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 2, 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. 27, 29, 31

CI confidence interval. 10

DSM Diagnostic and Statistical Manual of Mental Disorders. 2

DZ dizygotic. 9, 10

GCTA Genome-wide Complex Trait Analysis. 27, 31–33

GD Gestation Day. 48

GO Gene Ontology. 50, 54, 55

GWAS Genome Wide Association Study. 12, 15, 16, 28, 29

IL-6 Interleukin-6. 3

IQ intelligence quotient. 5

LD Linkage Disequilibrium. 12, 17, 21, 22, 24–27, 29, 31

LDSC LD SCore. 15, 27, 31–33

LPS lipopolysaccharide. 3

maf Minor Allele Frequency. 5, 12, 30–33

MAGMA Multi-marker Analysis of GenoMic Annotation. 50, 55

MIA maternal immune activation. 3

MZ monozygotic. 9–11

NCP non-centrality parameter. 22, 23

NGS next generation sequencing. 28

PC Principle Component. 50, 52

PGC Psychiatric Genomics Consortium. 50, 55

PolyI:C polyriboinosinic-polyribocytidilic acid. 3

RIN RNA integrity number. 48

RPKM Reads Per Kilobase per Million mapped reads. 48, 49, 51, 53

SCZ schizophrenia. 30, 32, 47

SE standard error. 22

SHREK SNP Heritability and Risk Estimation Kit. 15, 27, 31–35

SNP Single Nucleotide Polymorphism. 12, 16–18, 21, 26–33, 50

SVD Singular Value Decomposition. 25

tSVD Truncated Singular Value Decomposition. 25–27

WGCNA Weighted Gene Co-expression Network Analysis. 49, 50

WHO World Health Organization. 1, 2

YLD years lost due to disability. 1, 2

Contents

List of Figures

List of Tables

Chapter 1

Heritability Estimation

1.1 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.

1.2 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.

1.2.1 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

$$Y = \beta X + \epsilon \tag{1.1}$$

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. (1.1), one can then have

$$\begin{aligned} \mathrm{Var}(Y) &= \mathrm{Var}(\beta X) + \mathrm{Var}(\epsilon) \\ \mathrm{Var}(Y) &= \beta^{\mathrm{Var}}(X) \\ \beta^2 \frac{\mathrm{Var}(X)}{\mathrm{Var}(Y)} &= 1 \end{aligned} \tag{1.2}$$

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. (1.2). 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:

$$r = \frac{t}{\sqrt{n-2+t^2}} \tag{1.3}$$

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. (1.3)

$$r^2 = \frac{t^2}{n-2+t^2} \tag{1.4}$$

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. (1.4) as

$$r^2 = \frac{\chi^2}{n} \tag{1.5}$$

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

$$f = \frac{\chi^2 - 1}{n} \tag{1.6}$$

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:

$$f_{observed} = f_{true} + f_{LD} \tag{1.7}$$

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. (1.1) in matrix form:

$$\boldsymbol{Y} = \boldsymbol{\beta}^t \boldsymbol{X} + \epsilon \tag{1.8}$$

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

$$\begin{aligned} Heritability &= \frac{\mathrm{Var}(\boldsymbol{\beta}^t \boldsymbol{X})}{\mathrm{Var}(\boldsymbol{Y})} \\ &= \mathrm{Var}(\boldsymbol{\beta}^t \boldsymbol{X}) \end{aligned} \tag{1.9}$$

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

$$\begin{aligned} \mathrm{Var}(\boldsymbol{\beta}^t \boldsymbol{X}) &= \mathrm{E}_X \mathrm{Var}_{\beta|X}(\boldsymbol{X}^t \boldsymbol{\beta}) + \mathrm{Var}_X \mathrm{E}_{(\beta|X)}(\boldsymbol{\beta}^2 \boldsymbol{X}) \\ &= \mathrm{E}_X(\boldsymbol{X}^t \boldsymbol{\beta} \boldsymbol{\beta}^T \boldsymbol{X}) \\ &= \mathrm{E}_X(\boldsymbol{X}^t \boldsymbol{H} \boldsymbol{X}) \\ &= \mathrm{E}(\boldsymbol{X})^t \boldsymbol{H} \mathrm{E}(\boldsymbol{X}) + \mathrm{Tr}(\mathrm{Var}(\boldsymbol{X}\boldsymbol{H})) \\ &= \mathrm{Tr}(\mathrm{Var}(\boldsymbol{X}\boldsymbol{H})) \\ &= \sum_i h_i^2 \end{aligned} \tag{1.10}$$

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

$$\begin{aligned} \mathrm{Cov}(\boldsymbol{X}_i, \boldsymbol{Y}) &= \mathrm{Cov}(\boldsymbol{X}_i, \boldsymbol{\beta}^t \boldsymbol{X} + \epsilon) \\ &= \mathrm{Cov}(\boldsymbol{X}_i, \boldsymbol{\beta}^t \boldsymbol{X}) \\ &= \sum_j \mathrm{Cov}(\boldsymbol{X}_i, \boldsymbol{X}_j) \boldsymbol{\beta}_j \\ &= \boldsymbol{R}_i \boldsymbol{\beta}_j \end{aligned} \tag{1.11}$$

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

$$\rho_i = \boldsymbol{R}_i \boldsymbol{\beta}_j \tag{1.12}$$

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

$$\hat{\rho}_i = \rho_i + \frac{\epsilon_i}{\sqrt{n}} \tag{1.13}$$

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. (1.13) and eq. (1.12), 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

$$f_i = \frac{\chi_i^2 - 1}{n} \tag{1.14}$$

we got

$$\begin{aligned}\sum_i(\chi_i^2 - \mathrm{E}(\chi_i^2))^2 &= \sum_i(f_i - \boldsymbol{R}_i\boldsymbol{H}\boldsymbol{R}_i)^2 \\ &= \boldsymbol{f}\boldsymbol{f}^t - 2\boldsymbol{f}^t\boldsymbol{R}_{\boldsymbol{sq}}\hat{\boldsymbol{h}} + \hat{\boldsymbol{h}}^t\boldsymbol{R}_{\boldsymbol{sq}}^t\boldsymbol{R}_{\boldsymbol{sq}}\hat{\boldsymbol{h}}\end{aligned} \tag{1.15}$$

where $\boldsymbol{R_{sq}} = \boldsymbol{R} \circ \boldsymbol{R}$. By differentiating eq. (1.15) w.r.t $\hat{h}$ and set to 0, we get

$$\begin{aligned}2\boldsymbol{R}_{\boldsymbol{sq}}^t\boldsymbol{R}_{\boldsymbol{sq}}\hat{\boldsymbol{h^2}} - 2\boldsymbol{R}_{\boldsymbol{sq}}\boldsymbol{f} &= 0 \\ \boldsymbol{R}_{\boldsymbol{sq}}\hat{\boldsymbol{h^2}} &= \boldsymbol{f}\end{aligned} \tag{1.16}$$

