# 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 10, 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

**CEU** Northern Europeans from Utah. 45, 49, 51

**GCTA** Genome-wide Complex Trait Analysis. 33, 50–52, 54, 56–59, 61, 63, 65, 70–72

**GWAS** Genome Wide Association Study. 34, 35, 47–49

**LD** Linkage Disequilibrium. 34–36, 39, 42, 43, 45–51

**LDSC** LD SCore. 33, 46, 50–52, 54–61, 63, 65–67, 70–72

**maf** Minor Allele Frequency. 49–52

**MSE** mean squared error. 58, 59, 63

**NCP** non-centrality parameter. 41

**PCGC** Phenotype correlation - genotype correlation regression. 33

**PGC** Psychiatric Genomics Consortium. 48

**SE** standard error. 40, 41, 47

**SHREK** SNP Heritability and Risk Estimation Kit. 54–61, 63–66, 68, 70–72

**SNP** Single Nucleotide Polymorphism. 33–37, 39, 45–55, 57–59, 61–69, 71, 72

**SVD** Singular Value Decomposition. 43, 44

**tSVD** Truncated Singular Value Decomposition. 43–45

# 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(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, Lander, and Rosset, 2014). Although Golan, 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. $\epsilon$ 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

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

$\beta^2$ 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, $\beta^2$ 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 $\chi^2$ distribution.

Furthermore, when the effect size is small and $n$ is big, $r^2$ will be approximately $\chi^2$ distributed with mean $\sim 1$. 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 $\text{SNP}_i$ and the phenotype($\hat{\rho}_i$) to be

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

for some error $\epsilon_i$. The distribution of the correlation coefficient about the true correlation $\rho$ is approximately

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

By making the assumption that $\rho_i$ 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 $\chi^2$-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), $\chi^2$ 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 $\chi^2$ 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 $\mu_i$, 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 $\mathrm{SNP}_i$ and $\mathrm{SNP}_j$. 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 $\chi^2$ 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 $\chi^2$ distributed, we might view eq. (2.16) as a decomposition of a vector of $\chi^2$ 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 $\chi^2$ 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, Wray, and 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, Wray, and 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 $\chi^2_{CC}$ 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 (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*$(\sigma_i)$ 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 $\sigma_n$ 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 $\sigma_i$ is small, the error $\epsilon_i$ 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 $\sigma_i$ 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 $\sigma_k$ and $\sigma_{k+1}$. 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 $\sigma_{k+1}$ and subsequent singular values should be truncated, leading to unstable

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 $\epsilon$ 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.

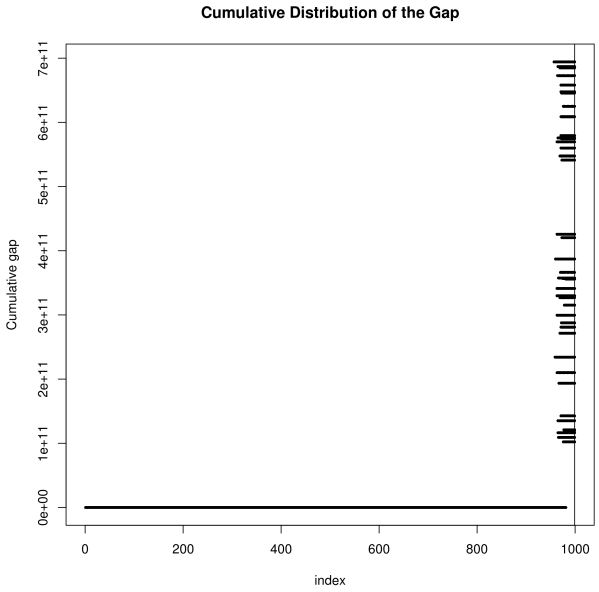
By employing the tSVD as a method for regularization, we were able to solve the ill-posed eq. (2.16), and obtain the estimated heritability.

### 2.2.6 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

**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




"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

$$\text{Var}(\beta) = \frac{h^2}{M}\boldsymbol{I} \tag{2.36}$$

$M =$ number of SNPs

$\beta =$ vector containing per normalized genotype effect sizes

$I =$ identity matrix

$h^2 =$ heritability

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 Hill, 1980; Wang and Thompson, 2007) have proposed different methods for the correction of sample $R^2$. We considered the following LD correction algorithms:

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

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

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

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

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

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 a simple quantitative polygenic trait simulation to compare the performance of our algorithm using different LD bias correction algorithms.

First, 5,000 SNPs with Minor Allele Frequency (maf) $\geq 0.1$ were randomly se-

lected 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 (S. 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. We then simulate 10 different population and repeat the whole process.

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
8. Repeat step 1-7 10 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 (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,

### 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.

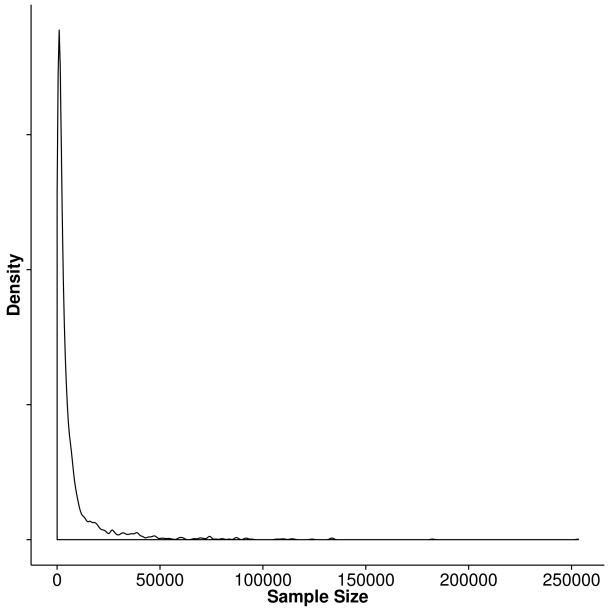



**Figure 2.2:** GWAS sample size distribution.

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 $\geq 0.01$. We therefore only consider scenario with "common" variants and only use SNPs with maf $\geq 0.1$ in the CEU haplotypes as an input to HAPGEN2. This will reduce the probability of having SNPs with maf $< 0.01$ in the final simulated sample sets.

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) similar to that in section 2.3.

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 similar to that in section 2.3. 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 $\geq 0.1$ in the CEU haplotypes as an input to HAPGEN2. We then randomly select 100 SNPs 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 100 effect size following eq. (2.42)
4. Randomly assign the effect size to 100 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 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

## 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`.

| Number of Causal SNPs | SHREK without LD Correction | SHREK with LD Correction |
|---|---|---|
| 5 | 0.0175 | 0.0352 |
| 10 | 0.0163 | 0.0337 |
| 50 | 0.0149 | 0.0305 |
| 100 | 0.0159 | 0.0319 |
| 250 | 0.0146 | 0.0319 |
| 500 | 0.0138 | 0.0293 |

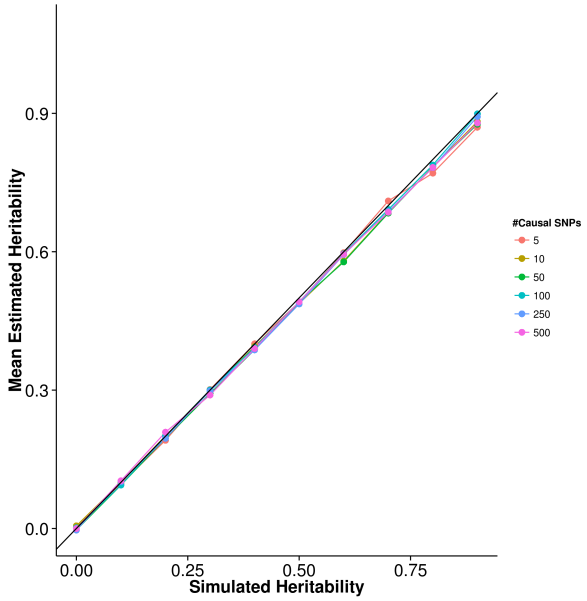
**Table 2.1:** MSE SHREK with and without LD correction. The MSE remains relatively stable for both algorithm with respect to the number of causal SNPs. However, SHREK with LD correction does have almost doubled the MSE when compared to SHREK without LD correction. This suggest that the LD correction might be have a negative impact to the heritability estimation in our simulation.

