to eq. (2.43)
6. Define case control status using the liability threshold and randomly select same number of case and controls for subsequent simulation

7. Perform heritability estimation using our algorithm, LDSC with fixed intercept, LDSC with intercept estimation and GCTA

8. Repeat step 5-7 50 times

9. Repeat step 1-8 10 times

2.4.6 Extreme Phenotype Selection

The simulation of extreme phenotype selection was the same as the quantitative trait simulation. The only difference being that instead of using all samples for heritability estimation, we only use the extreme 10% of samples among the population for the heritability estimation. In brief, instead of simulating 1,000 samples, we simulate 5,000 samples following the exact procedure in the quantitative trait simulation with random effect size. However, after simulation of the phenotype using eq. (2.43), we standardize the phenotype and only select the top 10% and bottom 10% samples (500 samples each) from the sample distribution. We then perform the same simulation procedure as in the quantitative trait simulation with random effect size.

It was noted that the extreme phenotype selection were not supported by the LDSC and GCTA. To allow comparison in such scenario, we apply the extreme phenotype adjustment from Pak C Sham and S. M. Purcell (2014) to the estimation obtained from LDSC and GCTA.

2.5 Simulation with Real Data

2.6 Application to Real Data

To test the performance of our algorithm under real life scenario, we apply our algorithm to the PGC data, including Bipolar (Psychiatric GWAS Consortium Bipolar Disorder Working Group, 2011), Major depression disorder (Stephan Ripke, Naomi R Wray, et al., 2013), Autism (Unpublished) and schizophrenia (S Ripke et al., 2013). We also performed LDSC alongside our algorithm to compare the results from the two algorithm. Unfortunately, as the sample genotypes were not provided, we cannot perform GCTA analysis, therefore we only considered our algorithm and LDSC. For the bipolar and major depression data, we

performed liftover (Hinrichs et al., 2006) to convert the genomic coordinates to genome version hg19 such that it is compatible with the data from 1000 genome.

The reference genome were downloaded from 1000 genome (Project et al., 2012) and were converted to plink binaries using plink --vcf function. We used the European super population which contain a total of 503 samples where singleton and non-biallelic SNPs were filtered out. To filter related samples, genotypes were first pruned before the identity by descent (IBD) were calculated. Samples pairs with pi hat larger than 0.125 were considered related, which roughly correspond to third degree relativeness. Samples were removed on a stepwise fashion where samples related to most samples were removed first, until none of the samples were related. In total, 57 samples were removed, leaving us with 446 reference samples. LD score was computed based on the 446 samples using a 1mb window size.

As all the studies were case control GWAS, the population prevalence of the trait has to be provided in order to adjust for the attenuation bias. Therefore we used prevalence of 0.15 for major depression disorder and 0.01 for schizophrenia, bipolar disorder and autism.

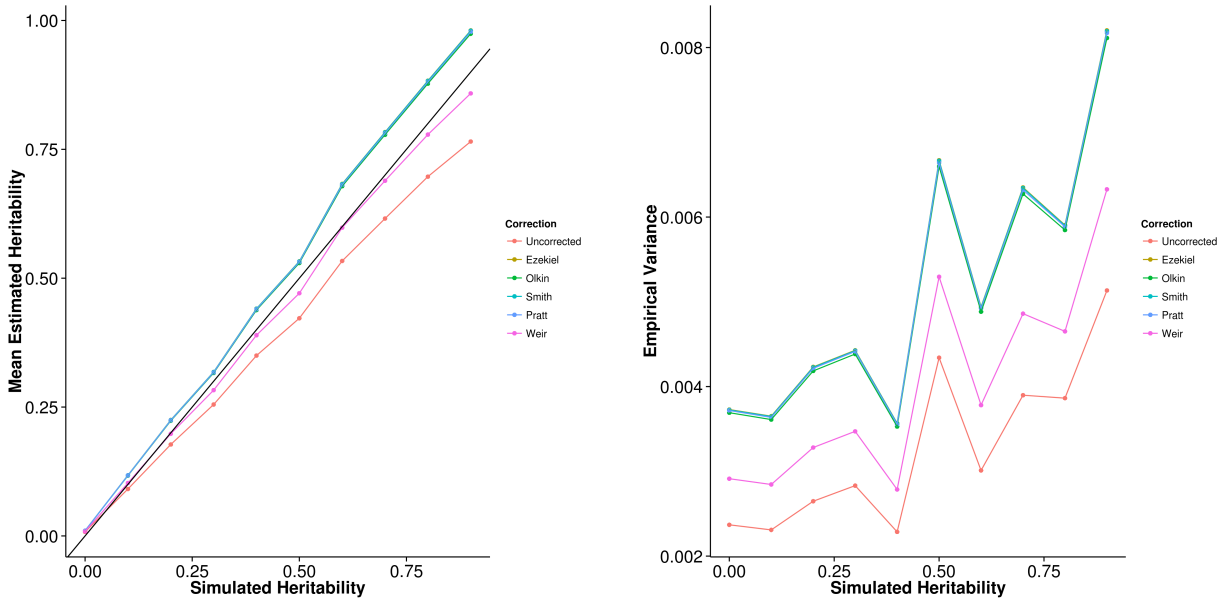
2.7 Result

The heritabilibty estimation were implemented in SNP Heritability and Risk Estimation Kit (SHREK) and is available on `https://github.com/choishingwan/shrek`.

2.7.1 LD Correction

First, we would like to assess the effect of LD correction on the heritability estimation and the impact of different bias correction algorithms. By performing the simulation using HAPGEN2, we were able to simulate sample with LD structure comparable to the LD of the 1000 genome CEU samples.

First, we would like to compare the performance of SHREK when different bias correction algorithms were applied (fig. 2.3a). From the graph, it was observed that when no bias correction was applied, the mean estimation were in general downwardly biased. This was consistent with our expectation of a general upward bias in sample R^2 which will downwardly penalize the resulting heritability estimation. On the other hand, the bias correction algorithms all worked as expected where they increases the mean estimation of heritability. By removing the upward bias in the sample R^2, the heritability estimation



(a) Mean Estimation

(b) Empirical Variance

Figure 2.3: Effect of LD correction to Heritability Estimation. We compared the performance of our algorithm when different R^2 bias correction algorithm was used. When no bias correction was carried out, a downward bias was observed. After the application of the bias correction algorithms, the mean estimations of all except in the case of Weir eq. (2.41) algorithms leads to an overestimation of heritability. On the other hand, the corrections all lead to increase in variance of the estimation.

should increase. However for most algorithms except for Weir's formula (eq. (2.41)) an over adjustment were observed, leading to a general upward bias in the estimation. Taking into account of the variance of estimation (fig. 2.3b), Weir's formula were the most suitable for SHREK where not only it reduces the bias in the final heritability estimation, it does not introduce too much additional variance into the estimation. As a result of that, we selected the Weir's formula as our default LD correction algorithm.

2.7.2 Comparing with Other Algorithms

Having selected the optimal LD correction algorithm, we then compared the performance of SHREK with existing algorithms to understand the relative of these algorithms under different conditions. First, we examined the performance of the algorithms under the quantitative trait scenario where we varies the trait heritability and the number of causal SNPs.

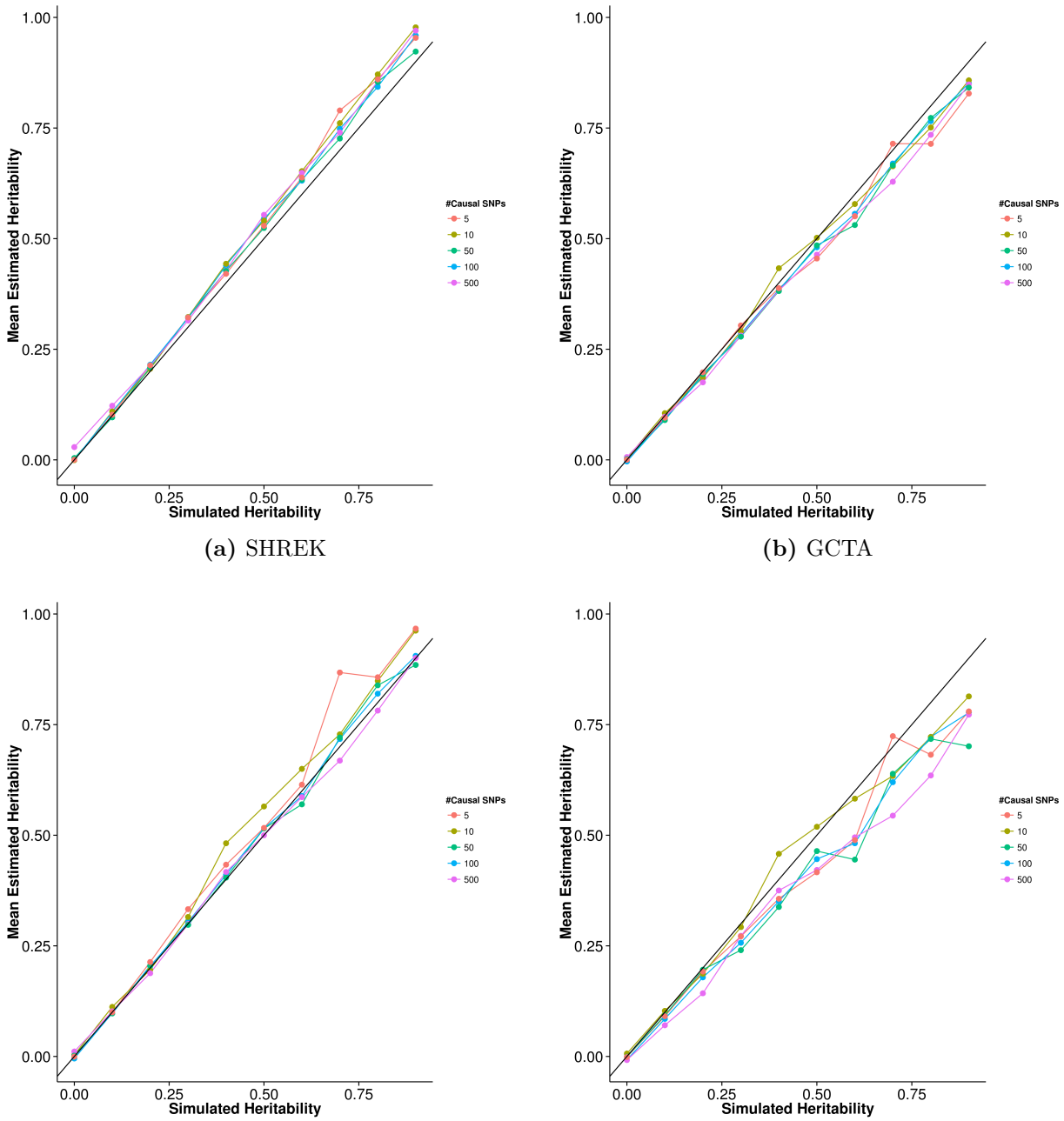
Quantitative Trait Simulation

In the simulation of quantitative trait scenario, the effect size were randomly drawn from the exponential distribution with $\lambda = 1$ and traits with different number of causal SNPs and different narrow sense heritability were simulated. The main aim of this simulation was to assess the effect of number of causal SNPs and trait heritability on the power of estimation of different algorithms.

First, the mean heritability estimation were compared to the simulated heritability in order to identify the bias in estimation for each algorithms. From the graph (fig. 2.4), it was observed that the mean estimations of SHREK has a small upward bias (fig. 2.4a). However, the bias was insensitive to the change in number of causal SNPs suggesting that SHREK is relatively robust to trait complexity. On the other hand, estimations form GCTA were moderately biased downward (fig. 2.4b), similar to the estimations from LDSC with intercept estimation (fig. 2.4d), but with a smaller variability. Finally, when the intercept is fixed, LDSC has the smallest bias when the trait is polygenic but an upward bias is also observed when the number of causal SNPs is small.

Furthermore, while comparing the empirical variance of the estimates (fig. 2.5), variance of estimations from LDSC were sensitive to the number of causal SNPs where as the number of causal SNPs decreases (figs. 2.5c and 2.5d), the variance increases, similar to what was reported by B. K. Bulik-Sullivan et al. (2015). The variance were also higher when intercept estimation was performed. On the other hand, although the variance of SHREK was relatively higher when compared to LDSC when the intercept was fixed, the variation of its estimations was insensitive to the number of causal SNPs, when the number of causal SNPs was small, the variance of estimation from SHREK can be even be lower than LDSC (fig. 2.5a). Finally, of all the algorithms, the estimations from GCTA has the lowest variation when compared to other algorithm (fig. 2.5b), except when it was the case of 5 causal SNPs where it has a slightly higher variance when compared to SHREK when the simulated heritability was high (e.g. ≥ 0.8).

Another important factor to consider was the estimation of the SE. Of all the algorithms, GCTA (fig. 2.6b) has the best estimate, follow by SHREK (fig. 2.6a). However, it was noted that a consistent underestimation of variance was observed with SHREK whereas GCTA only underestimate the variance when the number of causal SNPs is small. On the other hand, when the intercept was fixed (fig. 2.6c), LDSC cannot accurately estimate its variance and tends to overestimate, especially when the number of causal SNPs were small. When intercept estimations was performed (fig. 2.6d), the estimation of variance was



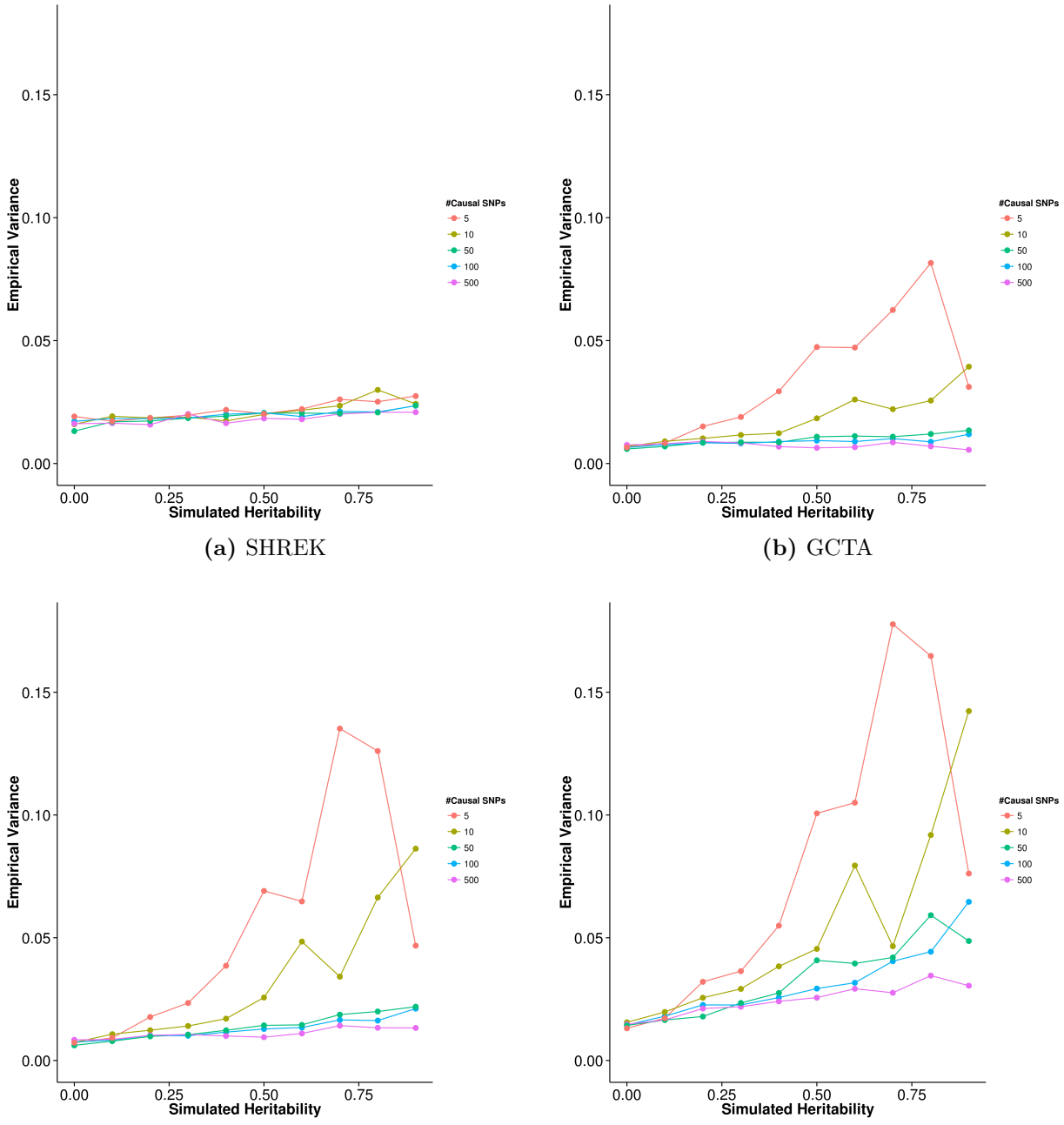
(a) SHREK

(b) GCTA

(c) LDSC with fix intercept

(d) LDSC with intercept estimation

Figure 2.4: Mean of results from quantitative trait simulation with random effect size simulation. Estimations form SHREK were slightly biased upwards whereas GCTA and LDSC with intercept estimations both biased downwards. On the other hand, LDSC with fixed intercept provides least biased estimates under polygenic conditions. However, when the number of causal SNPs is small (e.g. 5 or 10), an upward bias was observed.