And the heritability is then defined as

$$\hat{Heritability} = \mathbf{1}^t\boldsymbol{R}_{\boldsymbol{sq}}^{-1}\boldsymbol{f} \tag{1.17}$$

1.2.2 Calculating the Standard error

From eq. (1.17), we can derive the variance of heritability H as

$$\begin{aligned}\mathrm{Var}(H) &= \mathrm{E}[H^2] - \mathrm{E}[H]^2 \\ &= \mathrm{E}[(\mathbf{1}^t\boldsymbol{R}_{\boldsymbol{sq}}^{-1}\boldsymbol{f})^2] - \mathrm{E}[\mathbf{1}^t\boldsymbol{R}_{\boldsymbol{sq}}^{-1}\boldsymbol{f}](\mathrm{E}[\mathbf{1}^t\boldsymbol{R}_{\boldsymbol{sq}}^{-1}\boldsymbol{f}])^t \\ &= \mathrm{E}[\mathbf{1}^t\boldsymbol{R}_{\boldsymbol{sq}}^{-1}\boldsymbol{f}\boldsymbol{f}^t\boldsymbol{R}_{\boldsymbol{sq}}^{-1}\mathbf{1}] - \mathrm{E}[\mathbf{1}^t\boldsymbol{R}_{\boldsymbol{sq}}^{-1}\boldsymbol{f}](\mathrm{E}[\mathbf{1}^t\boldsymbol{R}_{\boldsymbol{sq}}^{-1}\boldsymbol{f}])^t \\ &= \mathbf{1}^t\boldsymbol{R}_{\boldsymbol{sq}}^{-1}\mathrm{E}[\boldsymbol{f}\boldsymbol{f}^t]\boldsymbol{R}_{\boldsymbol{sq}}^{-1}\mathbf{1} - \mathrm{E}[\mathbf{1}^t\boldsymbol{R}_{\boldsymbol{sq}}^{-1}\boldsymbol{f}](\mathrm{E}[\mathbf{1}^t\boldsymbol{R}_{\boldsymbol{sq}}^{-1}\boldsymbol{f}])^t \\ &= \mathbf{1}^t\boldsymbol{R}_{\boldsymbol{sq}}^{-1}\mathrm{Var}(\boldsymbol{f})\boldsymbol{R}_{\boldsymbol{sq}}^{-1}\mathbf{1} + \mathrm{E}[\mathbf{1}^t\boldsymbol{R}_{\boldsymbol{sq}}^{-1}\boldsymbol{f}](\mathrm{E}[\mathbf{1}^t\boldsymbol{R}_{\boldsymbol{sq}}^{-1}\boldsymbol{f}])^t - \mathrm{E}[\mathbf{1}^t\boldsymbol{R}_{\boldsymbol{sq}}^{-1}\boldsymbol{f}](\mathrm{E}[\mathbf{1}^t\boldsymbol{R}_{\boldsymbol{sq}}^{-1}\boldsymbol{f}])^t \\ &= \mathbf{1}^t\boldsymbol{R}_{\boldsymbol{sq}}^{-1}\mathrm{Var}(\boldsymbol{f})\boldsymbol{R}_{\boldsymbol{sq}}^{-1}\mathbf{1}\end{aligned} \tag{1.18}$$

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

$$\mathrm{Cov}(X_i^2, X_j^2) = 2R_{ij}^2 + 4R_{ij}\mu_i\mu_j \tag{1.19}$$

As we only have the *observed* expectation, we should re-define eq. (1.19) as

$$\mathrm{Cov}(X_i^2, X_j^2) = \frac{2R_{ij}^2 + 4R_{ij}\mu_i\mu_j}{n^2} \tag{1.20}$$

where n is the sample size.

By substituting eq. (1.20) into eq. (1.18), we will get

$$\mathrm{Var}(H) = \mathbf{1}^t \boldsymbol{R}_{\boldsymbol{sq}}^{-1} \frac{2\boldsymbol{R}_{\boldsymbol{sq}} + 4\boldsymbol{R} \circ \boldsymbol{z}\boldsymbol{z}^t}{n^2} \boldsymbol{R}_{\boldsymbol{sq}}^{-1} \mathbf{1} \tag{1.21}$$

where $\boldsymbol{z} = \sqrt{\boldsymbol{\chi^2}}$ from eq. (1.14), with the direction of effect as its sign and $\circ$ is the element-wise product (Hadamard product).

The problem with eq. (1.21) 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. (1.16) 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

$$\mathrm{Var}(H) = \frac{2(e + 2H)}{n^2} \tag{1.22}$$

eq. (1.22) should in theory gives us an heuristic estimation of the SE. Moreover, the direction of effect was not required for eq. (1.22), reducing the number of input required from the user.

1.2.3 Case Control Studies

When dealing with case control data, as the phenotype were usually discontinuous, we cannot directly use eq. (1.17) to estimate the heritability. Instead, we will need to employ the concept of liability threshold model from **??**.

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

$$\frac{NCP_{CC}}{NCP_{QT}} = \frac{i^2 v(1-v)N_{CC}}{(1-K)^2 N_{QT}} \tag{1.23}$$

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. (1.23) became

$$NCP_{QT} = \frac{NCP_{CC}(1-K)^2}{i^2 v(1-v)} \tag{1.24}$$

By combining eq. (1.24) and eq. (1.14), we can then have

$$f = \frac{(\chi^2_{CC} - 1)(1-K)^2}{ni^2v(1-v)} \tag{1.25}$$

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

$$\hat{Heritability} = \frac{(1-K)^2}{i^2v(1-v)}\mathbf{1}^t\boldsymbol{R}_{\boldsymbol{sq}}^{-1}\boldsymbol{f} \tag{1.26}$$

1.2.4 Extreme Phenotype Selections

When extreme phenotype selection were performed, the variance of the selected phenotype will not be representative of that in the population. Most notably, the variance of the post selection phenotype will tends to increase. 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):

$$\hat{Heritability} = \frac{V_{P'}}{V_P}\mathbf{1}^t\boldsymbol{R}_{\boldsymbol{sq}}^{-1}\boldsymbol{f} \tag{1.27}$$

1.2.5 Calculating the Linkage Disequilibrium matrix

To estimate the heritability, the population LD matrix is required. In reality, one can only obtain the LD matrix based on a subset of the population (e.g. the 1000 genome project(Project et al., 2012) or the HapMap project(Altshuler et al., 2010)). There are therefore sampling errors among the LD elements.