### 2.7.1 Performance

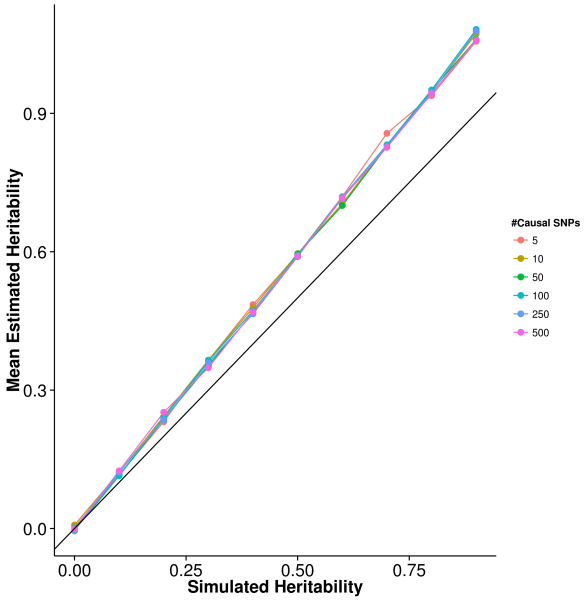
We first examine the performance of SHREK when estimating the narrow sense heritability of varies quantitative traits. Although the number of causal SNPs did not have a significant impact to the performance of SHREK, a general up-ward bias was observed (figure similar to fig. 2.3b). As the simulated heritability increases, the bias of the estimate systematically increases. Consider the systematic nature of the bias, we hypothesize that the bias might be a result of a simulation artifact which produces sample LD that violates our expectation of an inflated sample LD. Another possibility is that there is a general *underestimation* of the LD, therefore leads to an over estimation of heritability.

We re-ran the simulation on **??**, this time comparing the performance of SHREK with and without the LD correction and instead of simulating 50,000 SNPs on chromosome 1, we simulate 5,000 SNPs on chromosome 22 for a faster comparison. It was observed that when SHREK was performed without LD correction (fig. 2.3a), there was a downward bias where the magnitude of bias was much smaller when compared to the estimation with LD correction (fig. 2.3b). On top of that, although the variance of both algorithm was affected by the number of causal SNPs where a smaller amount of causal SNPs leads to higher variance, the variance of estimation was smaller when LD correction was not performed. All in all, the mean squared error (MSE) when LD correction was not performed was almost half of when LD correction was performed (table 2.1).

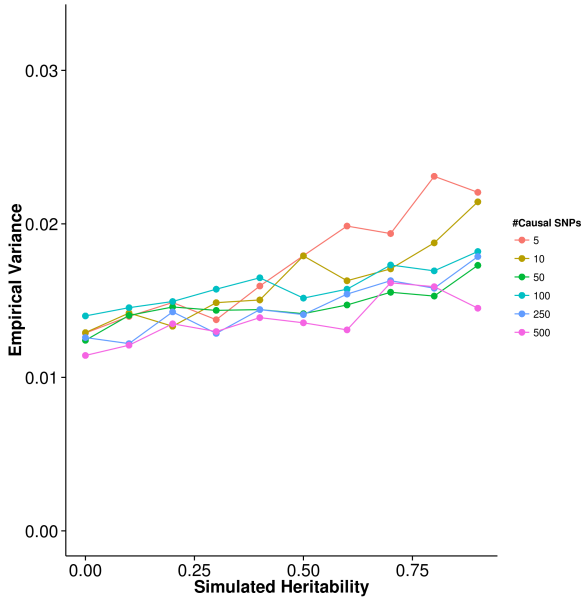
The result does suggest that there might be some problem with the LD, therefore lead to bias in estimation when the LD correction was used. To understand the source of the issue, we first inspect whether if the simulated LD matrix were as expected where the



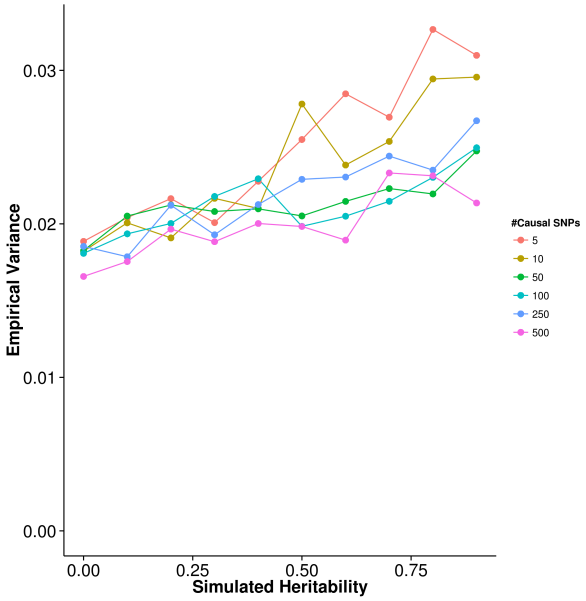

**(a)** Mean of Estimation without LD Correction




**(b)** Mean of Estimation with LD Correction




**(c)** Variance of Estimation without LD Correction




**(d)** Variance of Estimation with LD Correction

**Figure 2.3:** Performance of SHREK with and without LD correction. It was observed that there was an upward bias of SHREK when the LD correction was performed. Not only was the bias in estimation of SHREK lower when LD correction was not performed, the variance was also smaller, suggesting that the LD correction seems to have a negative impact to the heritability estimation.

sample $R^2$ should be more likely to be bigger than the population $R^2$.

When comparing the "population" $R^2$ and the "sample" $R^2$, the "sample" $R^2$ from the simulation were on average larger than "population" $R^2$ (0.00189 larger). On the other hand, when comparing the $R^2$ from the independent samples, the $R^2$ was 0.00180 larger than the "population" $R^2$. Finally, when comparing the "sample" $R^2$ from the $R^2$ of the independent samples, the "sample" $R^2$ was $9.052\times10^{-5}$ larger than the $R^2$ of the independent samples. The difference between the two were very small, suggesting that they were very similar. Also, the bias observed in the "sample" $R^2$ when compared to the "population" $R^2$ also suggest that the simulation does produce the expected results, ruling out the possibility of simulation artifact.

In conclusion, SHREK without LD correction seems to be superior in performance when compared to SHREK with LD correction. The default behavior of SHREK now is to avoid performing the LD correction, with the option to allow user to enable LD correction when it was required.