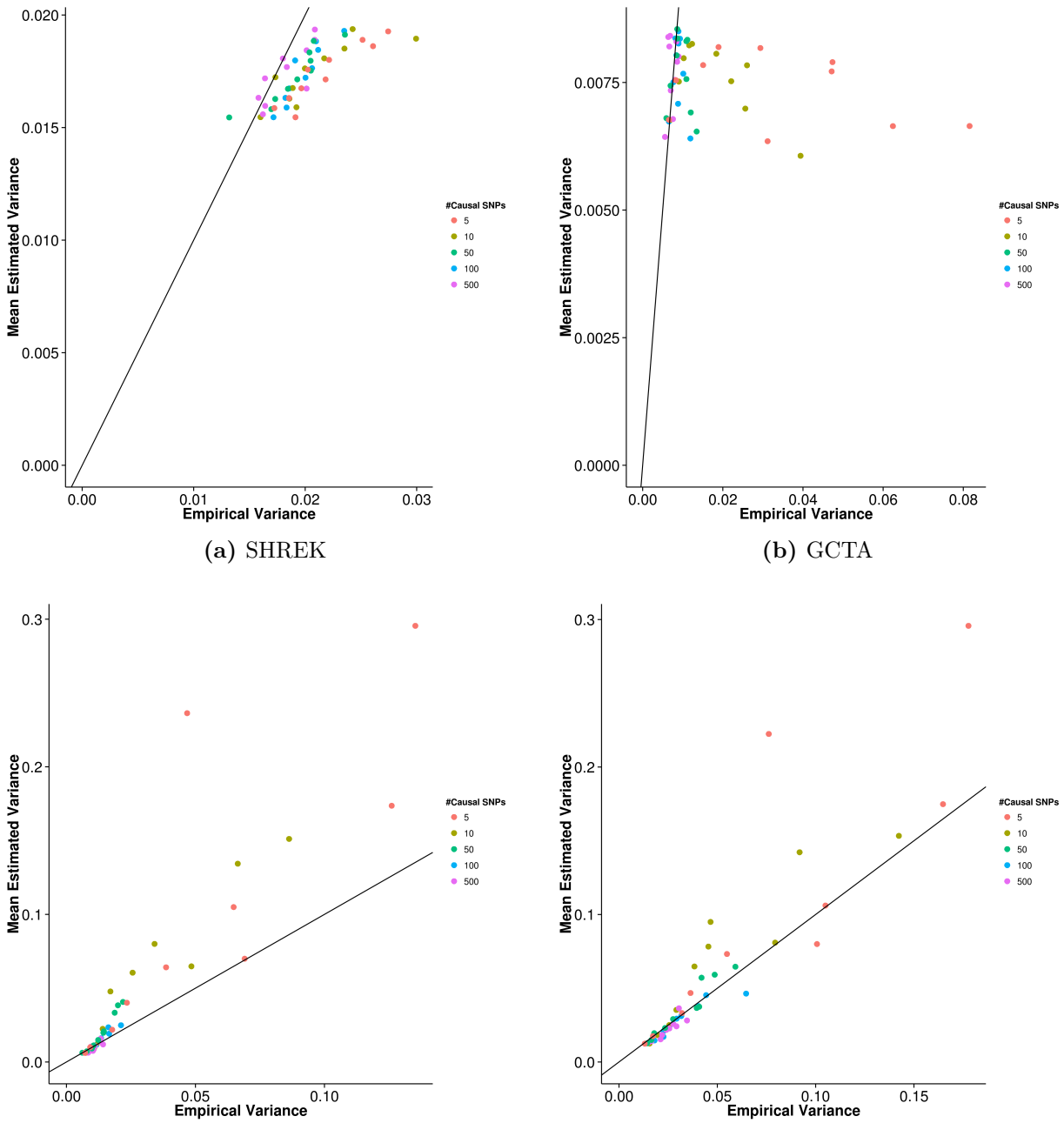
(a) SHREK

(b) GCTA

(c) LDSC with fix intercept

(d) LDSC with intercept estimation

Figure 2.5: Variance of results from quantitative trait simulation with random effect size simulation. Under the polygenic conditions, GCTA has the smallest variance, follow by LDSC. However, it was observed when the number of causal SNPs decreases, the variance of the estimation increases for all algorithm, with variance of the SHREK estimate being the least affected. In fact, under oligogenic conditions, SHREK has a lower empirical variance when compared to LDSC.



(a) SHREK

(b) GCTA

(c) LDSC with fix intercept

(d) LDSC with intercept estimation

Figure 2.6: Estimated variance of results from quantitative trait simulation with random effect size simulation when compared to the empirical variance. GCTA has the best estimate of its empirical variance under the polygenic conditions whereas SHREK tends to under-estimate its empirical variance. On the other hand, LDSC to over-estimate the variance especially when the number of causal SNPs is small.

Number of Causal SNPs	SHREK	LDSC	LDSC-In	GCTA
5	0.0235	0.0576	0.0828	0.0365
10	0.0231	0.0343	0.0555	0.0189
50	0.0196	0.0157	0.0494	0.0114
100	0.0210	0.0129	0.0363	0.00961
500	0.0205	0.0115	0.0308	0.00887

Table 2.1: Mean squared error (MSE) of quantitative trait simulation with random effect size. Of all the algorithms, GCTA has the lowest MSE except when there is only 5 causal SNPs. When comparing the performance of SHREK and LDSC with fixed intercept, the performance of SHREK is better under the oligogenic condition whereas LDSC with fixed intercept excels under the polygenic condition. On the other hand, when intercept estimation were performed, the MSE of LDSC increases, mainly due to the increased SE. Therefore SHREK out perform LDSC with intercept estimation when there are minimal confounding variables.

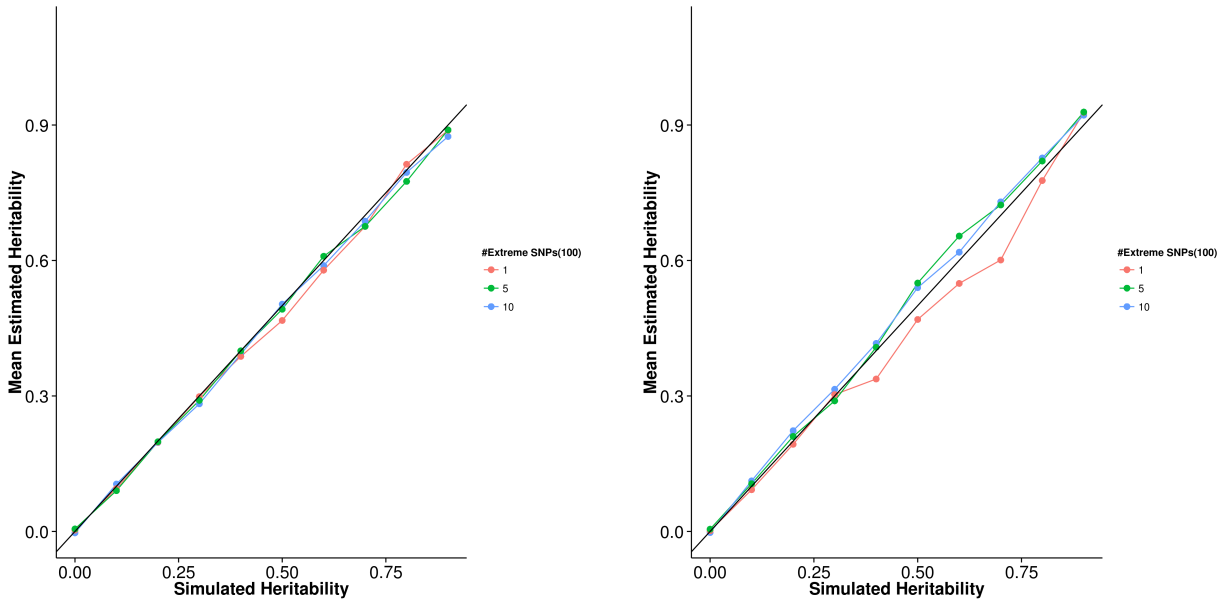
relatively better yet the overestimation were still observed when the number of causal SNPs is small.

By taking into consideration of both the bias and variance of the estimates, GCTA has the best overall performance. Under the oligogenic condition (e.g. number of causal SNPs ≤ 10), SHREK has relatively better performance when compared to LDSC. Whereas under the polygogenic condition, LDSC has better performance.

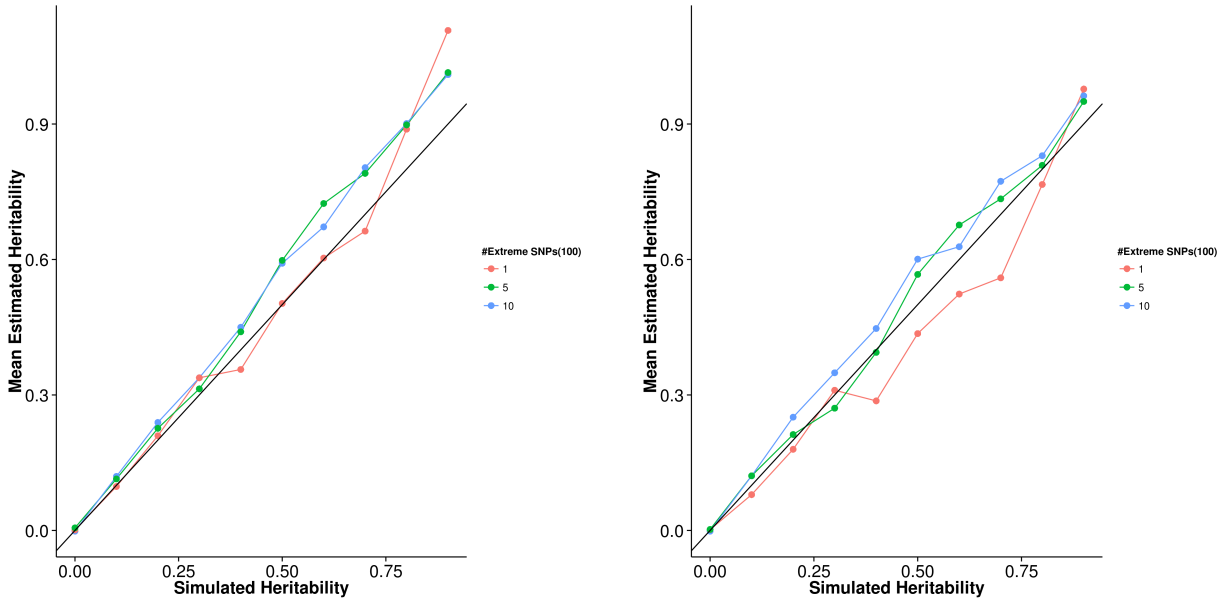
Quantitative Trait Simulation with Extreme Effect Size

Similarly, we were interested in the performance of the algorithms when a small number of SNPs account for majority of the effect. In this simulation, we simulated 100 causal SNPs of which 1, 5 or 10 of those SNPs account for majority of the effects.

When assessing the mean estimation of heritability (fig. 2.7), the performance of the algorithms were similar to that in the quantitative trait simulation. The only exception was when 1 SNP account for majority of effects during which the bias of estimation fluctuates in most algorithms except SHREK (fig. 2.7a). Similarly, the variance of the estimation (fig. 2.8) from GCTA and LDSC increases when only 1 SNP account for majority of effect. It was most obvious in the case of LDSC where the variance increased drastically as the heritability is high (fig. 2.8c). However, SHREK does not seems to be affected and were robust to the number of SNPs with extreme effect.

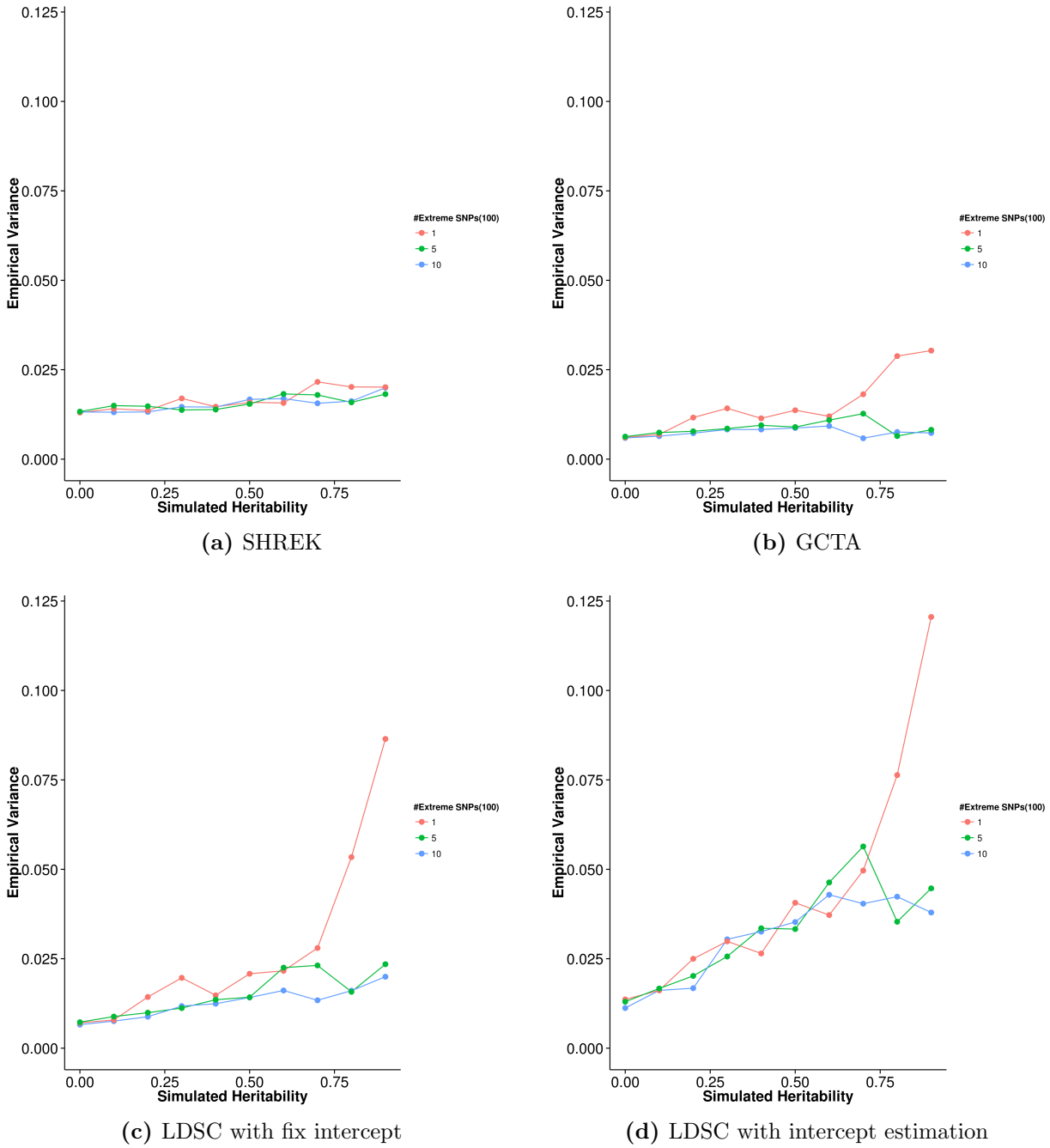


(a) SHREK **(b)** GCTA



(c) LDSC with fix intercept **(d)** LDSC with intercept estimation

Figure 2.7: Mean of results from quantitative trait simulation with extreme effect size simulation. 100 causal SNPs were simulated. It was observed that the mean estimation of heritability of all the tools were relatively unaffected by the number of SNPs representing a large portion of effect where SHREK has the least amount of bias.