Now if we consider eq. (1.17), the $\boldsymbol{R}_{\boldsymbol{sq}}$ matrix is required. As the squared LD is used, a positive bias is induced into our $\boldsymbol{R}_{\boldsymbol{sq}}$ matrix.

Based on Shieh (2010), one can correct for bias in the Pearson correlation ρ using

$$\rho = \rho\{1 + \frac{1-\rho^2}{2(N-4)}\} \tag{1.28}$$

where N is the number of sample used in the calculation of ρ. Similarly, there exists a bias

correction equation for ρ^2:

$$\rho^2 = 1 - \frac{N-3}{N-2}(1-\rho^2)\{1 + \frac{2(1-\rho^2)}{N-3.3}\} \tag{1.29}$$

Therefore, we corrected the $\boldsymbol{R_{sq}}$ based on eq. (1.29) such that the bias in estimation can be minimized.

1.2.6 Inverse of the Linkage Disequilibrium matrix

In order to obtain the heritability estimation, we will require to solve eq. (1.17). 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. (1.17) as calculating the sum of $\hat{\boldsymbol{h^2}}$ from eq. (1.16). This will involve solving for

$$\hat{\boldsymbol{h^2}} = \boldsymbol{R_{sq}^{-1}}\boldsymbol{f} \tag{1.30}$$

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. (1.30) 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 $\hat{\boldsymbol{h^2}}$.

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

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^t \tag{1.31}$$

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. (1.31), we can get the inverse of $\boldsymbol{A}$ as

$$\boldsymbol{A}^{-1} = \boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^t \tag{1.32}$$

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

$$\hat{\boldsymbol{B}}_{\boldsymbol{i}} = \boldsymbol{B}_{\boldsymbol{i}} + \epsilon_i \tag{1.33}$$

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

$$\begin{aligned} \boldsymbol{A}\boldsymbol{x} &= \hat{\boldsymbol{B}} \\ \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^t\boldsymbol{x} &= \hat{\boldsymbol{B}} \\ \boldsymbol{x} &= \boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^t\hat{\boldsymbol{B}} \end{aligned} \tag{1.34}$$

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. (1.34) that when the singular value σ_i is small, the error ϵ_i in eq. (1.33) 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. (1.31), the tSVD of $\boldsymbol{A}$ can be represented as

$$\boldsymbol{A}^{+} = \boldsymbol{U}\boldsymbol{\Sigma}_k\boldsymbol{V}^t \qquad \text{and} \qquad \boldsymbol{\Sigma}_k = \text{diag}(\sigma_1, \ldots, \sigma_k, 0, \ldots, 0) \tag{1.35}$$

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

$$\sigma_i = \begin{cases} \sigma_i & \text{for} \quad \sigma_i \geq t \\ 0 & \text{for} \quad \sigma_i < t \end{cases} \tag{1.36}$$

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 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:

$$gap = \sigma_k / \sigma_{k+1} \tag{1.37}$$

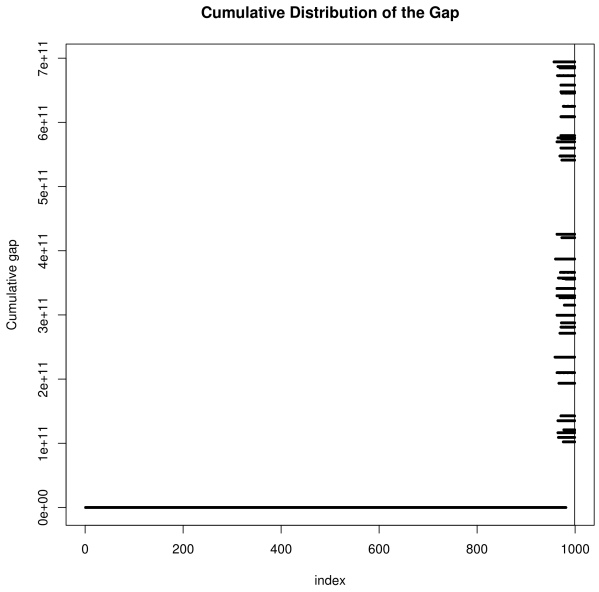
a large gap usually indicate a well-defined gap.

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. 1.1). 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. (1.16), and obtain the estimated heritability.

Figure 1.1: 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



1.2.7 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

$$\text{Var}(\beta) = \frac{h^2}{M}\boldsymbol{I} \tag{1.38}$$

M = number of SNPs

β = 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.

1.3 Assessing the Performance of Our Algorithm

First, we would like to test how well our algorithm works for heritability estimation under different scenarios. To account for different genetic architecture, we varies the heritability of the trait, the number of causal SNPs and the genotypes(therefore varies the LD pattern) during the quantitative trait simulation.

1.3.1 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. 1.2). 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.

1.3.2 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. Fr example, in the Psychiatric Genomics Consortium (PGC) schizophrenia GWAS, more than 9 million SNPs were included, with around 700,000 SNPs on chromosome 1. Although it would be idea to simulate 700,000 SNPs in our simulation, the time required for simulating the samples will become unrealistic.

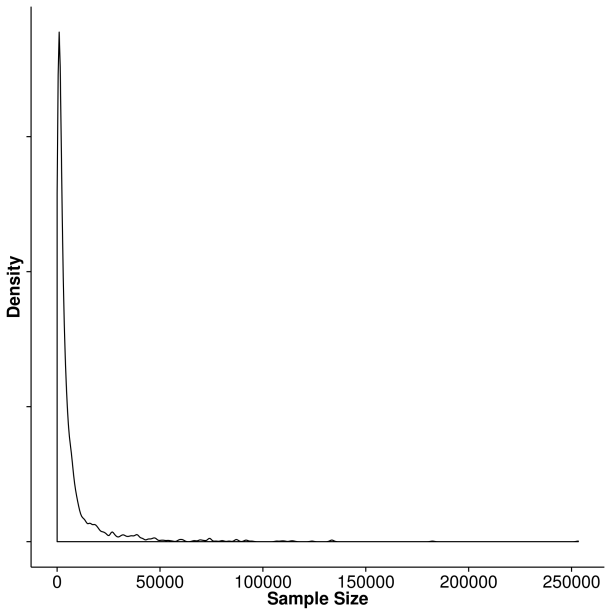


Figure 1.2: GWAS sample size distribution.

As the number of SNPs simulated grow, more time were required for the simulation of samples and more calculation will be required. Moreover, the increasing number of SNPs will lead to increased size of the LD matrix, requiring a long time for the inverse of the matrix. In reality, this should not be a real problem as one typically only calculate the heritability of the data set once and the speed of the algorithm is still relatively fast. However, in the case of simulation where we would like to repeat the same analysis many times, the small increment of time will lead to an escalation in total simulation time, making the simulation infeasible. 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.