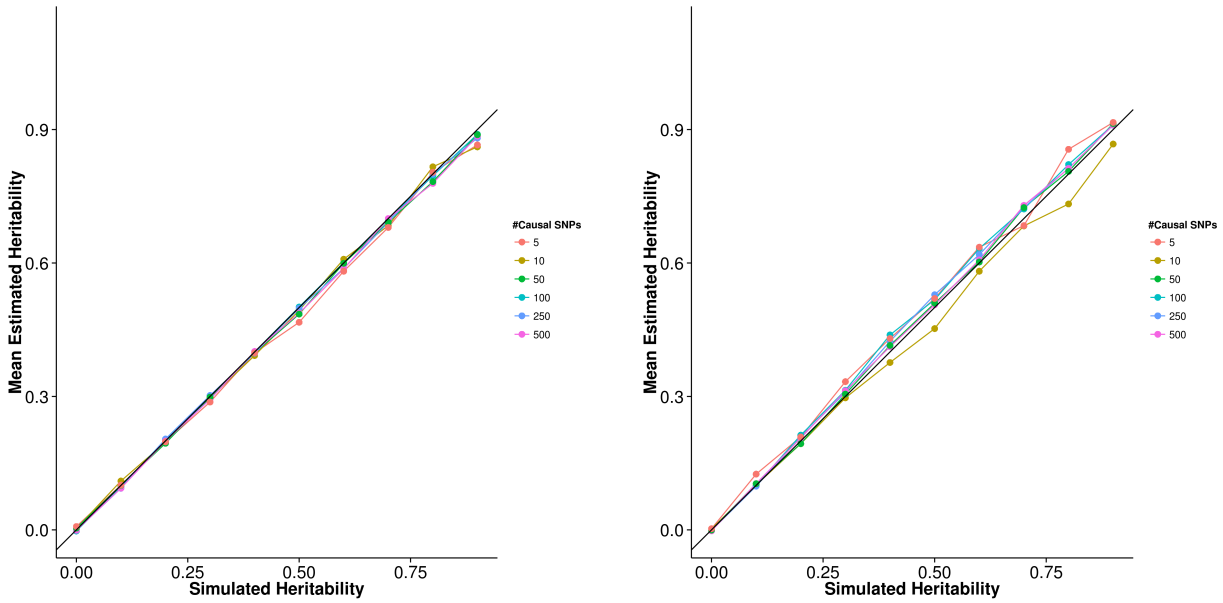
### 2.7.2 Comparing with Other Algorithms

It is important for us to compare our algorithm with existing algorithms to understand the relative performance of the 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.

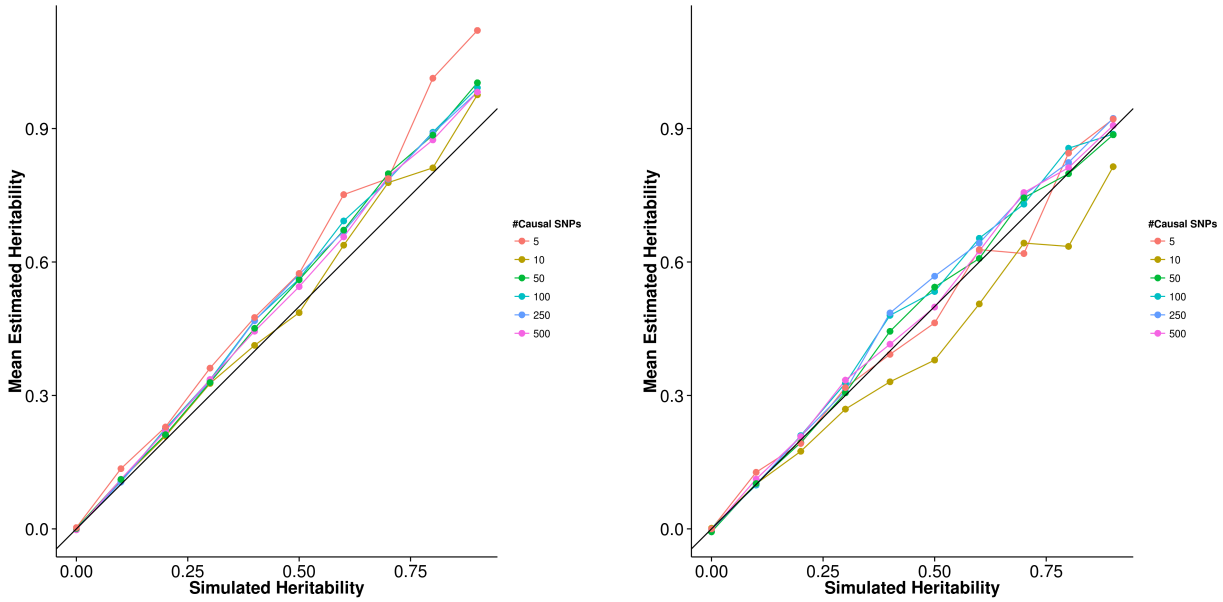
#### Quantitative Trait Simulation

In the simulation of quantitative trait scenario, the effect size were randomly drawn from the exponential distribution with $\lambda = 1$. Under this scenario, we would like the assess the performance of the algorithms when the heritability of the trait and the underlaying genetic architecture of the trait changes.

First, we would like to examine the mean heritability estimation when compared to the simulated heritability. From the graph (fig. 2.4), it was observed that the mean estimations of SHREK were very close to the simulated heritability (fig. 2.4a), with a slight downward bias. Moreover, the bias was insensitive to the change in number of causal SNPs suggesting that SHREK might be relatively robust to trait complexity. Similarly, estimations form GCTA were also accurate and were only moderately biased upward (fig. 2.4b).