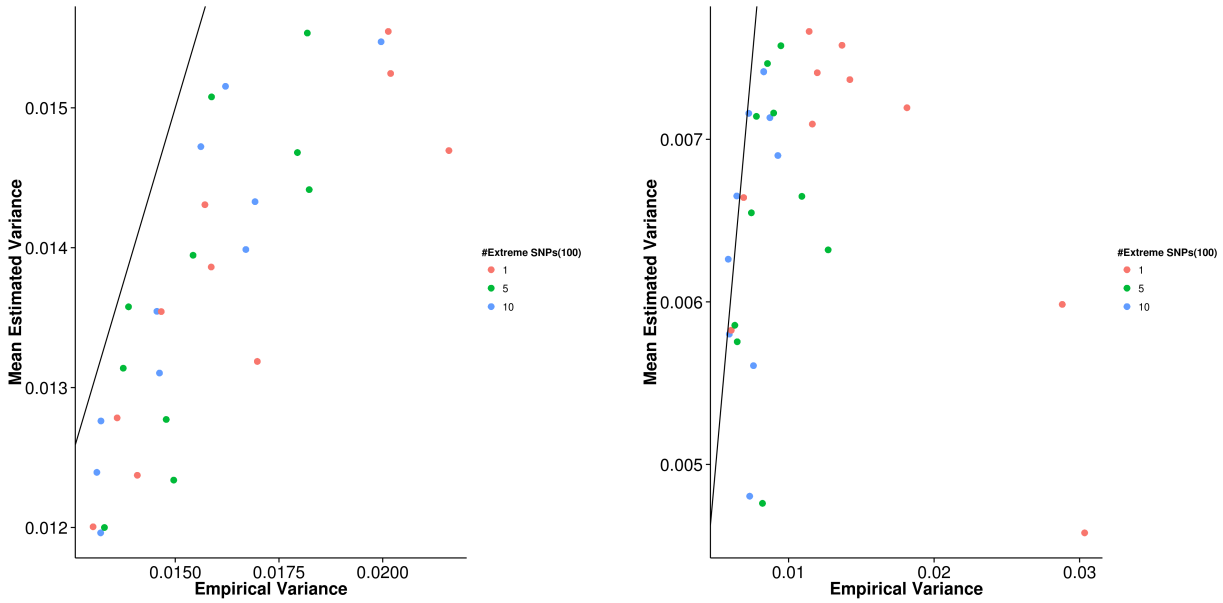
(a) SHREK

(b) GCTA

(c) LDSC with fix intercept

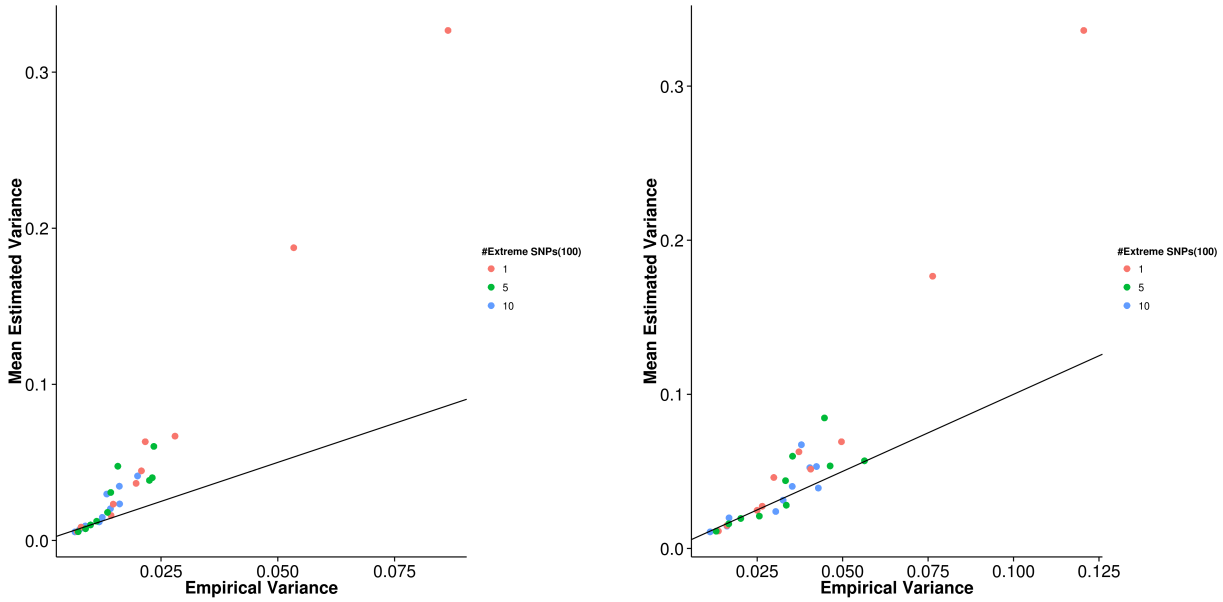
(d) LDSC with intercept estimation

Figure 2.8: Variance of results from quantitative trait simulation with extreme effect size simulation. 100 causal SNPs were simulated. GCTA has the smallest variance as with previous simulation. When compared to LDSC with fixed intercept, although the variance of SHREK was relatively higher, it was less sensitive to change in heritability and the number of SNPs explaining a large portion of effect. In situation where 1 SNP represent 50% of the effect, the variance of SHREK is actually lower than that of LDSC with fixed intercept once the heritability was ≥ 0.2.



(a) SHREK

(b) GCTA



(c) LDSC with fix intercept

(d) LDSC with intercept estimation

Figure 2.9: Estimated variance of results from quantitative trait simulation with extreme effect size simulation when compared to the empirical variance. 100 causal SNPs were simulated. SHREK generally under-estimate the variance whereas LDSC over-estimate the variance.

Number of Causal SNPs	SHREK	LDSC	LDSC-In	GCTA
1	0.168	0.329	0.485	0.171
5	0.158	0.208	0.340	0.0942
10	0.155	0.179	0.334	0.0800

Table 2.2: MSE of quantitative trait simulation with extreme effect size. Of all the algorithms, GCTA has the lowest MSE except when there is only 1 SNP with extreme effect. The performance of SHREK is in general better than LDSC and the performance of SHREK and LDSC with fixed intercept converges as the number of SNPs with extreme effect increases.

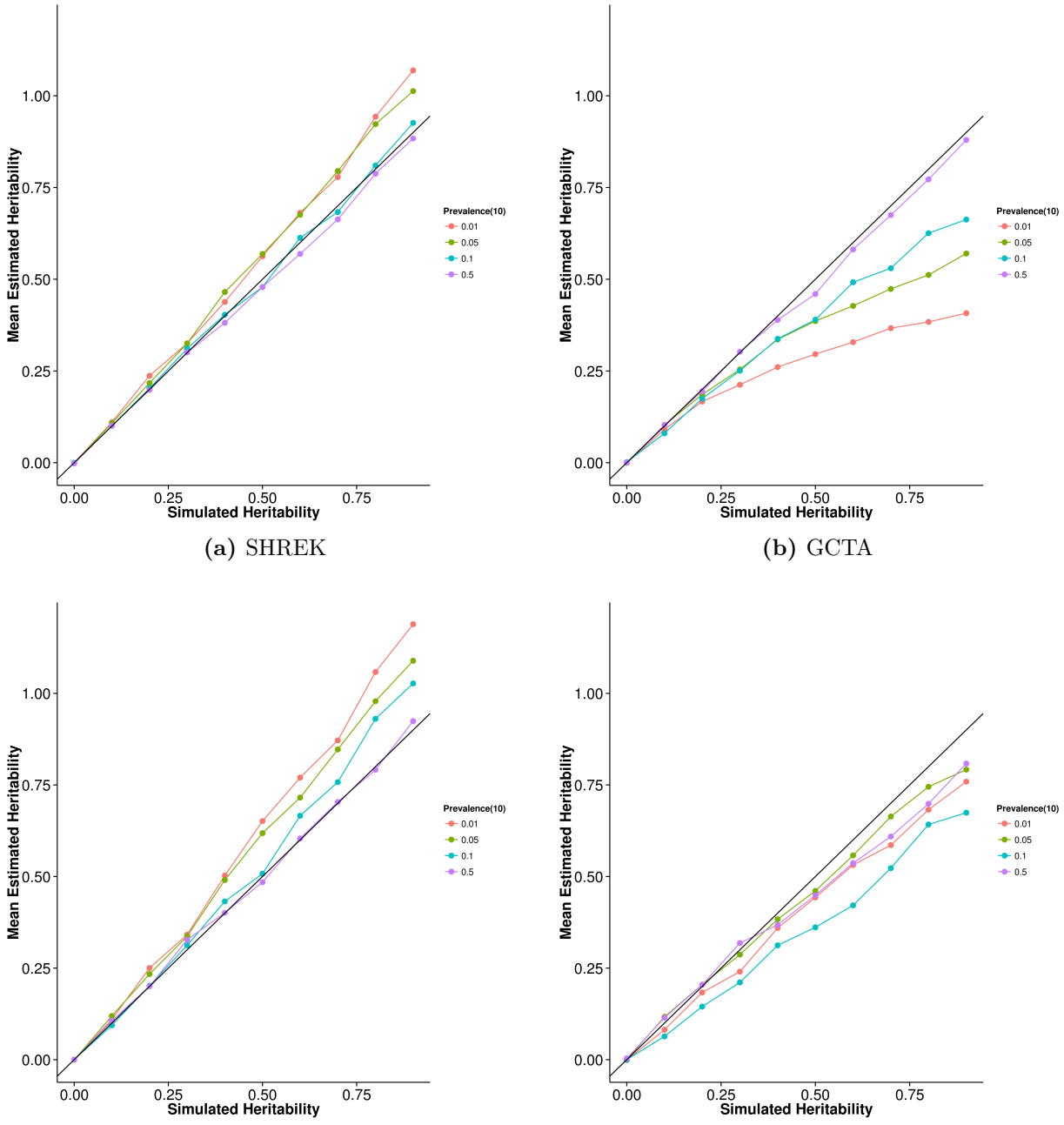
The estimated variance of LDSC were also affected by the number of SNPs with extreme effect where a smaller number of extreme effect SNPs the higher the estimated variance. A similar bias was also observed in SHREK and GCTA where the estimated variance differ more form the empirical variance when the number of SNPs with extreme effect is smaller.

To conclude, the performance of GCTA is superior to other algorithm except when there is 1 SNP with extreme effect where SHREK performs better (table 2.2). Again, the performance of SHREK was insensitive to the number of SNPs with extreme effect. Performance of LDSC gets better as the number of SNPs with extreme effect increases and performance with fixed intercept is better than when the intercept estimation function was used.

Case Control Simulation

Nowadays, most of the GWAS are Case Control studies, thus it is important to test the performance of the algorithms when dealing with case control samples. In the case control simulation, we varies the population prevalence and the trait heritability. We also varies the number of causal SNPs to assess the combine effect of these parameters to the performance of the algorithms.

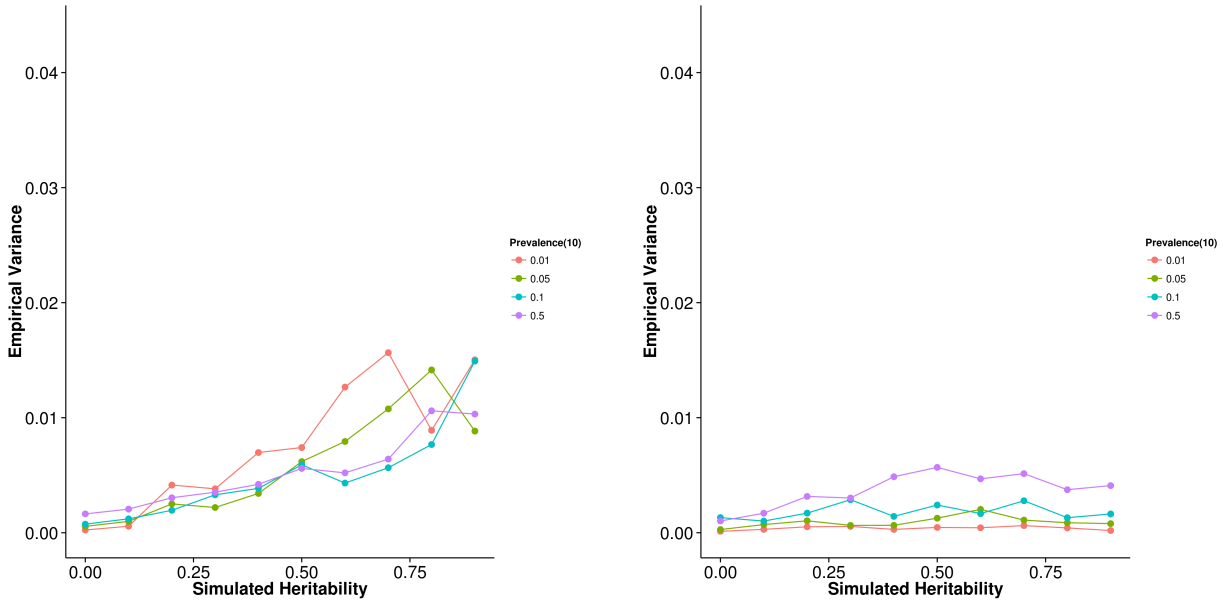
First, we simulated traits with 10 causal SNPs. From the graph, it is clear that the population prevalence has a significant impact to the performance of the algorithms (fig. 2.10). The performance of GCTA was as suggested by Golan, Eric S Lander, and Rosset (2014) where the degree of underestimation increases as the prevalence decreases. On the other hand, the opposite effect was observed for SHREK and LDSC with fixed intercept. Interestingly, when allow the estimate the intercept, the heritability estimated from LDSC becomes underestimated. The magnitude of the bias also decreases, suggesting



(a) SHREK **(b)** GCTA

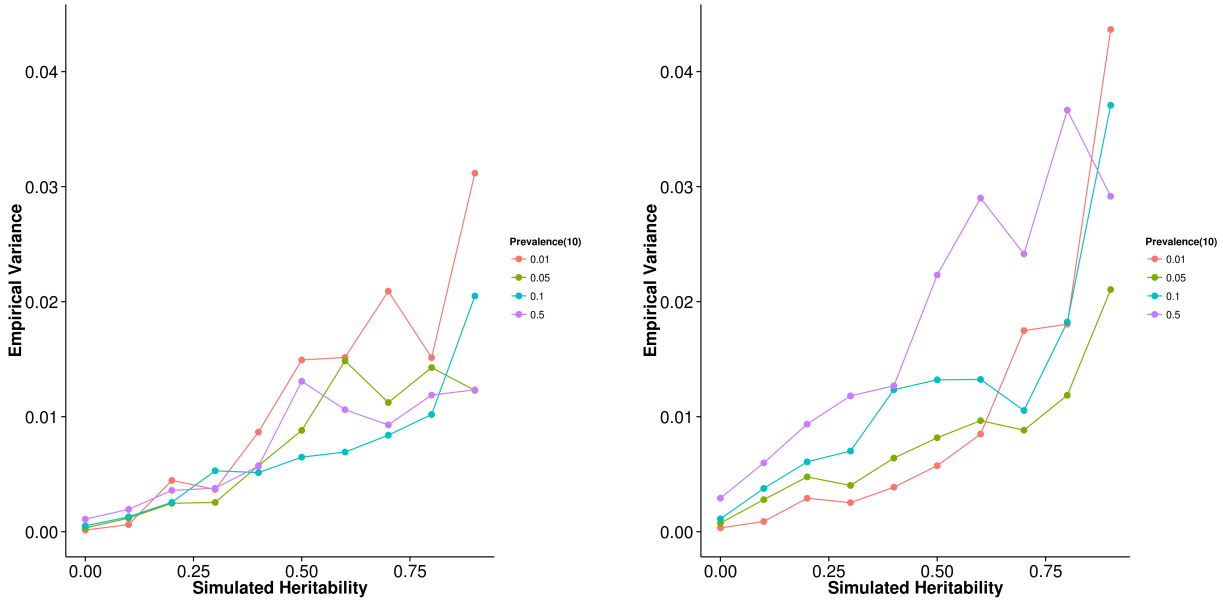
(c) LDSC with fix intercept **(d)** LDSC with intercept estimation

Figure 2.10: Mean of results from case control simulation with random effect size simulation with 10 causal SNPs. The performance of GCTA was as suggested by Golan, Eric S Lander, and Rosset (2014) where there was an underestimation as prevalence decreases. On the other hand, the upward bias of both LDSC with fixed intercept and SHREK increases as the prevalence decreases whereas LDSC with intercept estimation seems relatively robust to the change in prevalence.