1.3.3 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.

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 Northern Europeans from Utah (CEU) haplotypes as an input. In short, HAPGEN2 simulate 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). In a typical GWAS , one usually only have power in detecting "common variants", usually defined as variants with Minor Allele Frequency (maf) ≥ 0.01. We therefore only consider scenario with "common" variants and only use SNPs with maf ≥ 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 $k \in \{5, 10, 50, 100, 250, 500\}$. An important consideration in the simulation of causal SNPs is the effect size distribution. 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:

$$\begin{aligned} \theta &= \exp(\lambda = 1) \\ \beta &= \pm\sqrt{\frac{\theta \times h^2}{\sum \theta}} \end{aligned} \tag{1.39}$$

with a random direction of effect.

Given the normalized genotype as $\boldsymbol{X}$ and the simulated heritability as h^2, the

phenotype can then be calculated as

$$\begin{aligned}
\epsilon_i &\sim N(0, \sqrt{\mathrm{Var}(\boldsymbol{X}\boldsymbol{\beta})\frac{1-h^2}{h^2}}) \\
\boldsymbol{\epsilon} &= (\epsilon_1, \epsilon_2, ..., \epsilon_n)^t \\
\boldsymbol{y} &= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}
\end{aligned} \tag{1.40}$$

The test statistics were then calculated using the plink programme (S. Purcell et al., 2007) and input to our algorithm to estimate the heritability. An independent 500 samples were simulated as a reference panel for the calculation of LD matrix as in reality, we should not have the sample genotype for the construction of the LD matrix. The whole process will be repeated 50 times such that a distribution of the estimate can be obtained. The whole simulation process can be viewed as follow: 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.

In summary:

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. (1.39)
4. Randomly assign the effect size to k SNPs
5. Simulate 1,000 samples using HAPGEN2 and calculate their phenotype according to eq. (1.40)
6. Perform heritability estimation using our algorithm
7. Repeat step 5-6 50 times
8. Repeat step 1-7 10 times

1.4 Comparison with Other Algorithms

It is also important for us to compare our algorithm to existing methods for the performance in estimating the narrow sense heritability. We would also like to extend the simulation to other conditions such as quantitative traits with extreme effect size distribution, case control studies and quantitative traits with extreme phenotype selections.

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 considered as the golden standard 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.

1.4.1 Simulation

We first repeated the whole simulation process in section 1.3 by including GCTA, LDSC with fixed intercept and LDSC with intercept estimation on top of our algorithm. The sample genotype was provided to GCTA for the calculation of the genetic relationship matrix and the estimation of heritability. On the other hand, the LD score for LDSC were calculated using the same simulated reference panel as our algorithm to simulate conditions where the sample genotype was not available. So the simulation follows the following procedure:

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. (1.39)
4. Randomly assign the effect size to k SNPs
5. Simulate 1,000 samples using HAPGEN2 and calculate their phenotype according to eq. (1.40)

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

1.4.2 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

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

$$\begin{aligned} \beta_{eL} &= \pm\sqrt{\frac{0.5h^2}{m}} \\ \beta_{eS} &= \pm\sqrt{\frac{0.5h^2}{100-m}} \\ \beta &= \{\beta_{eL}, \beta_{eS}\} \end{aligned} \tag{1.41}$$

The effect size were then randomly assigned to 100 causal SNPs and phenotype will be calculated as in eq. (1.40). 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. (1.41), 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. (1.40)

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

1.4.3 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 100 SNPs with effect size simulated based on eq. (1.39). 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. (1.40). 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. (1.39)
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. (1.40)
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

1.4.4 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. (1.40), 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.

1.5 Result

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

1.5.1 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. As the simulated heritability increases, the bias of the estimate systematically increases. One possible reason for the general up-ward bias was that we under-estimate the R^2 during the analysis. When considering eq. (1.29), it was observed that the R^2 were downwardly corrected. Therefore we hypothesize that by not performing the bias correcting in the LD, we might get more accurate estimation of the heritability.

We re-run the simulation on section 1.3, this time comparing the performance of SHREK with and without the LD correction. It was observed that when SHREK was performed without LD correction, there was a downward bias. The magnitude of bias was much smaller when compared to the estimation with LD correction. On top of that, the variance of estimation was also smaller when LD correction was not performed. 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.

1.5.2 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

To study the performance of SHREK and LDSC in comparison to GCTA, we performed a variety of simulations to model scenarios with different number of causal SNPs, different effect size distribution and different type of traits.

First, we examined the performance of the algorithms under the quantitative trait scenario. In the quantitative trait scenario, we varies the number of causal SNPs and either assigned an equal effect size to each causal SNPs or assigned a per-allele effect sizes drawn from the squared root of the exponential distribution with $\lambda = 1$.

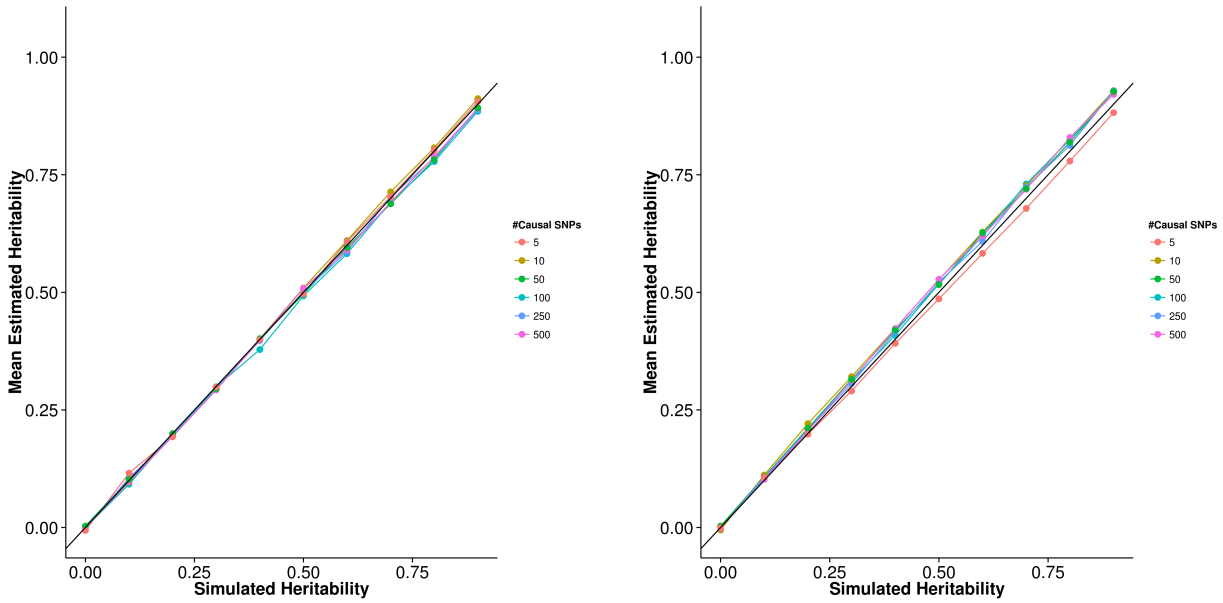
1.5.3 Quantitative Trait Simulation with Equal Effect Size

The simulation of equal effect size serves as a simplistic baseline model for the performance of the programmes. The first thing to look at is the mean estimation of heritability of the programmes. If there is any bias in the estimation of the programmes, one can easily visualize it by plotting the mean estimated heritability against the simulated heritability(fig. 1.3).