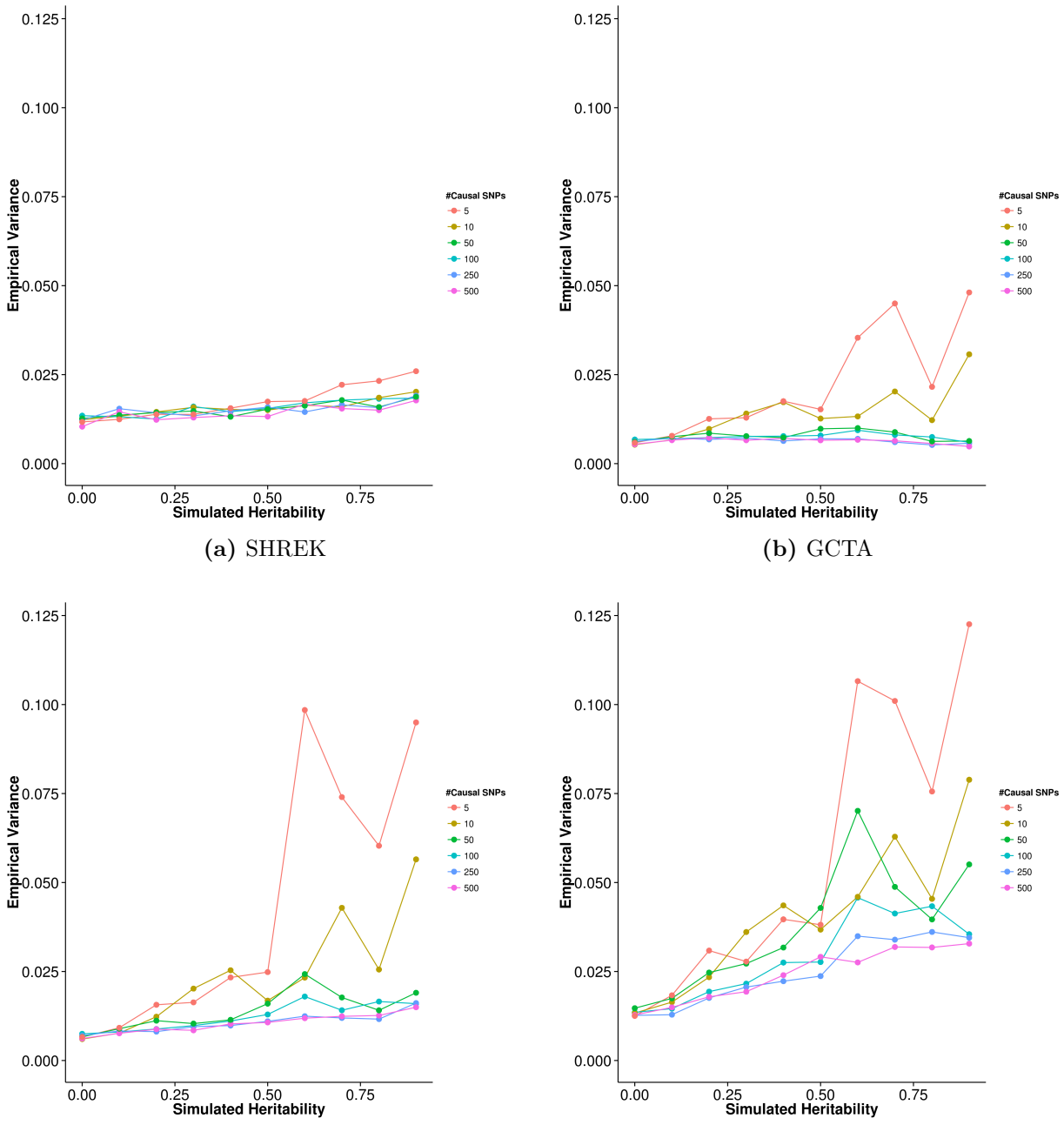



**(a)** SHREK **(b)** GCTA




**(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. SHREK was observed to be slightly biased downward. On the other hand, GCTA were biased upward except in the condition of 10 causal SNPs. Whereas a relatively higher upward bias was observed for LDSC.