(a) SHREK

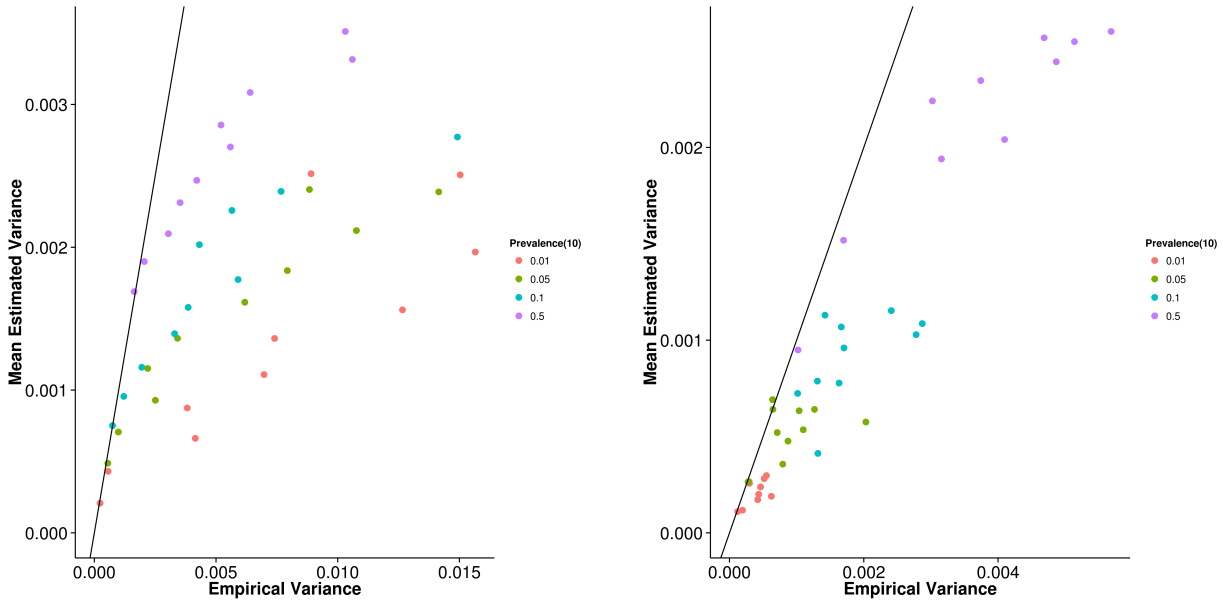
(b) GCTA



(c) LDSC with fix intercept

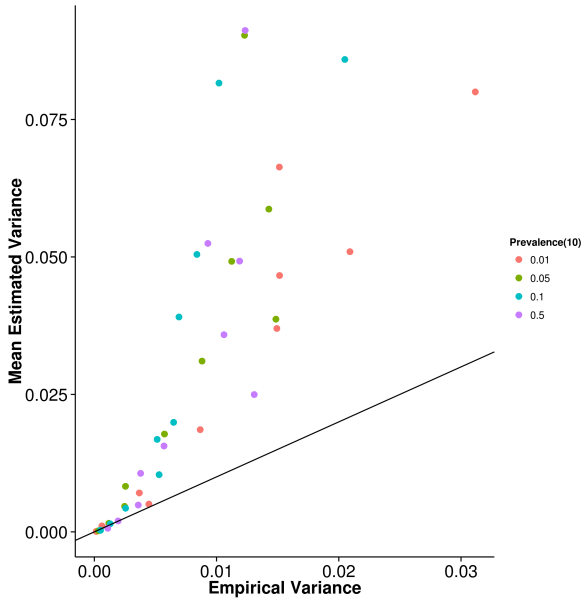
(d) LDSC with intercept estimation

Figure 2.11: Variance of results from case control simulation with random effect size simulation with 10 causal SNPs. There were no clear pattern as to how the prevalence affect the empirical variance of estimates from SHREK and LDSC. For GCTA, it seems like a larger prevalence tends to result in a larger empirical variance. Again, GCTA has the lowest variance, follow by SHREK and LDSC with fixed intercept. Nonetheless, it was important to remember that in case control simulation, a much smaller amount of SNPs was used, thus the results was not directly comparable to results from the quantitative simulation.

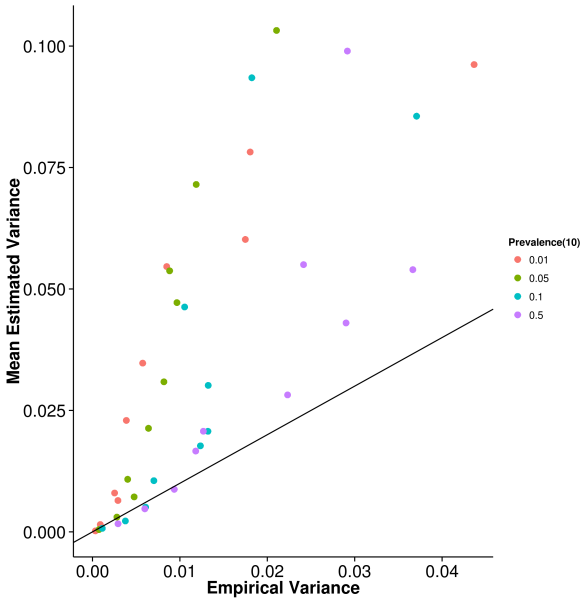


(a) SHREK

(b) GCTA



(c) LDSC with fix intercept



(d) LDSC with intercept estimation

Figure 2.12: Estimated variance of results from case control simulation with random effect size simulation when compared to empirical variance when 10 causal SNPs was simulated. A general underestimation was observed for SHREK and GCTA whereas a larger upward bias was observed for LDSC.

that the intercept estimation might have corrected for part of the bias of LDSC. The same pattern were also observed when the number of causal SNPs increases (figs. S1, S4 and S7), suggesting that the effect of number of causal SNPs were not the main contributor to the difference in bias.

As one inspect the empirical variance of the algorithms, GCTA clearly has the smallest average empirical variance among the algorithms (fig. 2.11b) where LDSC with intercept estimation has the largest empirical variance (fig. 2.11d). Unlike the quantitative trait simulation, the empirical variance of the estimates from SHREK (fig. 2.11a) seems to be very close to that of LDSC with fixed intercept (fig. 2.11c). When the heritability of the trait is high, the empirical variance of SHREK is even lower than that of LDSC with fixed intercept. As one increases the number of causal SNPs, the empirical variance of all algorithms decreases (figs. S2, S5 and S8) agreeing with the results from the quantitative trait simulation.

On the other hand, both SHREK (fig. 2.12a) and GCTA (fig. 2.12b) underestimates their empirical variance whereas LDSC overestimates its empirical variance no matter if the intercept estimation was performed (fig. 2.12). As the number of causal SNPs increases (figs. S3, S6 and S9), the bias of variance estimation remain unchanged for SHREK. However, for LDSC, the magnitude of bias of variance estimation reduces as the number of causal SNPs increases and were able to provide a relatively accurate estimation of its empirical variance when there were 500 causal SNPs (fig. S9c).

Taking into account of the bias and variance of the estimations (table 2.3), SHREK has the best average performance of all the algorithm tested. Interestingly, the performance of LDSC with intercept estimation were better than LDSC with fixed intercept when the prevalence is small. However, considering that we did not simulate any confounding variables, we would expect the intercept estimation to be unnecessary and will only increase the SE of the heritability estimation without improving the estimates. Yet from the simulation results, it suggests that the intercept estimation might have corrected some bias, leading to a better performance when intercept estimation was performed.

Another interesting observation was that although the MSE tends to decrease when the population prevalence increases for most algorithms, an inversed relationship was observed for LDSC when intercept estimation was performed. As the population prevalence increases, the MSE also increases for LDSC with intercept estimations and further investigation might be required to understand the specific reason of this behavior.

In general, the effect of the number of causal SNPs in the case control simulation

Population Prevalence	Number of Causal SNPs	SHREK	LDSC	LDSC-In	GCTA
0.01	10	**0.0145**	0.0361	0.0164	0.0675
0.01	50	0.0135	0.0254	**0.00791**	0.0702
0.01	100	0.0128	0.0227	**0.0102**	0.0698
0.01	500	**0.0126**	0.0214	0.0150	0.0710
0.05	10	0.0110	0.0201	**0.00983**	0.0302
0.05	50	**0.00453**	0.00974	0.0115	0.0299
0.05	100	**0.00569**	0.0113	0.00981	0.0304
0.05	500	**0.00540**	0.00999	0.0171	0.0305
0.1	10	**0.00512**	0.0109	0.0301	0.0165
0.1	50	**0.00381**	0.00824	0.0105	0.0152
0.1	100	**0.00418**	0.00802	0.0163	0.0148
0.1	500	**0.00400**	0.00740	0.0141	0.0155
0.5	10	0.00560	0.00749	0.0219	**0.00410**
0.5	50	0.00362	0.00528	0.0232	**0.00244**
0.5	100	0.00356	0.00460	0.0208	**0.00225**
0.5	500	0.00338	0.00365	0.0159	**0.00200**

Table 2.3: MSE of Case Control simulation. Algorithm with the best performance under each condition were bold-ed. When the population prevalence is 0.5, GCTA has the best performance, followed by SHREK. For most other conditions, SHREK has the best performance. Of all the algorithms, SHREK has the lowest average MSE. Interestingly, although most algorithms has a decreased MSE as the population prevalence increases, LDSC with intercept estimation behaves in the opposite way. Further investigation might be required to understand the reversed behavior of LDSC with intercept estimation when compared to other algorithms, especially when the same pattern was not observed for LDSC with fixed intercept estimation. On the other hand, one can observed that the MSE tends to decrease as the number of causal SNPs increases, as observed in the quantitative simulation.

agrees with what was observed in the quantitative trait simulations where as the number of causal SNPs increases the MSE tends to decrease for all algorithms, with SHREK least sensitive. Finally, it is important to note that for the case control simulation, a smaller amount of SNPs was simulated when compared to that in the quantitative trait simulation. The total sample number involved was also larger (2,000 samples with 1,000 cases and 1,000 controls). Thus, the result from this case control simulation was not directly comparable to the results from the quantitative trait simulation.

2.7.3 Extreme Phenotype Simulation

2.7.4 Real Data Simulation

2.7.5 Application to Real Data

We applied our methods to the PGC schizophrenia (SCZ), major depression disorder, autism and bipolar data sets.

2.8 Discussion

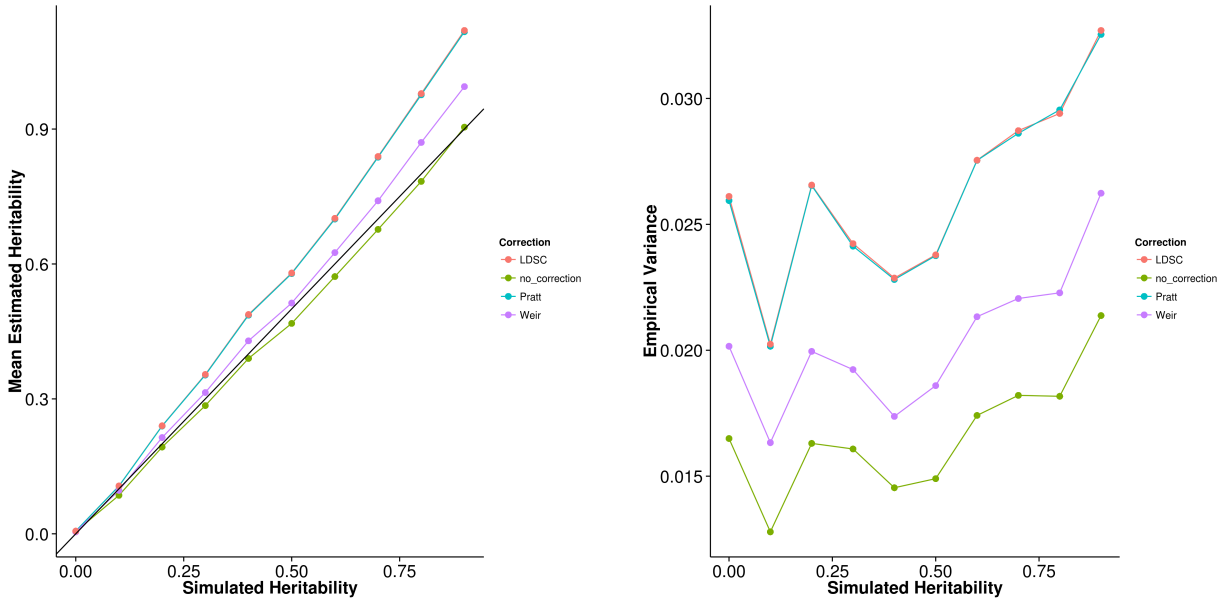
In this chapter, we introduced an alternative method to LDSC for the estimation of heritability using only the test statistics and have made available for download a standalone C++ programme SHREK.

2.8.1 LD Correction

As SHREK relies on the LD information to estimate the heritability, it is important for us to correct for possible sampling bias form the LD matrix. When testing for the bias correction algorithms, it was observed that majority of the algorithms, except that of eq. (2.41), inflates the heritability estimated. Therefore, our initial decision was to use eq. (2.41) as our LD sampling bias correction algorithm. However, in subsequent simulation of the quantitative traits, a general overestimation of the heritability was again observed. Because the pattern of overestimation is close to the one observed in the LD correction simulation, we hypothesize eq. (2.41) might also have introduced bias into the heritability estimation but the bias is only noticeable when the number of SNPs increases. When compared the SHREK, LDSC only overestimates the heritability when the number of causal SNPs is small. When inspecting their algorithm, it was observed that LDSC also corrected for the sampling bias on the R^2, but with yet another equation:

$$\text{LDSC} : \tilde{R^2} = \hat{R^2} - \frac{1 - \hat{R^2}}{n - 2} \tag{2.45}$$

It is therefore important to test whether if the application of eq. (2.45) might help to improve the performance of SHREK and whether if the performance of eq. (2.41) does decrease when



(a) Mean Estimation

(b) Empirical Variance

Figure 2.13: Effect of LD correction to Heritability Estimation when 50,000 SNPs were simulated. As an overestimation was observed in the quantitative trait simulation, we performed a short simulation to assess the impact of LD correction to the heritability of SHREK when the there is a larger number of SNPs. From the graph, it was observed that all LD correction algorithms inflate the heritability estimation when large number of SNPs were simulated. In fact, the bias was the smallest when no LD correction was performed.

the number of SNPs increases. We therefore repeated the simulation on LD correction but instead of simulating 5,000 SNPs on chromosome 22, we simulated 50,000 SNPs on chromosome 1. To reduce the run time of the simulation, we only compared the performance of SHREK when eq. (2.45), eq. (2.41) and eq. (2.39) were used for the LD correction.

Surprisingly, when large amount of SNPs were estimated, the bias of heritability estimation was the smallest when no LD correction was performed (fig. 2.13) where only a small downward bias was observed. The performance of eq. (2.45) was comparable to eq. (2.39) and eq. (2.41) introduce least bias to the estimation. It was clear that the genomic data structure (e.g. 0,1,2) violates some of the underlying assumption of the LD correction algorithms (e.g. normality) thus general LD correction algorithms might not be applicable to our situation. Further research is therefore required to find an optimal solution to correct for the LD sampling bias. Meanwhile, we allow users the freedom to disable the LD correction in SHREK.

2.8.2 Simulation Results

Chapter 3

Conclusion

Bibliography

Aberg, Karolina et al. (2010). "Genomewide association study of movement-related adverse antipsychotic effects." eng. In: *Biological psychiatry* 67.3, pp. 279–282. DOI: 10.1016/j.biopsych.2009.08.036.

Adkins, D E et al. (2011). "Genomewide pharmacogenomic study of metabolic side effects to antipsychotic drugs." eng. In: *Molecular psychiatry* 16.3, pp. 321–332. DOI: 10.1038/mp.2010.14.

Aghajanian, G K and G J Marek (1999). "Serotonin and hallucinogens." eng. In: *Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology* 21.2 Suppl, 16S–23S. DOI: 10.1016/S0893-133X(98)00135-3.

Alkelai, Ana et al. (2009). "Genome-wide association study of antipsychotic-induced parkinsonism severity among schizophrenia patients." eng. In: *Psychopharmacology* 206.3, pp. 491–499. DOI: 10.1007/s00213-009-1627-z.

Altshuler, David M et al. (2010). "Integrating common and rare genetic variation in diverse human populations." In: *Nature* 467.7311, pp. 52–58. DOI: 10.1038/nature09298 (cit. on pp. 45, 48).