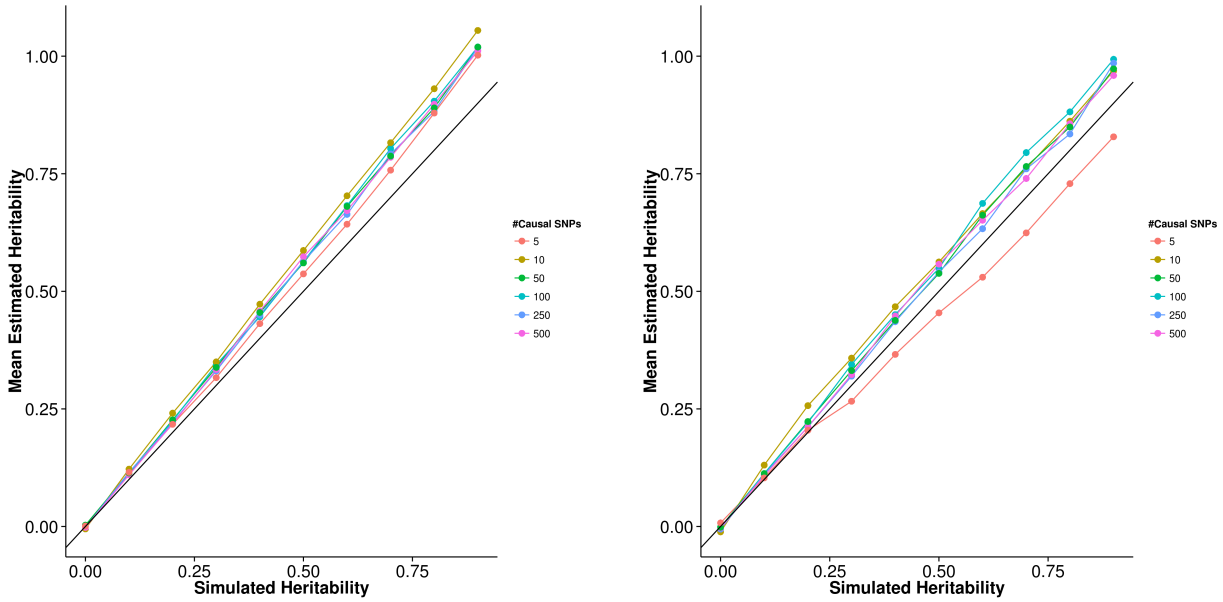
From the graph, it is clear that there was a slight over estimation for LDSC with fixed intercept(fig. 1.3c). The over estimation seems to be a function of the simulated heritability, where a large inflation was observed when a larger heritability was simulated. On the other hand, when allow for the estimation of intercept, less bias was observed for LDSC except for the scenario where only 5 causal SNPs was simulated where the estimation was downwardly biased.

Comparing to LDSC, SHREK has a smaller bias and tends to slightly underestimate(fig. 1.3a). However, the bias of SHREK is insensitive to the simulated heritability, making it robust to traits with different heritability. Similarly, the bias of GCTA is also smaller than that of LDSC(fig. 1.3a), with a slight upward bias in the estimation except when 5 causal SNPs was simulated. Again, the estimate of GCTA is also relatively insensitive to the simulated heritability. Overall, there was no clear pattern as to how the number of causal SNPs affects the mean estimation.

Next, we examine the empirical variance of the programmes(fig. 1.4). As can be seen from the graph, there is a clear pattern where the decrease in number of causal SNP generally increases the variance for all the programmes, with SHREK least affected. For LDSC, the simulated heritability also have a large impact to its empirical variance, with the empirical variance increases as the simulated heritability increases. In general, LDSC