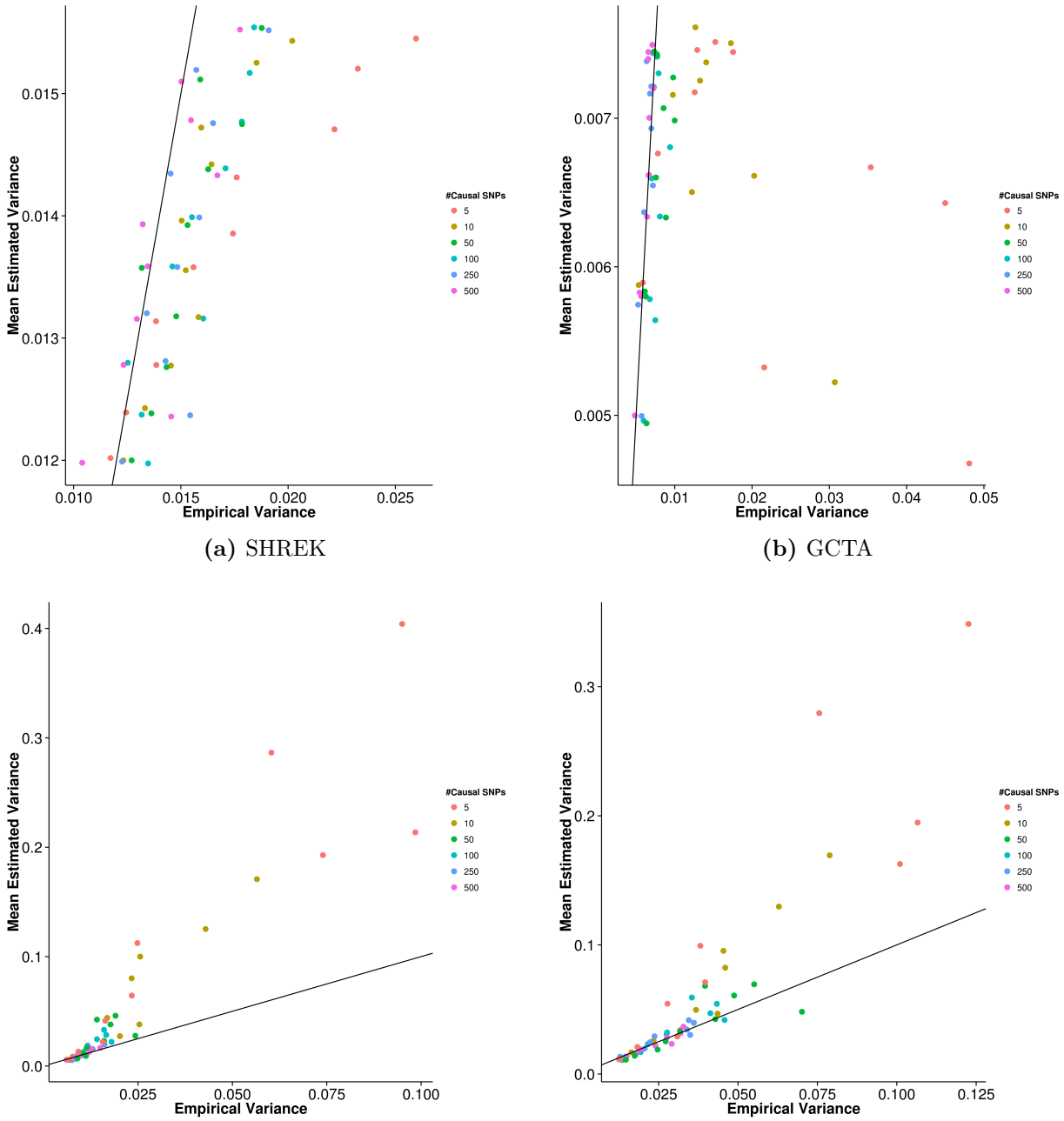

**(a)** SHREK

**(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. 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.




**(a)** SHREK

**(b)** GCTA

**(c)** LDSC with fix intercept

**(d)** LDSC with intercept estimation

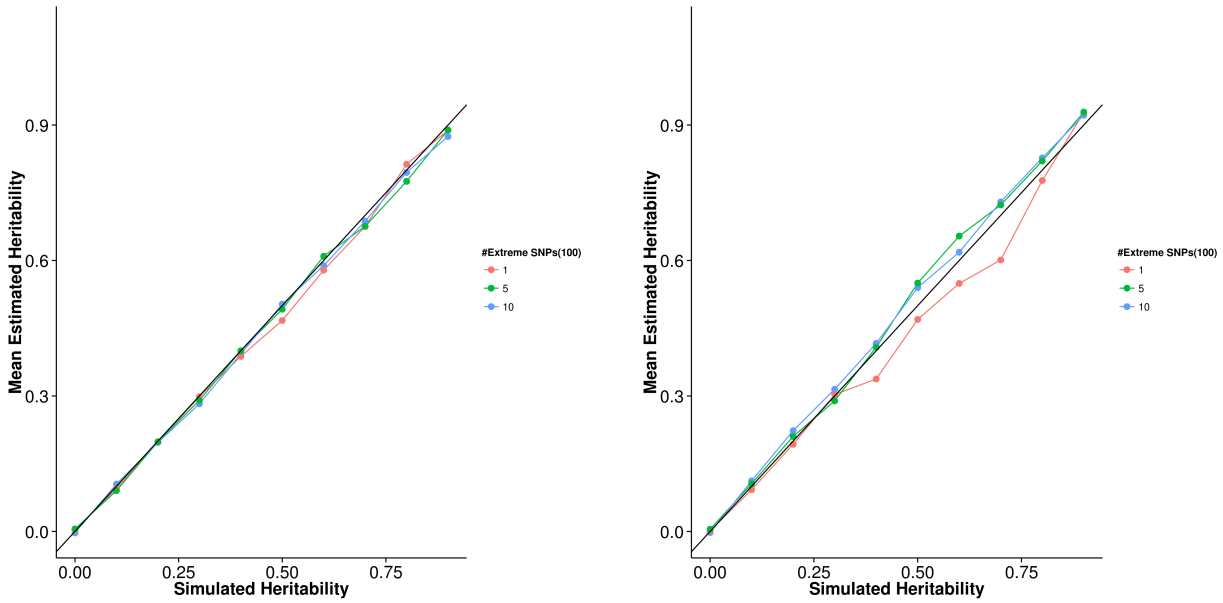
**Figure 2.6:** Estimated variance of results from quantitative trait simulation with random effect size simulation when compared to the empirical variance. GCTA has the best estimate of its empirical variance whereas SHREK tends to under-estimate its empirical variance. On the other hand, LDSC to over-estimate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

On the other hand, the bias observed in the estimations from LDSC were relatively large when compared to SHREK and GCTA. No matter if the intercept was estimated or fixed, estimations from LDSC were upwardly biased and increases as the trait heritability increases (figs. 2.4c and 2.4d). The only exception was when there were only 5 causal SNPs in which a large downward bias was detected when the intercept estimations function was used.

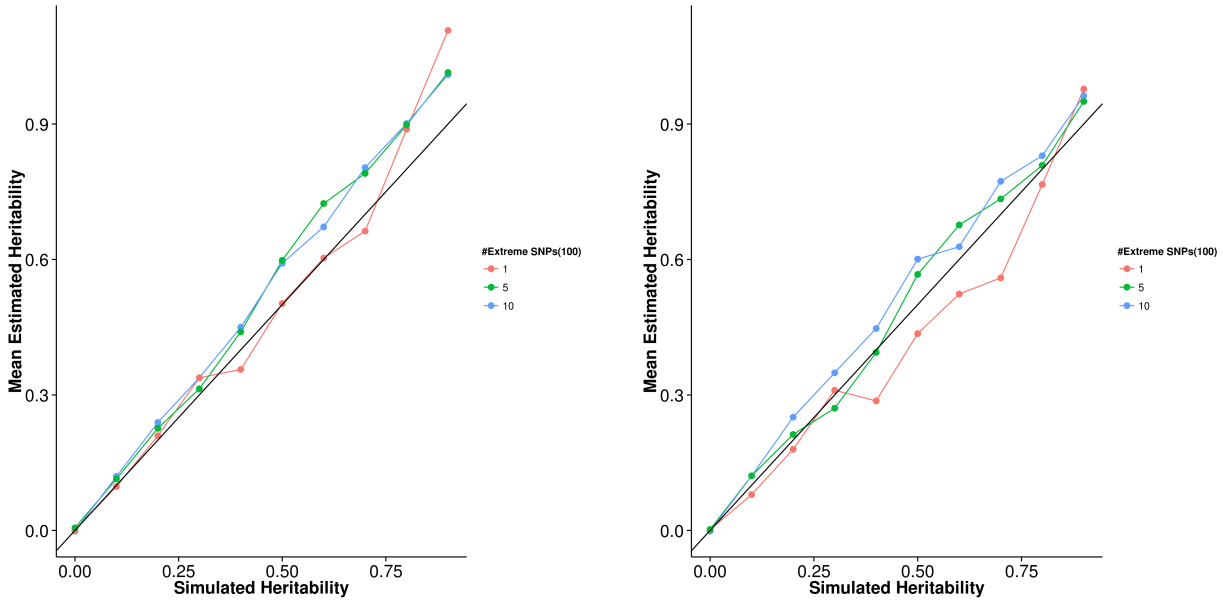
Furthermore, while comparing the empirical variance of the estimates (fig. 2.5), variance of estimations from LDSC were sensitive to the number of causal SNPs where as the number of causal SNPs decreases (figs. 2.5c and 2.5d), the variance increases, similar to what was reported by Bulik-Sullivan et 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. $\geq 0.8$).

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 SHREK tends to under-estimate the variance ($\sim 0.9$ fold) and its estimates were slightly affected by the number of causal SNPs. On the other hand, LDSC cannot accurately estimate its variance especially when the number of causal SNPs were small. When intercept estimations was performed (fig. 2.6d), the estimation of variance was relatively better ($\sim 1.25$ fold) when compared to the case where fixed intercept was used ($\sim 1.65$ fold) instead (fig. 2.6c).

Overall, by taking into consideration of both the bias and variance of the estimates, GCTA has the best performance, follow by SHREK. The LDSC with fixed intercept performs better than LDSC with intercept estimation except in the case of variance estimation. An interesting property of SHREK is that it is insensitive to number of causal SNP of the trait, making it idea for the estimation of narrow sense heritability when one is uncertain of the genetic architecture of the trait and when the sample genotypes are unavailable.