Alvir, J M et al. (1993). "Clozapine-induced agranulocytosis. Incidence and risk factors in the United States." eng. In: *The New England journal of medicine* 329.3, pp. 162–167. DOI: 10.1056/NEJM199307153290303.

American Psychiatric Association (2013). *Diagnostic and Statistical Manual of Mental Disorders*. American Psychiatric Publishing, p. 991. DOI: 10.1176/appi.books.9780890425596.744053.

Arinami, T et al. (1997). "A functional polymorphism in the promoter region of the dopamine D2 receptor gene is associated with schizophrenia." eng. In: *Human molecular genetics* 6.4, pp. 577–582.

Arranz, M J and J de Leon (2007). "Pharmacogenetics and pharmacogenomics of schizophrenia: a review of last decade of research". In: *Mol Psychiatry* 12.8, pp. 707–747.

Arranz, Maria J and Janet C Munro (2011). "Toward understanding genetic risk for differential antipsychotic response in individuals with schizophrenia". In: *Expert Review of Clinical Pharmacology* 4.3, pp. 389–405. DOI: 10.1586/ecp.11.16.

Bernstein, Bradley E et al. (2010). "The NIH Roadmap Epigenomics Mapping Consortium." eng. In: *Nature biotechnology* 28.10, pp. 1045–1048. DOI: 10.1038/nbt1010-1045.

Bouchard, Thomas J (2013). "The Wilson Effect: the increase in heritability of IQ with age." In: *Twin research and human genetics : the official journal of the International Society for Twin Studies* 16.5, pp. 923–30. DOI: 10.1017/thg.2013.54.

Brown, A S and E J Derkits (2010). "Prenatal infection and schizophrenia: a review of epidemiologic and translational studies". eng. In: *Am J Psychiatry* 167.3, pp. 261–280. DOI: appi.ajp.2009.09030361[pii]10.1176/appi.ajp.2009.09030361.

Buckley, Peter F et al. (2009). "Psychiatric Comorbidities and Schizophrenia". In: *Schizophrenia Bulletin* 35.2, pp. 383–402. DOI: 10.1093/schbul/sbn135.

Bulik-Sullivan, Brendan K et al. (2015). "LD Score regression distinguishes confounding from polygenicity in genome-wide association studies". In: *Nature Genetics* 47.3, pp. 291–295. DOI: 10.1038/ng.3211 (cit. on pp. 33, 50, 59).

Bulik-Sullivan, Brendan et al. (2015). "An atlas of genetic correlations across human diseases and traits". In: *Nat Genet* advance online publication.

Cacabelos, Ramon, Ryota Hashimoto, and Masatoshi Takeda (2011). "Pharmacogenomics of antipsychotics efficacy for schizophrenia." eng. In: *Psychiatry and clinical neurosciences* 65.1, pp. 3–19. DOI: 10.1111/j.1440-1819.2010.02168.x.

Consortium, The International HapMap (2005). "A haplotype map of the human genome". In: *Nature* 437, pp. 1299–1320. DOI: 10.1038/nature04226.

Deverman, B E and P H Patterson (2009). "Cytokines and CNS development". eng. In: *Neuron* 64.1, pp. 61–78. DOI: 10.1016/j.neuron.2009.09.002S0896-6273(09)00680-1[pii].

ENCODE Project Consortium (2012). "An integrated encyclopedia of DNA elements in the human genome". In: *Nature* 489.7414, pp. 57–74.

Falconer, Douglas S (1965). "The inheritance of liability to certain diseases, estimated from the incidence among relatives". In: *Annals of Human Genetics* 29.1, pp. 51–76. DOI: 10.1111/j.1469-1809.1965.tb00500.x.

Falconer, Douglas S and Trudy F C Mackay (1996). *Introduction to Quantitative Genetics (4th Edition)*. Vol. 12, p. 464.

Feuk, Lars, Andrew R Carson, and Stephen W Scherer (2006). "Structural variation in the human genome". In: *Nat Rev Genet* 7.2, pp. 85–97.

Finucane, Hilary K et al. (2015). "Partitioning heritability by functional annotation using genome-wide association summary statistics". In: *Nat Genet* advance online publication.

Gardner, E L, L S Walker, and W Paredes (1993). "Clozapine's functional mesolimbic selectivity is not duplicated by the addition of anticholinergic action to haloperidol: a brain stimulation study in the rat." eng. In: *Psychopharmacology* 110.1-2, pp. 119–124.

Giovanoli, S. et al. (2013). "Stress in puberty unmasks latent neuropathological consequences of prenatal immune activation in mice". eng. In: *Science* 339.6123, pp. 1095–1099. DOI: 10.1126/science.1228261339/6123/1095[pii].

Golan, David, Eric S Lander, and Saharon Rosset (2014). "Measuring missing heritability: Inferring the contribution of common variants". In: *Proceedings of the National Academy of Sciences* 111.49, E5272–E5281. DOI: 10.1073/pnas.1419064111 (cit. on pp. 33, 67, 68).

Gottesman, I I and J Shields (1967). "A polygenic theory of schizophrenia". In: *Proceedings of the National Academy of Sciences* 58.1, pp. 199–205.

Gottesman, II (1991). *Schizophrenia genesis: The origins of madness.* WH Freeman/Times Books/Henry Holt & Co.

Gottesman, II and J Shields (1967). "A polygenic theory of schizophrenia". In: *Proceedings of the National Academy of Sciences* 58.1, pp. 199–205.

Gottesman, Irving I and James Shields (1982). *Schizophrenia: The Epigenetic Puzzle*. Cambridge University Press.

Guennebaud, Gaël, Beno\^it Jacob, et al. (2010). *Eigen v3*. http://eigen.tuxfamily.org (cit. on pp. 43, 46).

Guey, Lin T. et al. (2011). "Power in the phenotypic extremes: A simulation study of power in discovery and replication of rare variants". In: *Genetic Epidemiology* 35.4, pp. 236–246. DOI: 10.1002/gepi.20572 (cit. on p. 42).

Gui, Hongsheng et al. (2013). "RET and NRG1 interplay in Hirschsprung disease." eng. In: *Human genetics* 132.5, pp. 591–600. DOI: 10.1007/s00439-013-1272-9 (cit. on p. 53).

Hansen, Per Christian (1987). "The truncated SVD as a method for regularization". In: *Bit* 27.4, pp. 534–553. DOI: 10.1007/BF01937276 (cit. on pp. 44, 45).

Harrison, P J and D R Weinberger (2005). "Schizophrenia genes, gene expression, and neuropathology: on the matter of their convergence." In: *Molecular psychiatry* 10.1, 40–68, image 5. DOI: 10.1038/sj.mp.4001686.

Harvey, Philip D. et al. (2012). "Diagnosis of schizophrenia: Consistency across information sources and stability of the condition". In: *Schizophrenia Research* 140.1-3, pp. 9–14. DOI: 10.1016/j.schres.2012.03.026.

Heston, Leonard L (1966). "Psychiatric Disorders in Foster Home Reared Children of Schizophrenic Mothers". In: *The British Journal of Psychiatry* 112.489, pp. 819–825.

Hinrichs, A S et al. (2006). "The UCSC Genome Browser Database: update 2006." eng. In: *Nucleic acids research* 34.Database issue, pp. D590–8. DOI: 10.1093/nar/gkj144 (cit. on p. 57).

Ho, Nghia (2011). *OPENCV VS. ARMADILLO VS. EIGEN ON LINUX* (cit. on p. 46).

Howes, O D and S Kapur (2009). "The dopamine hypothesis of schizophrenia: version III–the final common pathway". eng. In: *Schizophr Bull* 35.3, pp. 549–562. DOI: 10.1093/schbul/sbp006sbp006[pii].

Jablensky, Assen (2010). "The diagnostic concept of schizophrenia: its history, evolution, and future prospects." In: *Dialogues in clinical neuroscience* 12.3, pp. 271–87.

Jeanneteau, Freddy et al. (2006). "A functional variant of the dopamine D3 receptor is associated with risk and age-at-onset of essential tremor." eng. In: *Proceedings of the National Academy of Sciences of the United States of America* 103.28, pp. 10753–10758. DOI: 10.1073/pnas.0508189103.

Jones, Peter B et al. (2006). "Randomized controlled trial of the effect on Quality of Life of second- vs first-generation antipsychotic drugs in schizophrenia: Cost Utility of the Latest Antipsychotic Drugs in Schizophrenia Study (CUtLASS 1)." eng. In: *Archives of general psychiatry* 63.10, pp. 1079–1087. DOI: 10.1001/archpsyc.63.10.1079.

Jorgensen, Andrea L and Paula R Williamson (2008). "Methodological quality of pharmacogenetic studies: Issues of concern". In: *Statistics in Medicine* 27.30, pp. 6547–6569. DOI: 10.1002/sim.3420.

Kapur, Shitij and David Mamo (2003). "Half a century of antipsychotics and still a central role for dopamine D2 receptors." eng. In: *Progress in neuro-psychopharmacology & biological psychiatry* 27.7, pp. 1081–1090. DOI: 10.1016/j.pnpbp.2003.09.004.

Kay, Stanley R, Abraham Fiszbein, and Lewis A Opler (1987). "The Positive and Negative Syndrome Scale (PANSS) for Schizophrenia". In: *Schizophrenia Bulletin* 13.2, pp. 261–276. DOI: 10.1093/schbul/13.2.261.

Kelly, C and R G McCreadie (1999). "Smoking habits, current symptoms, and premorbid characteristics of schizophrenic patients in Nithsdale, Scotland." eng. In: *The American journal of psychiatry* 156.11, pp. 1751–1757.

Knapp, Martin, Roshni Mangalore, and Judit Simon (2004). "The global costs of schizophrenia." In: *Schizophrenia bulletin* 30.2, pp. 279–293.

Kumar, Subodh and Suprakash Chaudhury (2014). "Efficacy of amisulpride and olanzapine for negative symptoms and cognitive impairments: An open-label clinical study". In: *Industrial Psychiatry Journal* 23.1, pp. 27–35. DOI: 10.4103/0972-6748.144953.

Lander, E S et al. (2001). "Initial sequencing and analysis of the human genome." eng. In: *Nature* 409.6822, pp. 860–921. DOI: 10.1038/35057062.

Lehmann, H E and T A Ban (1997). "The history of the psychopharmacology of schizophrenia." eng. In: *Canadian journal of psychiatry. Revue canadienne de psychiatrie* 42.2, pp. 152–162.

Li, Miao-Xin Xin et al. (2011). "Evaluating the effective numbers of independent tests and significant p-value thresholds in commercial genotyping arrays and public imputation reference datasets". In: *Human Genetics* 131.5, pp. 747–756. DOI: 10.1007/s00439-011-1118-2 (cit. on p. 41).

Li, Na and Matthew Stephens (2003). "Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data." eng. In: *Genetics* 165.4, pp. 2213–2233 (cit. on p. 49).

Lichtenstein, Paul et al. (2009). "Common genetic determinants of schizophrenia and bipolar disorder in Swedish families: a population-based study". In: *The Lancet* 373.9659, pp. 234–239. DOI: 10.1016/S0140-6736(09)60072-6.

Lieberman, Jeffrey A et al. (2005). "Effectiveness of Antipsychotic Drugs in Patients with Chronic Schizophrenia". In: *New England Journal of Medicine* 353.12, pp. 1209–1223. DOI: 10.1056/NEJMoa051688.

Malhotra, A K, J-P Zhang, and T Lencz (2012). "Pharmacogenetics in psychiatry: translating research into clinical practice". In: *Mol Psychiatry* 17.8, pp. 760–769.

Mata, I et al. (2001). "Olanzapine: concordant response in monozygotic twins with schizophrenia". In: *The British Journal of Psychiatry* 178.1, p. 86.

McClay, J L et al. (2011). "Genome-wide pharmacogenomic analysis of response to treatment with antipsychotics." eng. In: *Molecular psychiatry* 16.1, pp. 76–85. DOI: 10.1038/mp.2009.89.

McClellan, Jon M, Ezra Susser, and Mary-Claire King (2007). "Schizophrenia: a common disease caused by multiple rare alleles". In: *The British Journal of Psychiatry* 190.3, pp. 194–199.

McGrath, John et al. (2008). "Schizophrenia: A Concise Overview of Incidence, Prevalence, and Mortality". In: *Epidemiologic Reviews* 30.1, pp. 67–76. DOI: 10.1093/epirev/mxn001.

Mednick (1958). "Schizophrenia Following Prenatal Exposure to an Influenza Epidemic". In: 1111.1.

Meltzer, Herbert Y (1991). "The Mechanism of Action of Novel Antipsychotic Drugs". In: *Schizophrenia Bulletin* 17.2, pp. 263–287. DOI: 10.1093/schbul/17.2.263.

Meltzer, Herbert Y (1999). "The Role of Serotonin in Antipsychotic Drug Action". In: *Neuropsychopharmacology* 21.S1, 106S–115S.

Meyer, U, J Feldon, and B K Yee (2009). "A review of the fetal brain cytokine imbalance hypothesis of schizophrenia". eng. In: *Schizophr Bull* 35.5, pp. 959–972. DOI: 10.1093/schbul/sbn022sbn022[pii].

Meyer, U, B K Yee, and J Feldon (2007). "The neurodevelopmental impact of prenatal infections at different times of pregnancy: the earlier the worse?" eng. In: *Neuroscientist* 13.3, pp. 241–256. DOI: 13/3/241[pii]10.1177/1073858406296401.

Meyer, Urs, Joram Feldon, and S Hossein Fatemi (2009). "In-vivo rodent models for the experimental investigation of prenatal immune activation effects in neurodevelopmental brain disorders". In: *Neuroscience & Biobehavioral Reviews* 33.7, pp. 1061–1079. DOI: http://dx.doi.org/10.1016/j.neubiorev.2009.05.001.

Müller, Norbert and Markus J Schwarz (2010). "Immune System and Schizophrenia". In: *Current immunology reviews* 6.3, pp. 213–220.

Neumaier, Arnold (1998). "Solving Ill-Conditioned and Singular Linear Systems: A Tutorial on Regularization". In: *SIAM Review* 40.3, pp. 636–666. DOI: 10.1137/S0036144597321909 (cit. on p. 43).

Onore, C E et al. (2014). "Maternal immune activation leads to activated inflammatory macrophages in offspring". eng. In: *Brain Behav Immun* 38, pp. 220–226. DOI: 10.1016/j.bbi.2014.02.007S0889-1591(14)00053-1[pii].

Oord, Edwin J C G van den et al. (2009). "A systematic method for estimating individual responses to treatment with antipsychotics in CATIE". In: *Schizophrenia Research* 107.1, pp. 13–21. DOI: http://dx.doi.org/10.1016/j.schres.2008.09.009.

Ording, Anne Gulbech et al. (2013). "Comorbid Diseases Interact with Breast Cancer to Affect Mortality in the First Year after Diagnosis—A Danish Nationwide Matched Cohort Study". In: *PLoS ONE* 8.10, e76013.

Orr, H Allen (1998). "The Population Genetics of Adaptation: The Distribution of Factors Fixed during Adaptive Evolution". In: *Evolution* 52.4, pp. 935–949 (cit. on p. 49).

Paolicelli, R C et al. (2011). "Synaptic pruning by microglia is necessary for normal brain development". eng. In: *Science* 333.6048, pp. 1456–1458. DOI: 10.1126/science.1202529science.1202529[pii].

Pirmohamed, Munir (2001). "Pharmacogenetics and pharmacogenomics". In: *British Journal of Clinical Pharmacology* 52.4, pp. 345–347. DOI: 10.1046/j.0306-5251.2001.01498.x.

Project, Genomes et al. (2012). "An integrated map of genetic variation from 1,092 human genomes". In: *Nature* 491.7422, pp. 56–65. DOI: http://www.nature.com/nature/

journal/v491/n7422/abs/nature11632.html\#supplementary-information (cit. on pp. 48, 57).

Psychiatric GWAS Consortium Bipolar Disorder Working Group (2011). "Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4." eng. In: *Nature genetics* 43.10, pp. 977–983. DOI: 10.1038/ng.943 (cit. on p. 56).

Purcell, S, S S Cherny, and P C Sham (2003). "Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits". en. In: *Bioinformatics* 19, pp. 149–150. DOI: 10.1093/bioinformatics/19.1.149.

Purcell, Shaun M et al. (2009). "Common polygenic variation contributes to risk of schizophrenia and bipolar disorder." eng. In: *Nature* 460.7256, pp. 748–752. DOI: 10.1038/nature08185.