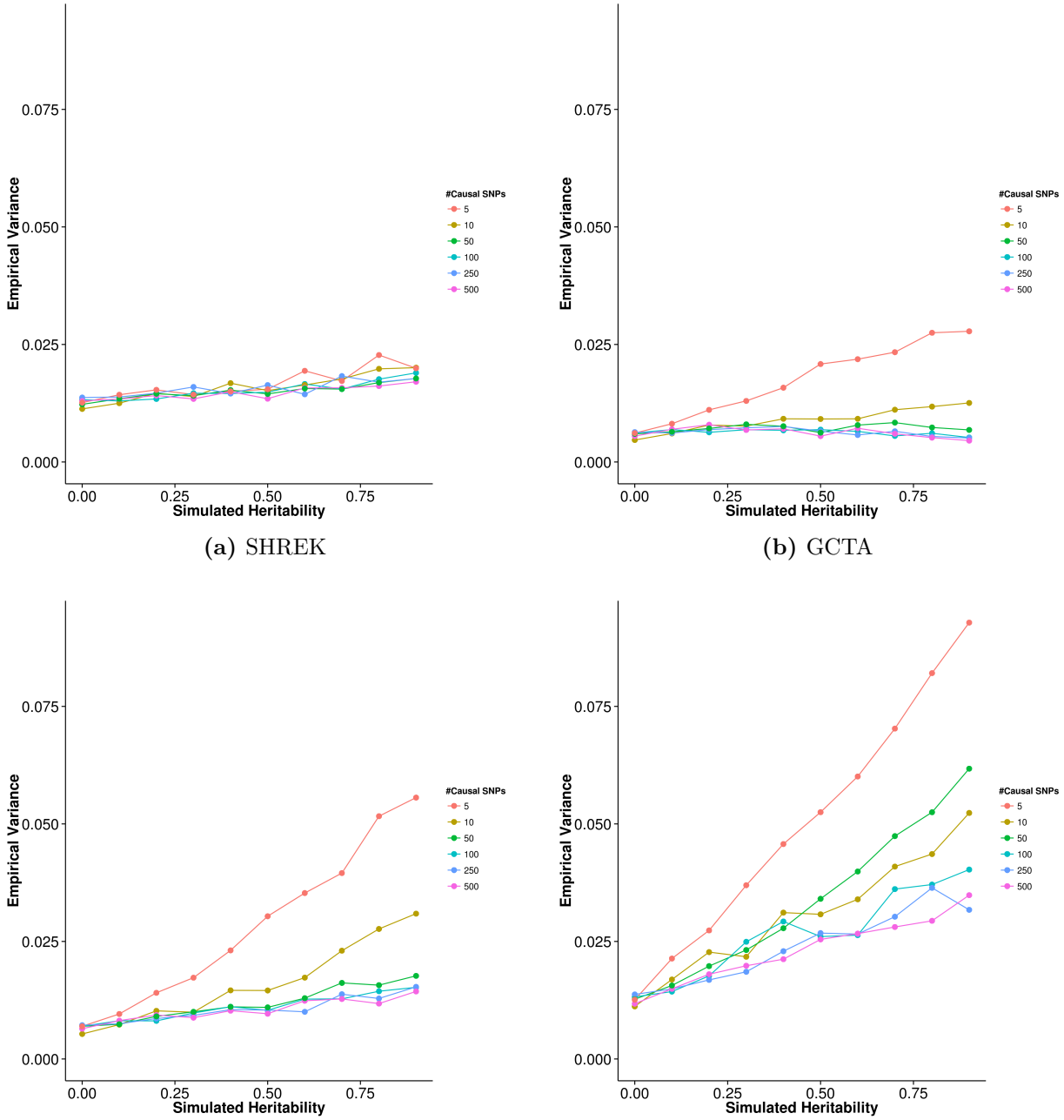


(a) SHREK **(b)** GCTA



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

Figure 1.3: Mean of results from quantitative trait simulation with equal effect size simulation. SHREK was found to be less biased of all the tools whereas there was a slight upward bias for LDSC when the intercept was fixed, especially when the number of causal SNPs was small.



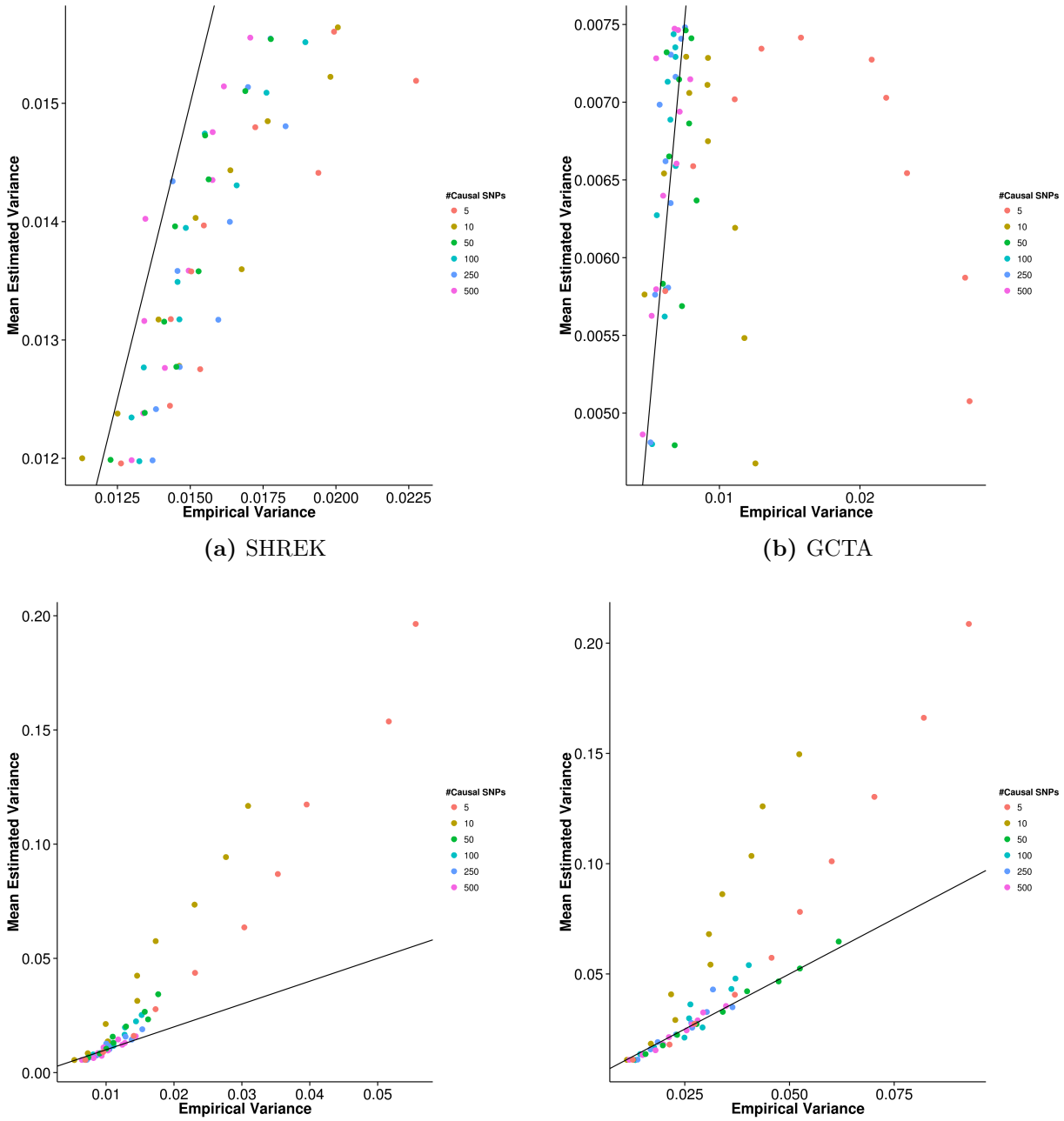
(a) SHREK

(b) GCTA

(c) LDSC with fix intercept

(d) LDSC with intercept estimation

Figure 1.4: Variance of results from quantitative trait simulation with equal effect size simulation. Of all the programmes, GCTA was found to have the lowest variance, follow by LDSC with fixed intercept. The variance of SHREK was slightly higher than that of LDSC with fixed intercept and is lower than that of LDSC with intercept estimation. Unlike LDSC, the variance of SHREK was less sensitive to change in total heritability.



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

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

Figure 1.5: Estimated variance of results from quantitative trait simulation with equal effect size simulation compared to the empirical variance. The estimated variances of all the tools were rather sensitive to the number of causal SNPs, where LDSC tends to over-estimate the variance as the number of causal SNPs decreases and SHREK and GCTA tends to under-estimate the variance.

with fixed intercept(fig. 1.4c) has a lower variance when compared to LDSC with intercept estimation(fig. 1.4d). Moreover, when the number of causal SNP is large, the variance of LDSC with fixed intercept(fig. 1.4c) is lower than SHREK(fig. 1.4a). However, SHREK is more robust to change in the number of causal SNPs and simulated heritabiliy when compared ot LDSC.