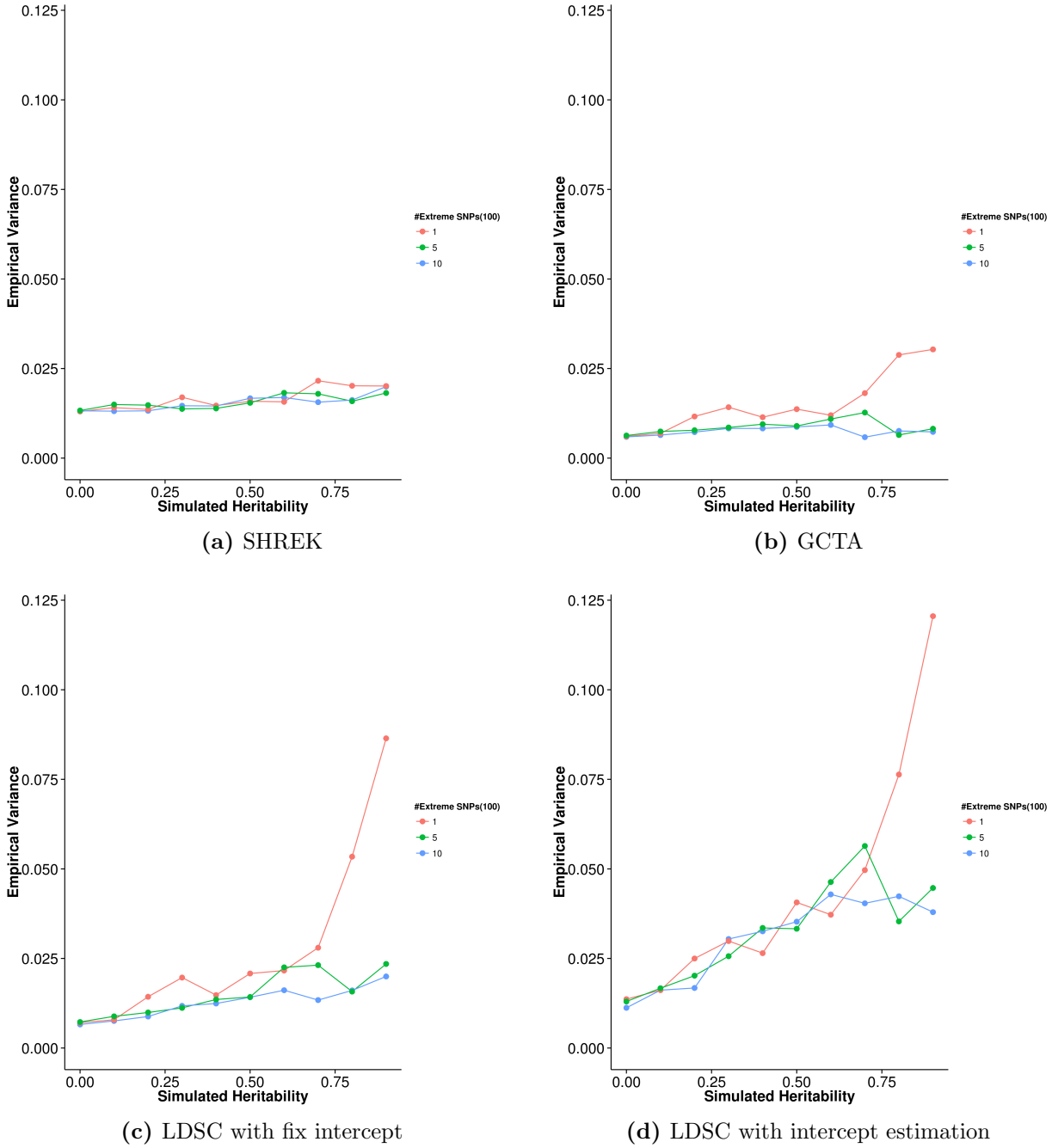



**(a)** SHREK **(b)** GCTA




**(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.



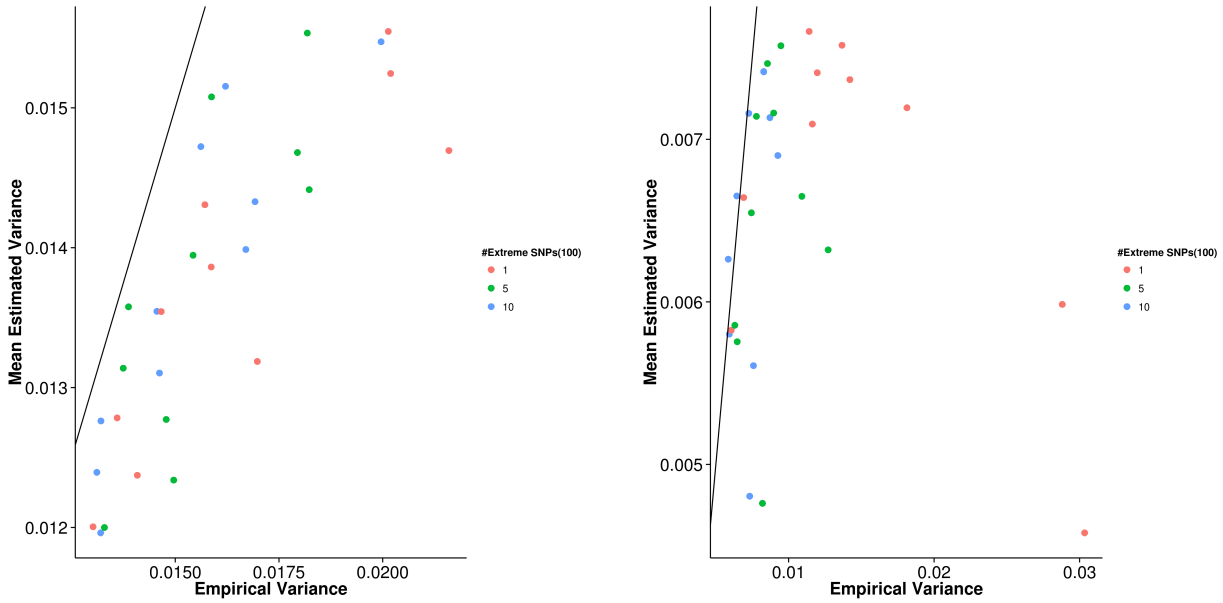

**(a)** SHREK

**(b)** GCTA

**(c)** LDSC with fix intercept

**(d)** LDSC with intercept estimation

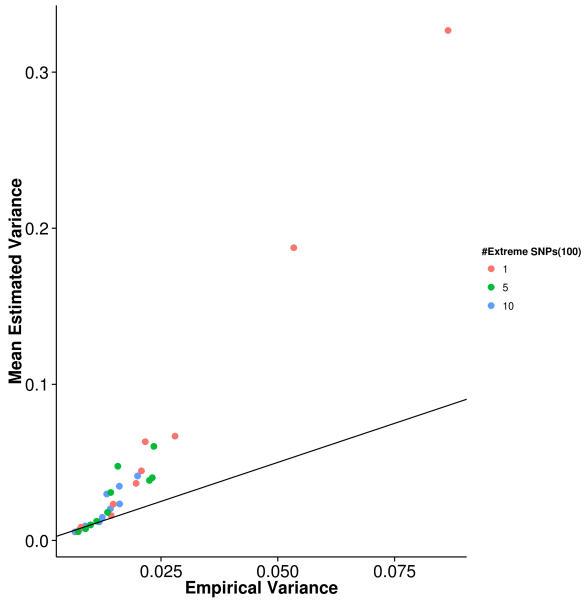
**Figure 2.8:** Variance of results from quantitative trait simulation with extreme effect size simulation. 100 causal SNPs were simulated. GCTA has the smallest variance as with previous simulation. When compared to LDSC with fixed intercept, although the variance of SHREK was relatively higher, it 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 $\geq 0.2$.