Purcell, Shaun et al. (2007). "PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses". In: *The American Journal of Human Genetics* 81.3, pp. 559–575. DOI: 10.1086/519795 (cit. on p. 49).

Ravyn, Dana et al. (2013). "CYP450 Pharmacogenetic treatment strategies for antipsychotics: A review of the evidence". In: *Schizophrenia Research* 149.1–3, pp. 1–14. DOI: http://dx.doi.org/10.1016/j.schres.2013.06.035.

Remington, Gary et al. (2013). "Clozapine's role in the treatment of first-episode schizophrenia." eng. In: *The American journal of psychiatry* 170.2, pp. 146–151. DOI: 10.1176/appi.ajp.2012.12060778.

Rijsdijk, Fruhling V and Pak C Sham (2002). "Analytic approaches to twin data using structural equation models." eng. In: *Briefings in bioinformatics* 3.2, pp. 119–133.

Riley, Brien and Kenneth S Kendler (2006). "Molecular genetic studies of schizophrenia." In: *European journal of human genetics : EJHG* 14.6, pp. 669–680. DOI: 10.1038/sj.ejhg.5201571.

Ripke, Stephan, Benjamin M. Neale, et al. (2014). "Biological insights from 108 schizophrenia-associated genetic loci". In: *Nature* 511, pp. 421–427. DOI: 10.1038/nature13595.

Ripke, Stephan, Naomi R Wray, et al. (2013). "A mega-analysis of genome-wide association studies for major depressive disorder." eng. In: *Molecular psychiatry* 18.4, pp. 497–511. DOI: 10.1038/mp.2012.21 (cit. on p. 56).

Ripke, S et al. (2013). "Genome-wide association analysis identifies 13 new risk loci for schizophrenia". eng. In: *Nat Genet* 45.10, pp. 1150–1159. DOI: 10.1038/ng.2742ng.2742[pii] (cit. on p. 56).

Risch, N (1990a). "Linkage strategies for genetically complex traits. I. Multilocus models." In: *American Journal of Human Genetics* 46.2, pp. 222–228.

Risch, N (1990b). "Linkage strategies for genetically complex traits. II. The power of affected relative pairs." In: *American Journal of Human Genetics* 46.2, pp. 229–241.

Saha, Sukanta, David Chant, and John Mcgrath (2007). "A Systematic Review of Mortality in Schizophrenia". In: *Archives of general psychiatry* 64.10, pp. 1123–1131. DOI: 10.1001/archpsyc.64.10.1123.

Sanderson, Conrad (2010). *Armadillo: An Open Source C++ Linear Algebra Library for Fast Prototyping and Computationally Intensive Experiments*. Tech. rep. (cit. on p. 46).

Schultz, Stephen H., Stephen W. North, and Cleveland G. Shields (2007). "Schizophrenia: A review". In: *American Family Physician* 75.12, pp. 1821–1829.

Sham, Pak C and Shaun M Purcell (2014). "Statistical power and significance testing in large-scale genetic studies." In: *Nature reviews. Genetics* 15.5, pp. 335–46. DOI: 10.1038/nrg3706 (cit. on pp. 42, 56).

Smith, S E et al. (2007). "Maternal immune activation alters fetal brain development through interleukin-6". eng. In: *J Neurosci* 27.40, pp. 10695–10702. DOI: 27/40/10695[pii]10.1523/JNEUROSCI.2178-07.2007.

Søgaard, Mette et al. (2013). "The impact of comorbidity on cancer survival: a review". In: *Clinical Epidemiology* 5.Suppl 1, pp. 3–29. DOI: 10.2147/CLEP.S47150.

Sokoloff, P et al. (2006). "The dopamine D3 receptor: a therapeutic target for the treatment of neuropsychiatric disorders." eng. In: *CNS & neurological disorders drug targets* 5.1, pp. 25–43.

Su, Zhan, Jonathan Marchini, and Peter Donnelly (2011). "HAPGEN2: Simulation of multiple disease SNPs". In: *Bioinformatics* 27.16, pp. 2304–2305. DOI: 10.1093/bioinformatics/btr341 (cit. on pp. 49, 52).

Sullivan, Patrick F (2005). "The Genetics of Schizophrenia". In: *PLoS Med* 2.7, e212.

Sullivan, Patrick F, Kenneth S Kendler, and Michael C Neale (2003). "Schizophrenia as a Complex Trait". In: *Archives of general psychiatry* 60, pp. 1187–1192. DOI: 10.1001/archpsyc.60.12.1187.

Szatkiewicz, J P et al. (2014). "Copy number variation in schizophrenia in Sweden". In: *Mol Psychiatry* 19.7, pp. 762–773.

Talkowski, Michael E et al. (2007). "Dopamine Genes and Schizophrenia: Case Closed or Evidence Pending?" In: *Schizophrenia Bulletin* 33.5, pp. 1071–1081.

Tandon, Rajiv (2007). *Antipsychotic treatment of schizophrenia: two steps forward, one step back.* eng.

Tienari, Pekka et al. (2004). "Genotype-environment interaction in schizophrenia-spectrum disorder". In: *The British Journal of Psychiatry* 184.3, pp. 216–222.

Tsuang, Ming T., William S. Stone, and Stephen V. Faraone (2000). "Toward reformulating the diagnosis of schizophrenia". In: *American Journal of Psychiatry* 157.7, pp. 1041–1050. DOI: 10.1176/appi.ajp.157.7.1041.

Üçok, A L P and Wolfgang Gaebel (2008). "Side effects of atypical antipsychotics: a brief overview". In: *World Psychiatry* 7.1, pp. 58–62.

Visscher, Peter M, William G Hill, and Naomi R Wray (2008). "Heritability in the genomics era [mdash] concepts and misconceptions". In: *Nat Rev Genet* 9.4, pp. 255–266.

Vojvoda, Dolores et al. (1996). "Monozygotic twins concordant for response to clozapine". In: *The Lancet* 347.8993, p. 61. DOI: http://dx.doi.org/10.1016/S0140-6736(96)91594-9.

Vuillermot, Stéphanie et al. (2010). "A longitudinal examination of the neurodevelopmental impact of prenatal immune activation in mice reveals primary defects in dopaminergic development relevant to schizophrenia". eng. In: *J Neurosci* 30.4, pp. 1270–1287. DOI: 10.1523/JNEUROSCI.5408-09.201030/4/1270[pii].

Wang, Zhongmiao and Bruce Thompson (2007). "Is the Pearson r 2 Biased, and if So, What Is the Best Correction Formula?" In: *The Journal of Experimental Education* 75.2, pp. 109–125. DOI: 10.3200/JEXE.75.2.109-125 (cit. on p. 48).

Weir, B S and W G Hill (1980). "EFFECT OF MATING STRUCTURE ON VARIATION IN LINKAGE DISEQUILIBRIUM". In: *Genetics* 95.2, pp. 477–488 (cit. on p. 48).

Welter, Danielle et al. (2014). "The NHGRI GWAS Catalog, a curated resource of SNP-trait associations". In: *Nucleic Acids Research* 42.D1, pp. 1001–1006. DOI: 10.1093/nar/gkt1229 (cit. on p. 51).

World Health Organization (2013). *WHO methods and data sources for global burden of disease estimates*. Tech. rep. Geneva.

Yang, Jian, Michael N Weedon, et al. (2011). "Genomic inflation factors under polygenic inheritance". In: *Eur J Hum Genet* 19.7, pp. 807–812.

Yang, Jian, Naomi R. Wray, and Peter M. Visscher (2010). "Comparing apples and oranges: Equating the power of case-control and quantitative trait association studies". In: *Genetic Epidemiology* 34.3, pp. 254–257. DOI: 10.1002/gepi.20456 (cit. on p. 41).

Yang, J et al. (2011). "GCTA: a tool for genome-wide complex trait analysis". eng. In: *Am J Hum Genet* 88.1, pp. 76–82. DOI: 10.1016/j.ajhg.2010.11.011S0002-9297(10)00598-7[pii] (cit. on p. 50).

Zhang, Jian-Ping and Anil K Malhotra (2011). "Pharmacogenetics and antipsychotics: therapeutic efficacy and side effects prediction." eng. In: *Expert opinion on drug metabolism & toxicology* 7.1, pp. 9–37. DOI: 10.1517/17425255.2011.532787.

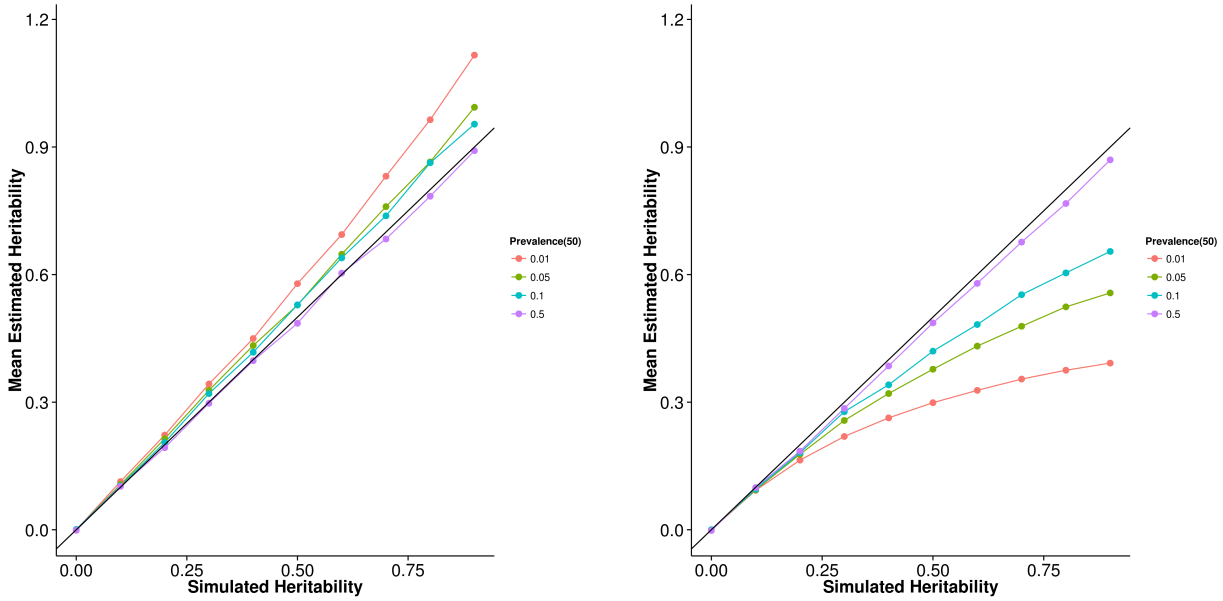
Zhang, Jian-Ping and Anil K Malhotra (2013). "Pharmacogenetics of antipsychotics: recent progress and methodological issues." eng. In: *Expert opinion on drug metabolism & toxicology* 9.2, pp. 183–191. DOI: 10.1517/17425255.2013.736964.

Zhao, B and J P Schwartz (1998). "Involvement of cytokines in normal CNS development and neurological diseases: recent progress and perspectives". eng. In: *J Neurosci Res* 52.1, pp. 7–16. DOI: 10.1002/(SICI)1097-4547(19980401)52:1<7::AID-JNR2>3.0.CO;2-I[pii].

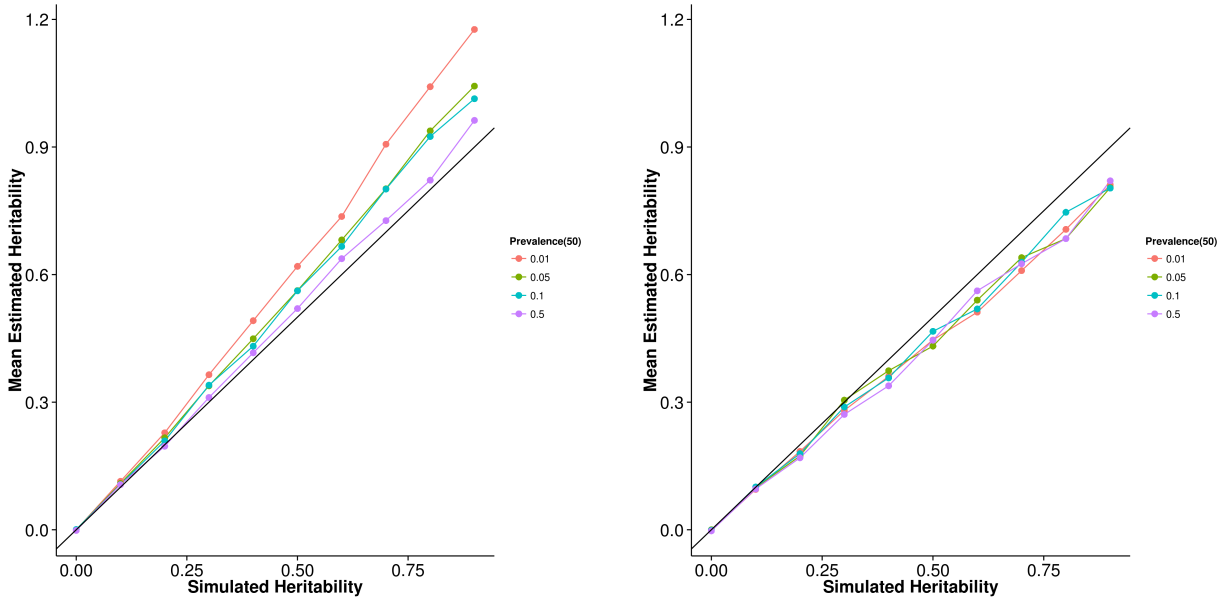
Zheng, Gang, Boris Freidlin, and Joseph L Gastwirth (2006). "Robust genomic control for association studies." eng. In: *American journal of human genetics* 78.2, pp. 350–356. DOI: 10.1086/500054.

Supplementary Materials

Appendix

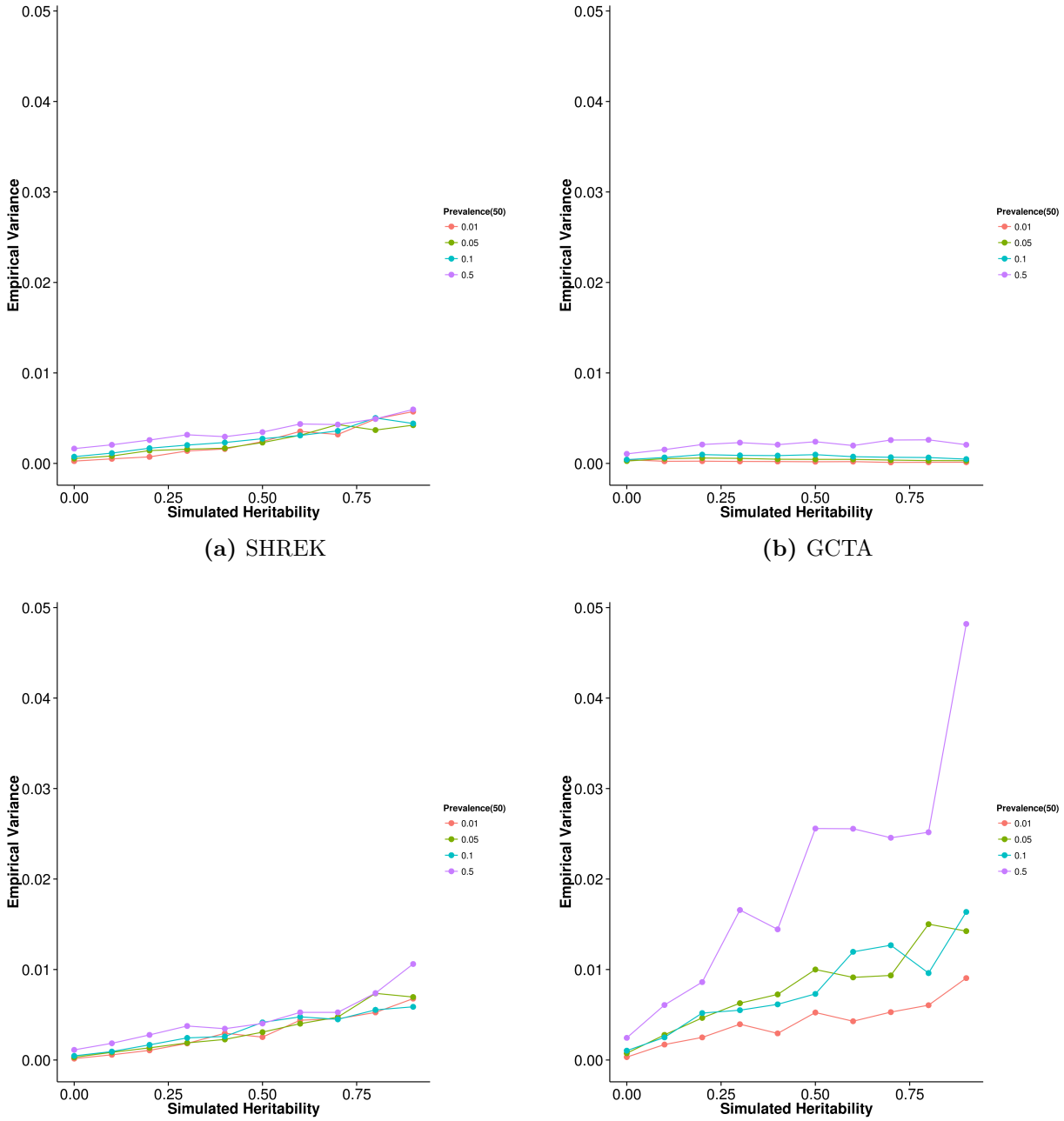


(a) SHREK **(b)** GCTA



(c) LDSC with fix intercept **(d)** LDSC with intercept estimation

Figure S1: Mean of results from case control simulation with random effect size simulation with 50 causal SNPs. In general, the results were similar to the scenario with 10 causal SNPs with the only exception that the estimates from LDSC with intercept estimates seems to be less affected by the change in prevalence of the trait.