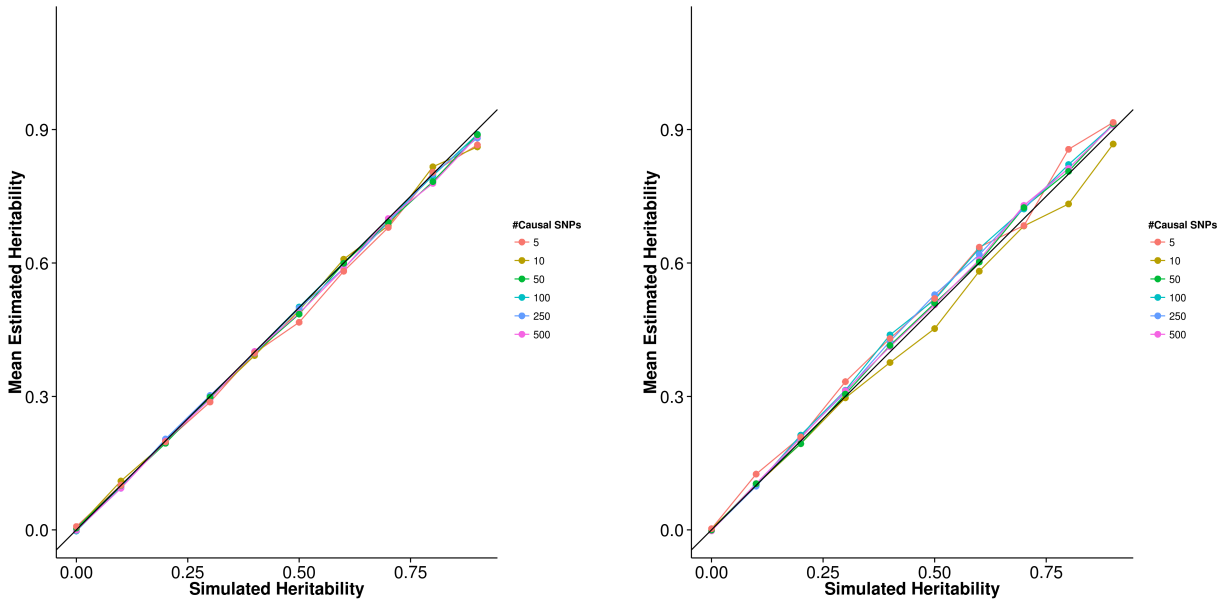
Of all the programmes, GCTA has the best performance(fig. 1.4b) except when the trait only contains 5 causal SNPs. Not only does it has the smallest variance, its empirical variances was almost invariant to change in simulated heritability. However, the case with 5 causal SNPs serves as an out-lier. It was most obvious when inspecting the relationship between the estimated variance and the empirical variance of GCTA(fig. 1.5b). Comparing the estimated variance and the empirical variance, it was clear that GCTA can, in most case accurately estimate its variance. In the case of 5 causal SNPs however, GCTA underestimates its variance. It was also observed in the case of 10 causal SNPs, there was already a slight under-estimation of the variance, suggesting that there might be an increase in empirical variance that was not capture by GCTA.

In the case of the programmes using the test statistic, it was observed that SHREK in general under-estimate the empirical variance(fig. 1.5a) for an average of 0.9 fold. On the other hand, LDSC over-estimates the variance for roughly 1.5 times when a fixed intercept(fig. 1.5c) was used and roughly 1.2 times when the intercept was estimated(fig. 1.5d).

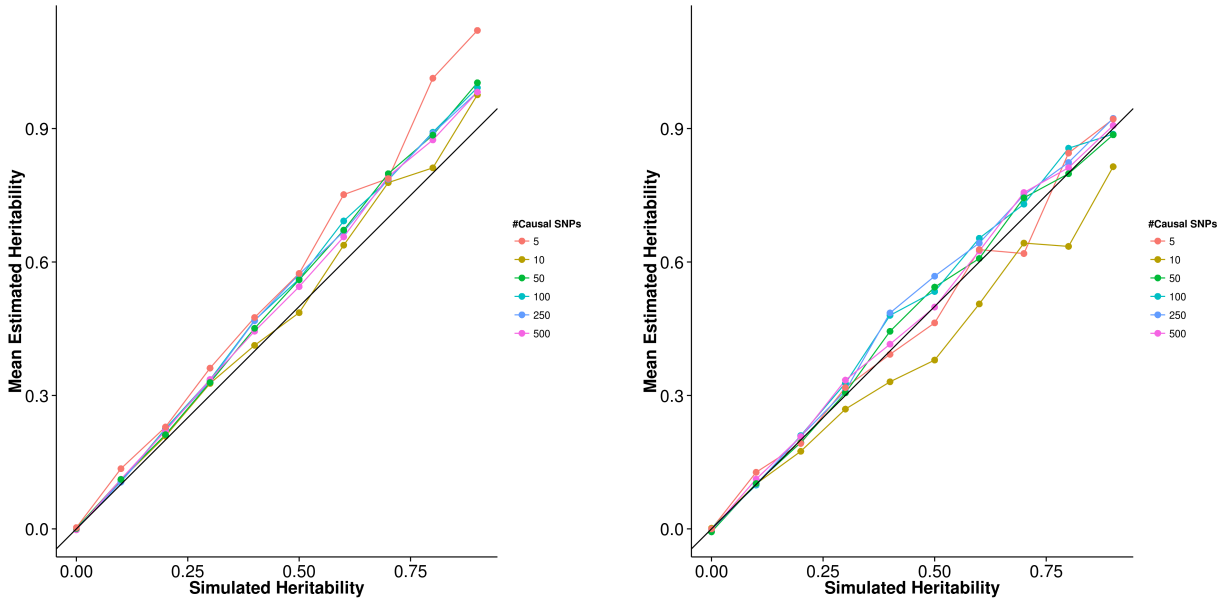
To summarize the results, we calculate the mean squared error (MSE) of the estimation of heritability of the programmes under different simulation condition(table 1.1). With the exception of the 5 causal SNPs scenario, GCTA has the best performance, has a almost 2 fold smaller MSE when compared to SHREK. As the number of casual SNPs increases, the performance of LDSC with fixed intercept and SHREK converges where in general, SHREK has a smaller MSE. Interestingly, unlike LDSC, SHREK was insensitive to change in number of causal SNPs and its performance were relatively stable.

1.5.4 Quantitative Trait Simulation with Random Effect Size

Next, we simulate quantitative trait with random effect size assigned to the causal SNPs. The exponential distribution with $\lambda = 1$ was selected because it was suggested that it may serve as a heuristic expectation the genetic architecture of adaptation(Orr, 1998). There might be many other distribution that can be used, however due to limitation in resources, we will only focus on the exponential distribution with $\lambda = 1$.