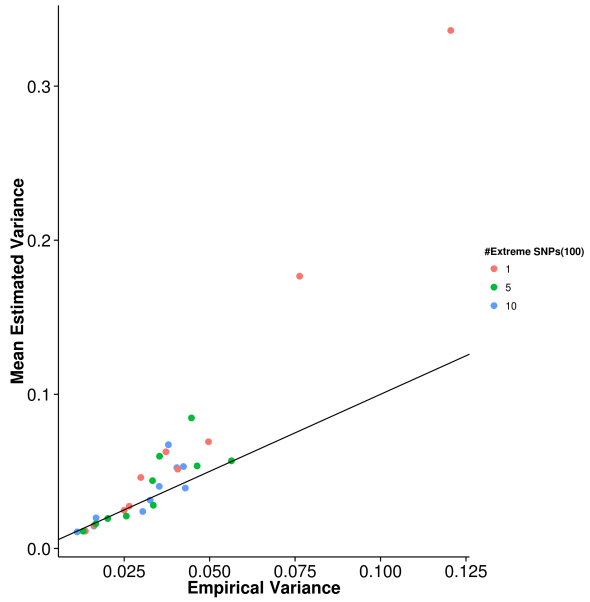

**(a)** SHREK




**(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 |
|---|---|---|---|---|
| 5 | 0.177 | 0.565 | 0.584 | 0.230 |
| 10 | 0.159 | 0.251 | 0.470 | 0.151 |
| 50 | 0.153 | 0.179 | 0.378 | 0.0796 |
| 100 | 0.157 | 0.166 | 0.305 | 0.0794 |
| 250 | 0.152 | 0.144 | 0.266 | 0.0674 |
| 500 | 0.143 | 0.134 | 0.247 | 0.0646 |

**Table 2.2:** 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 and the performance of SHREK and LDSC with fix intercept converges as number of causal SNPs increases. LDSC with fix intercept even surpassed SHREK's performance when the number of causal SNPs was as high as 500.

### 2.7.3 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.

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 2.3). 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

| 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 2.3:** 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.

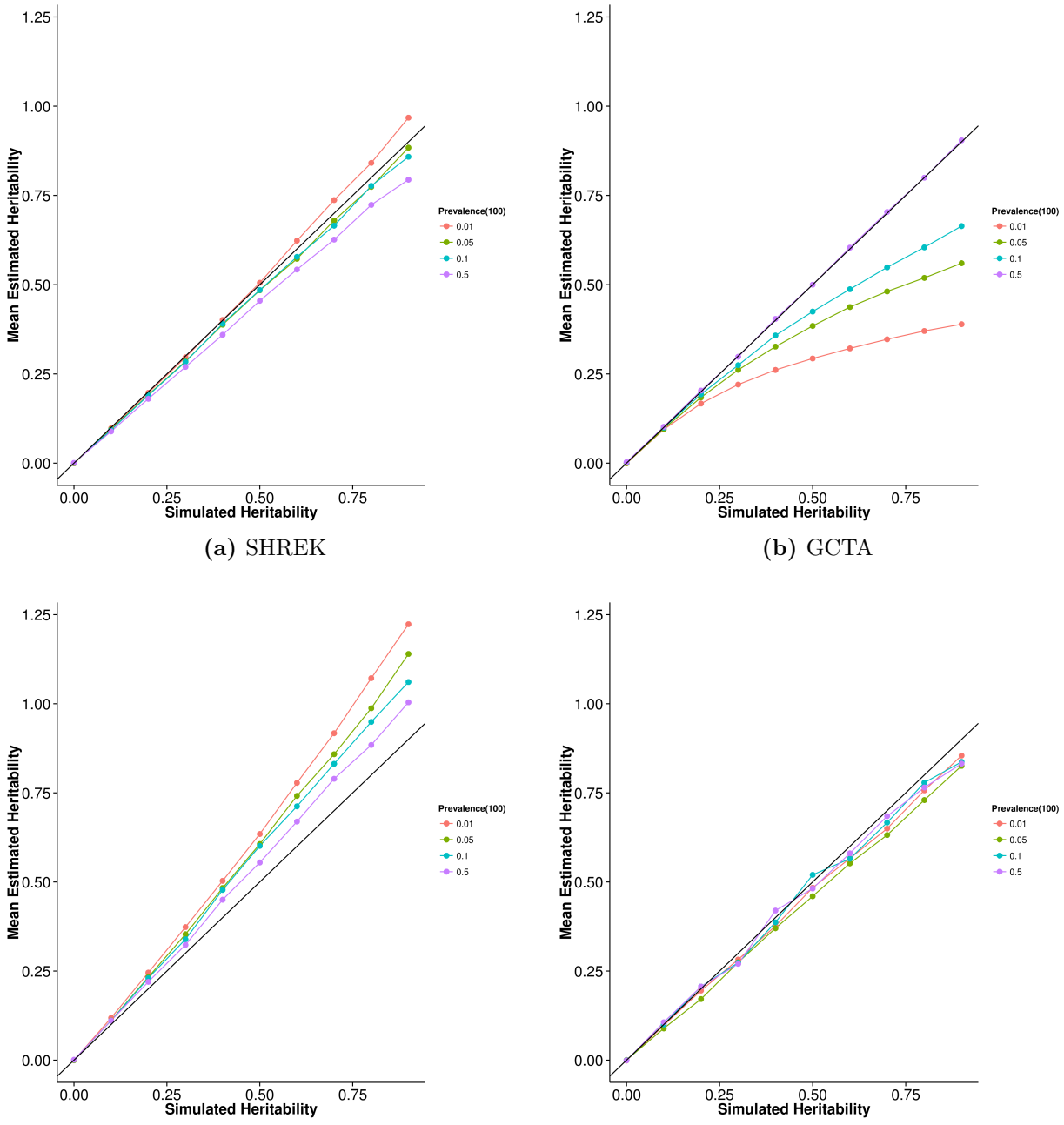
performance with fixed intercept is better than when the intercept estimation function was used.