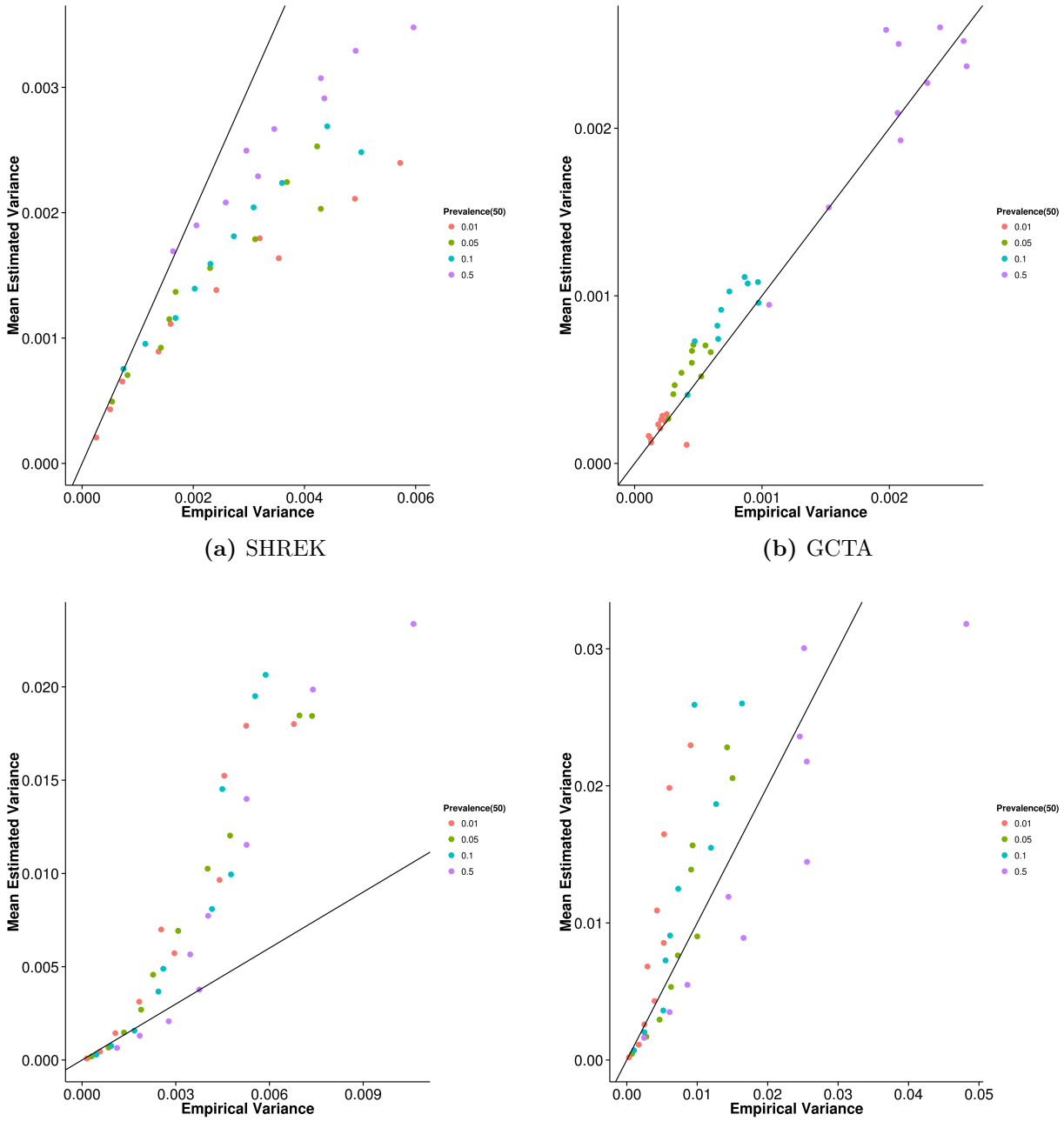
(a) SHREK

(b) GCTA

(c) LDSC with fix intercept

(d) LDSC with intercept estimation

Figure S2: Variance of results from case control simulation with random effect size simulation with 50 causal SNPs. For most algorithm except that of LDSC with fixed intercept, the empirical variance of the estimates increases as the population prevalence of the trait increases, with the estimations from LDSC with intercept estimation display the largest variance.



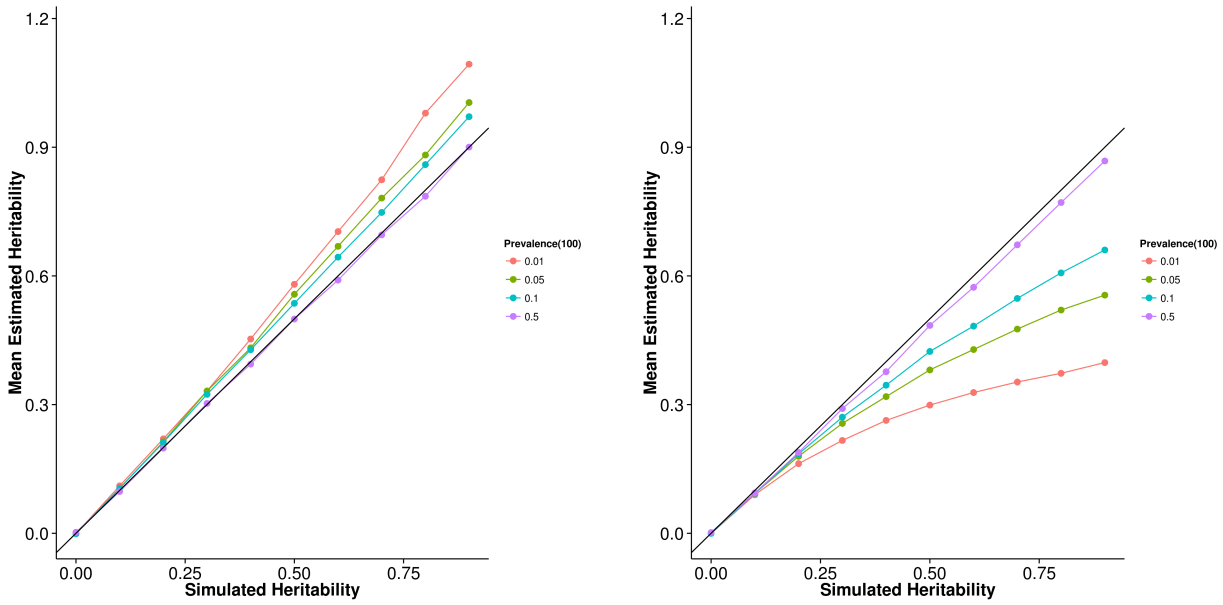
(a) SHREK

(b) GCTA

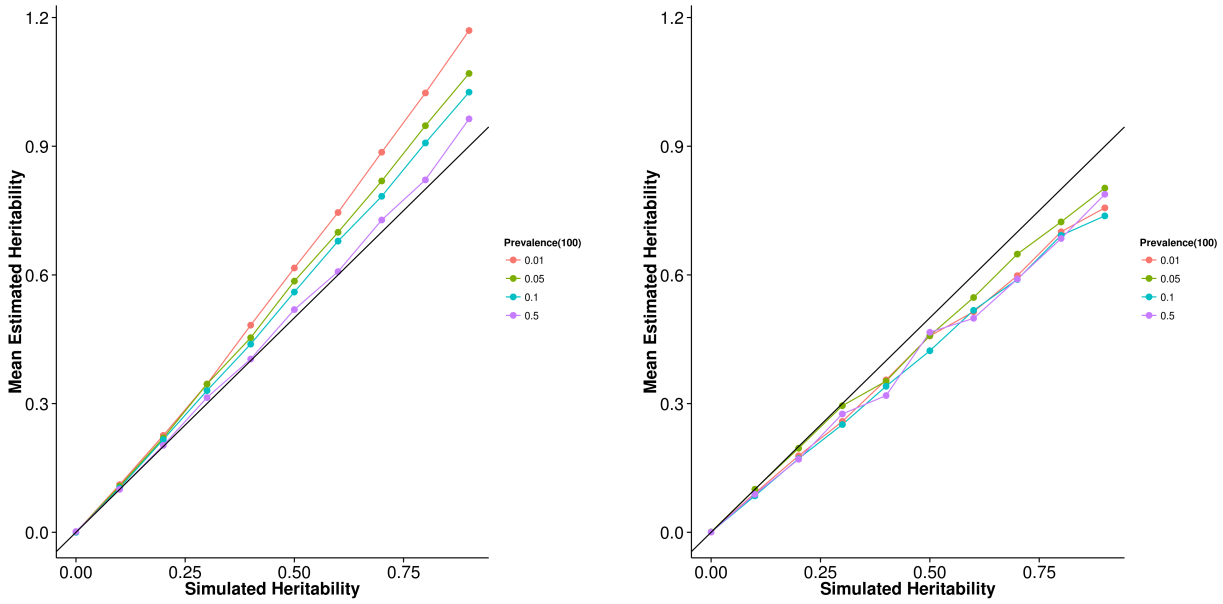
(c) LDSC with fix intercept

(d) LDSC with intercept estimation

Figure S3: Estimated variance of results from case control simulation with random effect size simulation when compared to empirical variance when 50 causal SNPs was simulated. Again, the estimation of variance from SHREK tends to be downwardly biased and LDSC with fixed intercept tends to be upwardly biased. However, when intercept estimation was performed, the estimation of variance of LDSC improved.

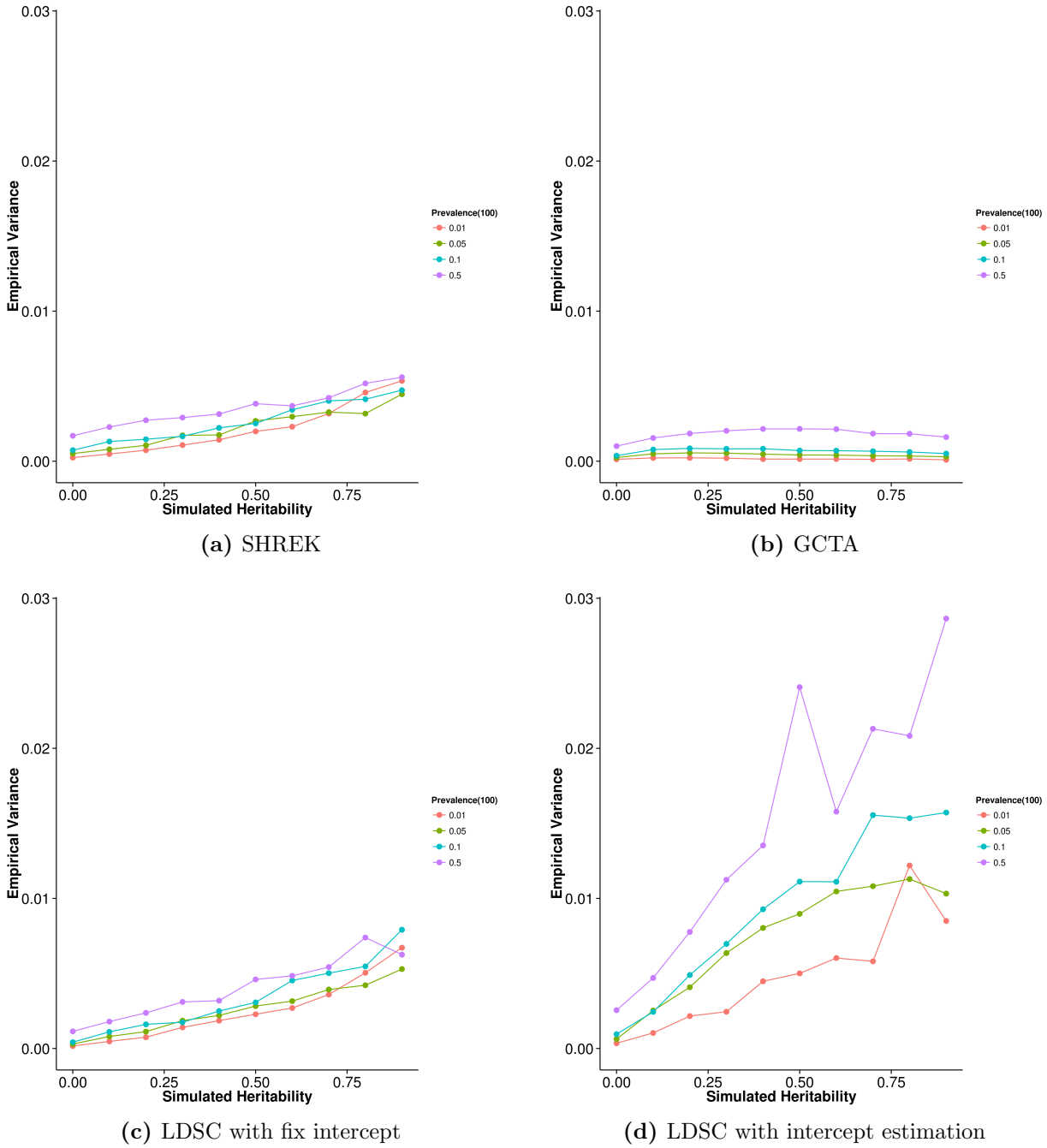


(a) SHREK **(b)** GCTA



(c) LDSC with fix intercept **(d)** LDSC with intercept estimation

Figure S4: Mean of results from case control simulation with random effect size simulation with 100 causal SNPs. The bias seems to be unaffected by the number of causal SNPs and were the same as what was observed when there were 10 or 50 causal SNPs.