### 2.7.4 Case Control Simulation

Finally, for the case control simulation, we simulated

### 2.7.5 Extreme Phenotype Simulation

## 2.8 Discussion



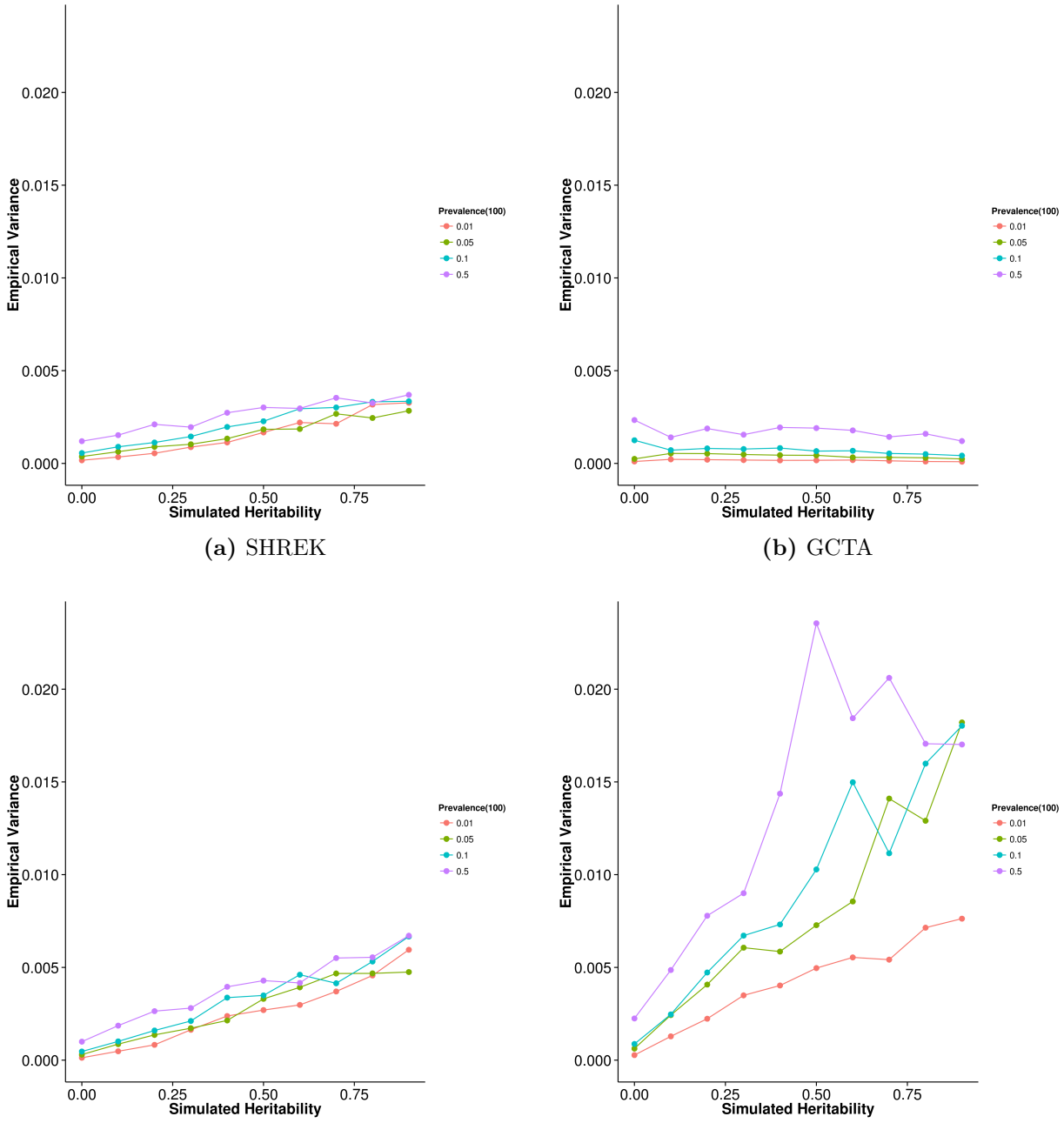

**(a)** SHREK

**(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. The performance of GCTA was as suggested by Golan, Lander, and Rosset (2014) where there was an underestimation as prevalence decreases. On the other hand, LDSC were upwardly biased when a fixed intercept was used and this bias was corrected when an estimation of intercept was allowed. SHREK does not seems to as sensitive to change in prevalence and the estimation were relatively robust.



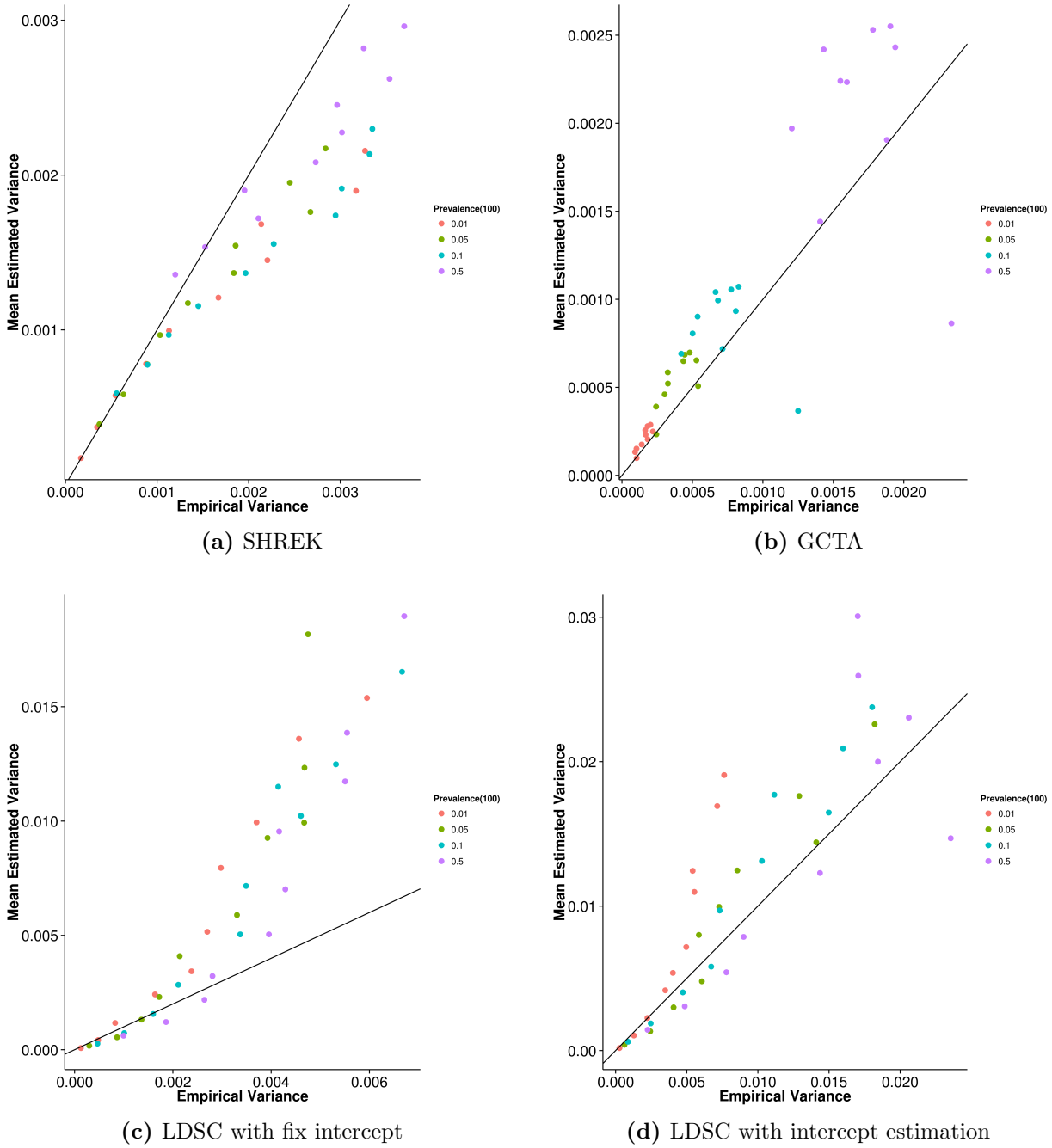

**(a)** SHREK

**(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. It was clear that the prevalence affects the variance of estimation where a larger variance tends to increase the variance of estimation. Again, GCTA has the lowest variance, however, unlike in the quantitative trait simulation, SHREK has a lower average variance when compared to 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. From the quantitative trait simulation with random effect size(fig. 2.6), it was observed that the variance estimation of SHREK and GCTA were rater accurate. Similarly, in the case control simulation with 100 causal SNPs, it was observed that the variance estimation of SHREK and GCTA were close to the empirical variance with slight bias. A large up-ward bias was observed for LDSC with fixed intercept estimation but the bias was less when LDSC was allowed to estimate the intercept.s

# Chapter 5

# Conclusion

# Bibliography


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, 47).

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, 49, 62).

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, 68).

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

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. 52).

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).

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. 48).

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).

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. 48).

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 p. 47).

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. 48).

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, 55).

Shieh, G (2010). "Estimation of the simple correlation coefficient". eng. In: *Behav Res Methods* 42.4, pp. 906–917. DOI: 10.3758/BRM.42.4.90642/4/906[pii].

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. 48, 51).

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. 47).

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. 47).

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. 50).

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. 49).

# Supplementary Materials

# Appendix