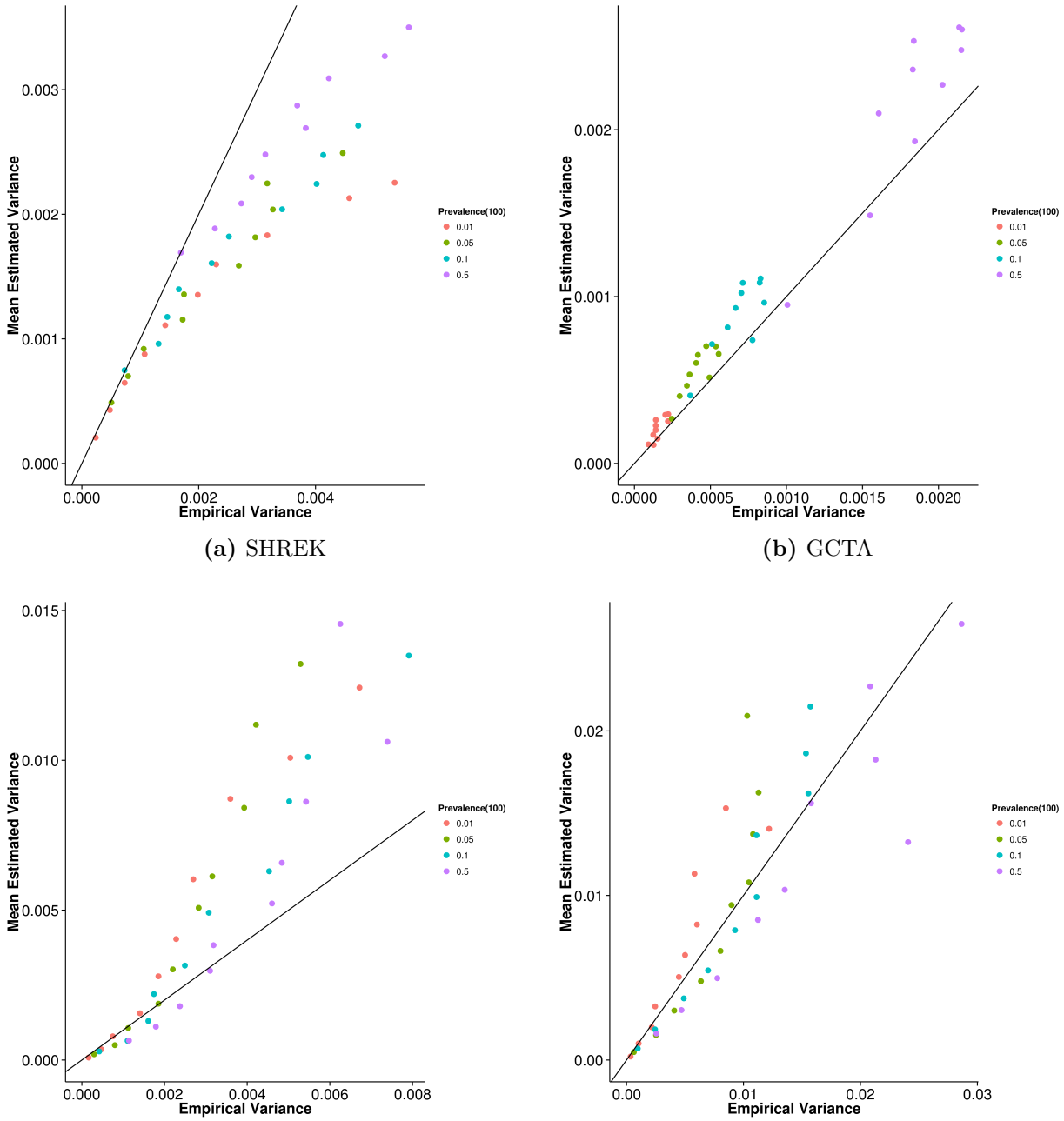
(a) SHREK

(b) GCTA

(c) LDSC with fix intercept

(d) LDSC with intercept estimation

Figure S5: Variance of results from case control simulation with random effect size simulation with 100 causal SNPs. As the number of causal SNPs increased to 100, the relationship between the population prevalence and the empirical variance of the algorithms become clear where as the population prevalence increases, the empirical variance of all algorithm increases. Again, LDSC with intercept estimation has the largest variation of all the algorithms and the empirical variance of LDSC with fix intercept is only slightly higher than that of SHREK.



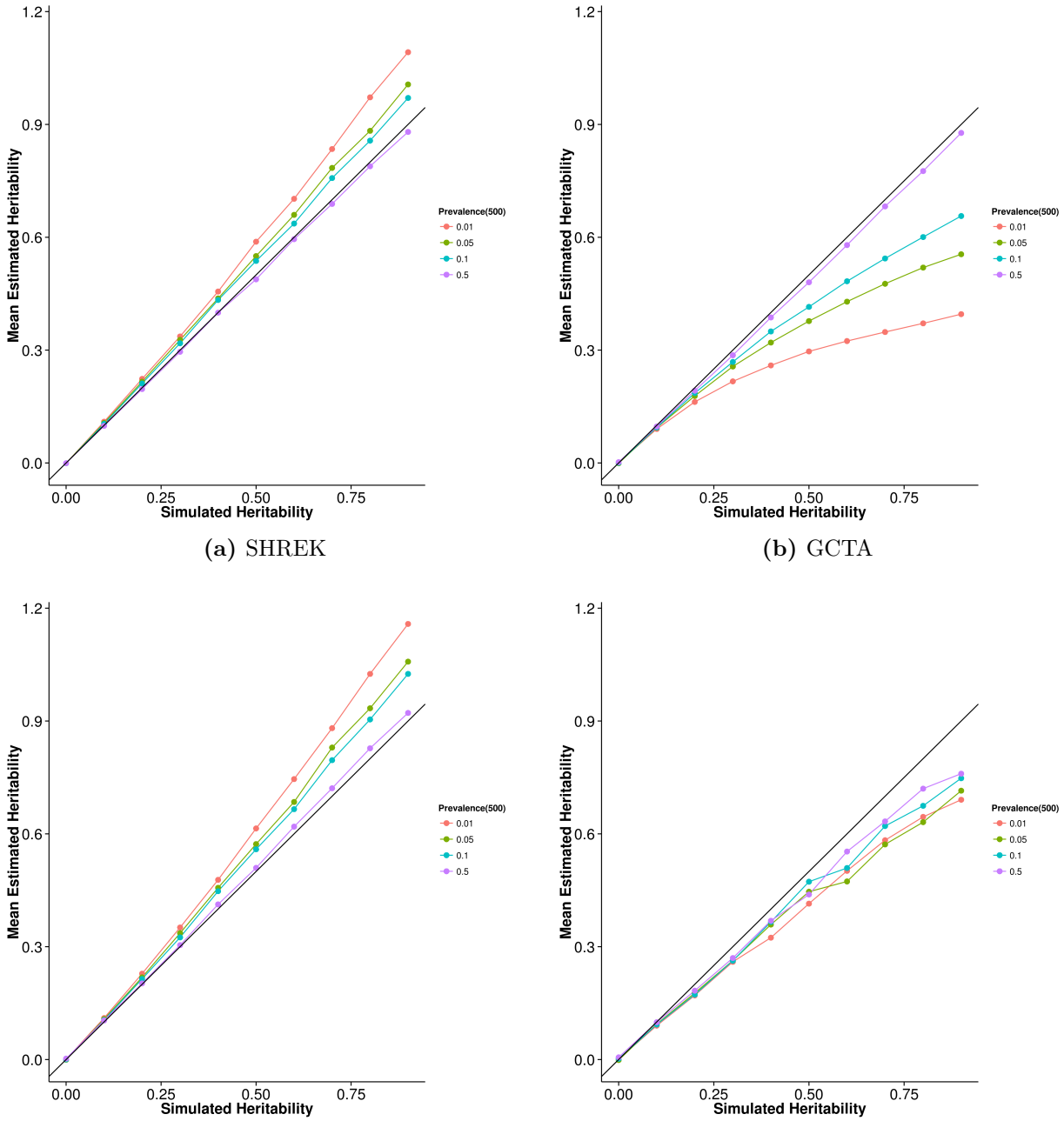
(a) SHREK

(b) GCTA

(c) LDSC with fix intercept

(d) LDSC with intercept estimation

Figure S6: Estimated variance of results from case control simulation with random effect size simulation when compared to empirical variance when 100 causal SNPs was simulated. Once again, SHREK underestimated its empirical variance and LDSC with fixed intercept overestimates its empirical variance. However, the magnitude of overestimation of LDSC with fixed intercept decreased when compared to previous conditions.



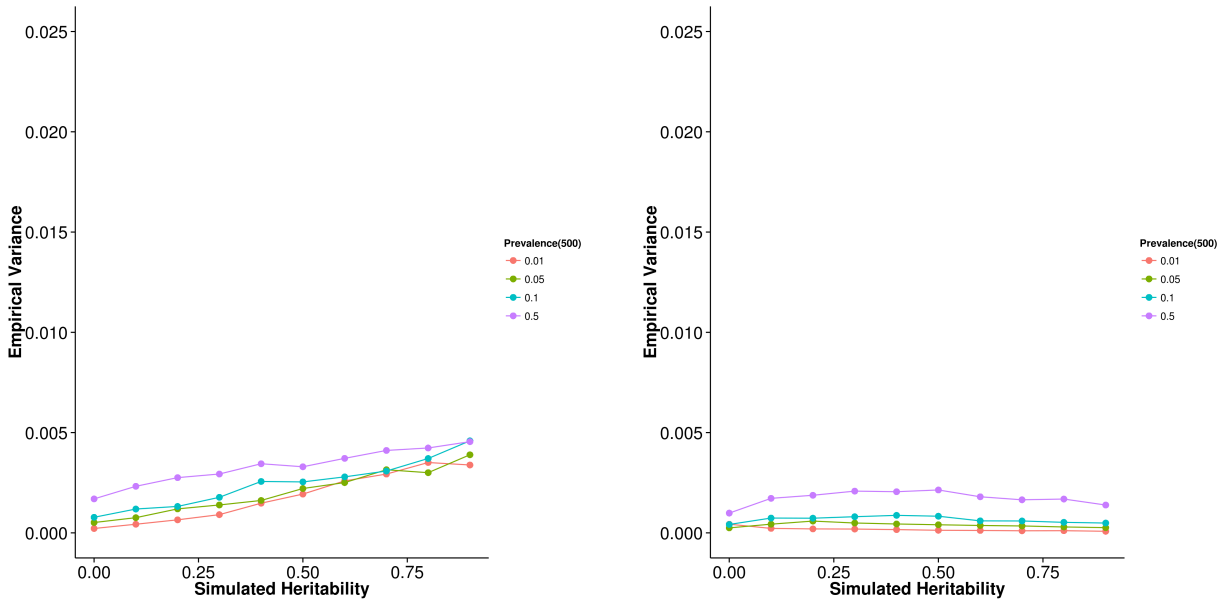
(a) SHREK

(b) GCTA

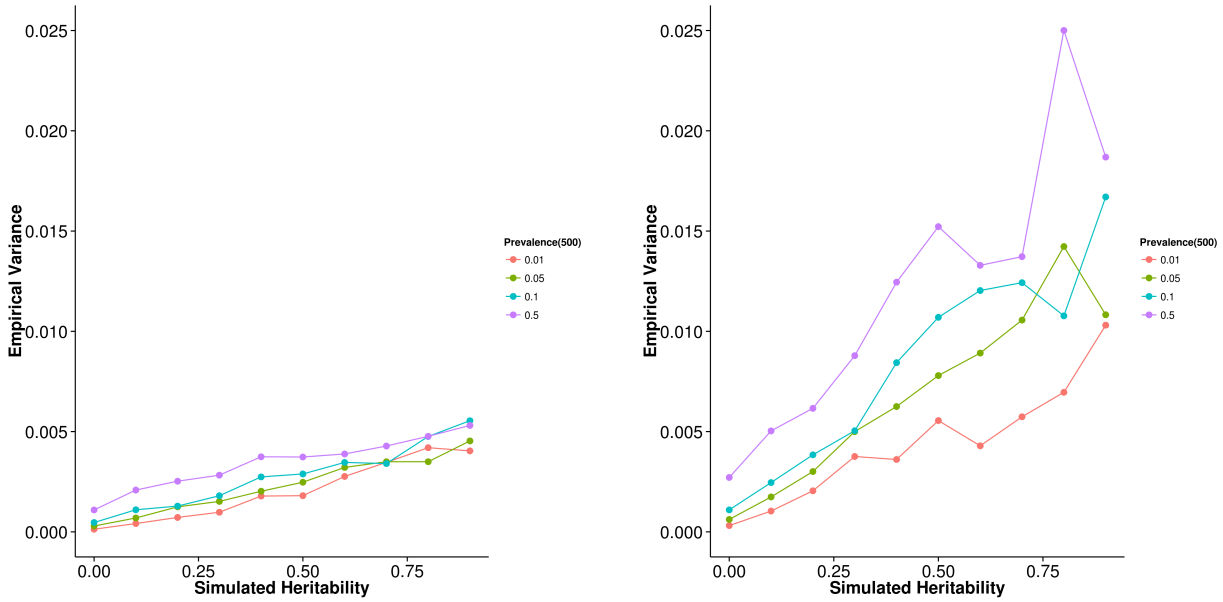
(c) LDSC with fix intercept

(d) LDSC with intercept estimation

Figure S7: Mean of results from case control simulation with random effect size simulation with 500 causal SNPs. Again, a clear pattern of underestimation was observed for GCTA and LDSC with intercept estimation whereas estimations from SHREK and LDSC with fixed intercepts tends to be upwardly biased, with the magnitude of bias increases as the population prevalence decreases.

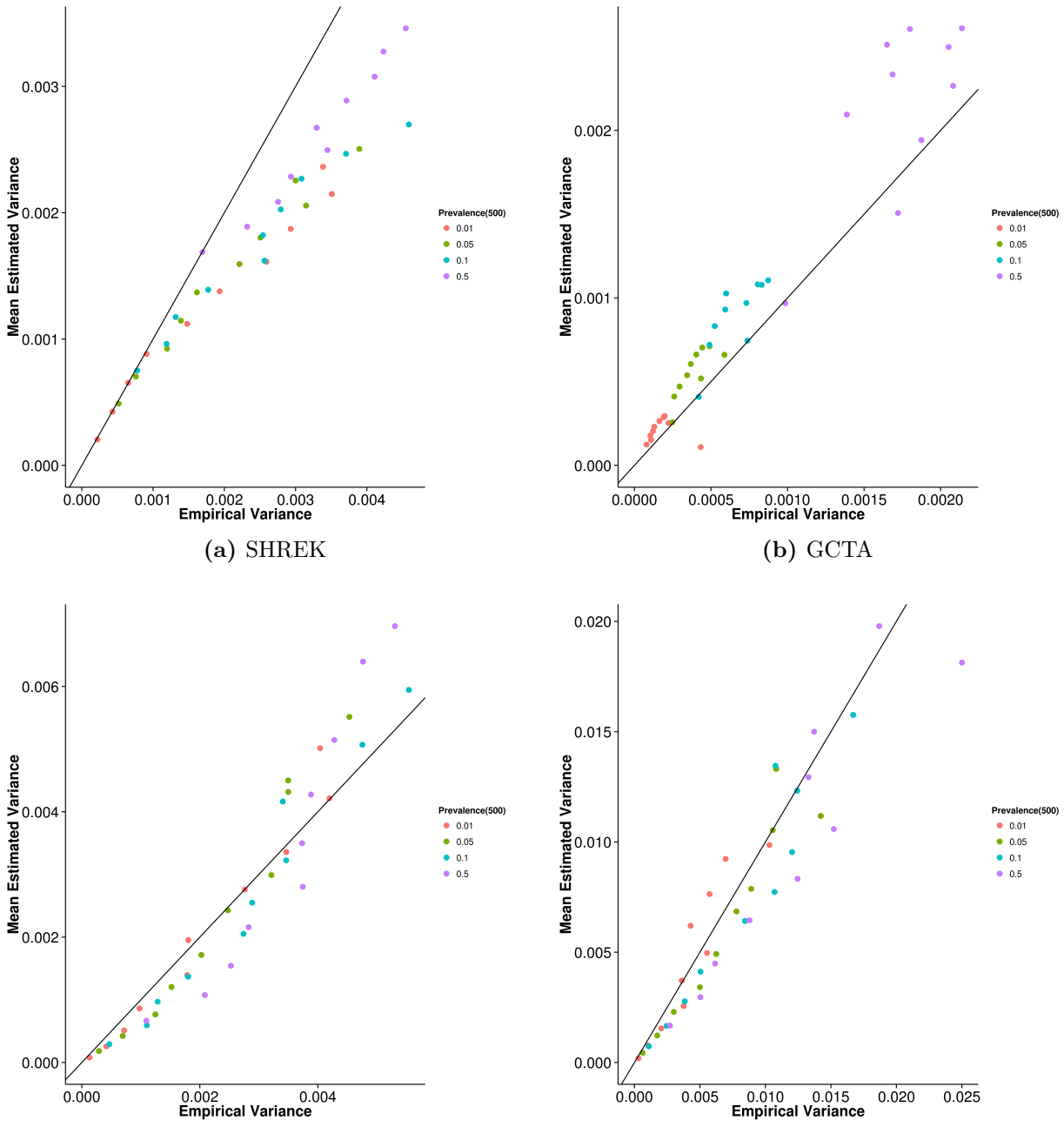


(a) SHREK **(b)** GCTA



(c) LDSC with fix intercept **(d)** LDSC with intercept estimation

Figure S8: Variance of results from case control simulation with random effect size simulation with 500 causal SNPs. As the number of causal SNPs increased to 500, the empirical variance of SHREK and LDSC with fixed intercept converges. However, the empirical variance of LDSC with intercept estimations remains high.



(a) SHREK **(b)** GCTA

(c) LDSC with fix intercept **(d)** LDSC with intercept estimation

Figure S9: Estimated variance of results from case control simulation with random effect size simulation when compared to empirical variance when 500 causal SNPs was simulated. When the trait contains 500 causal SNPs, LDSC begins to provide a good estimation of its own empirical variance both with and without intercept estimation. On the other hand, SHREK's estimation of its own empirical variance remains consistently lower than the true empirical variance.