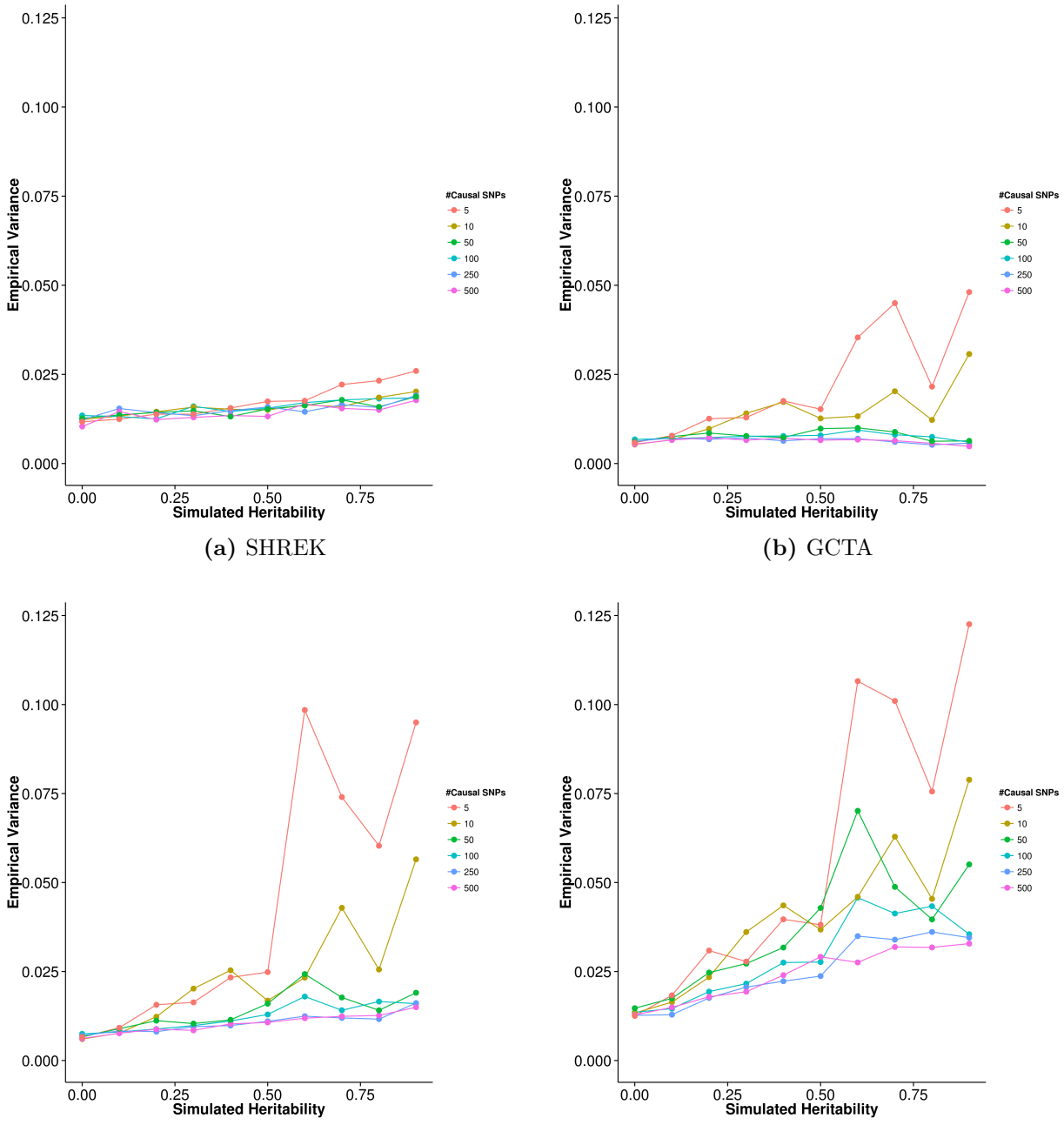


(a) SHREK **(b)** GCTA



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

Figure 1.6: Mean of results from quantitative trait simulation with random effect size simulation. The result was very much similar to the condition where a equal effect size was simulated. Again, SHREK has the most accurate mean estimate when compared to other tools, with LDSC slightly inflated.



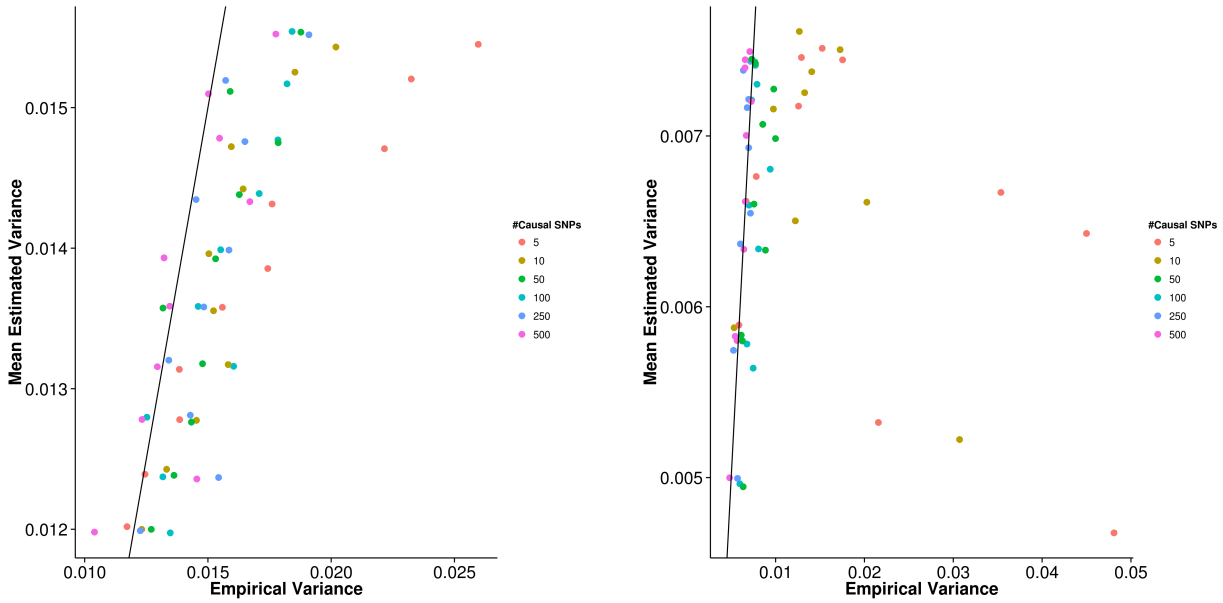
(a) SHREK

(b) GCTA

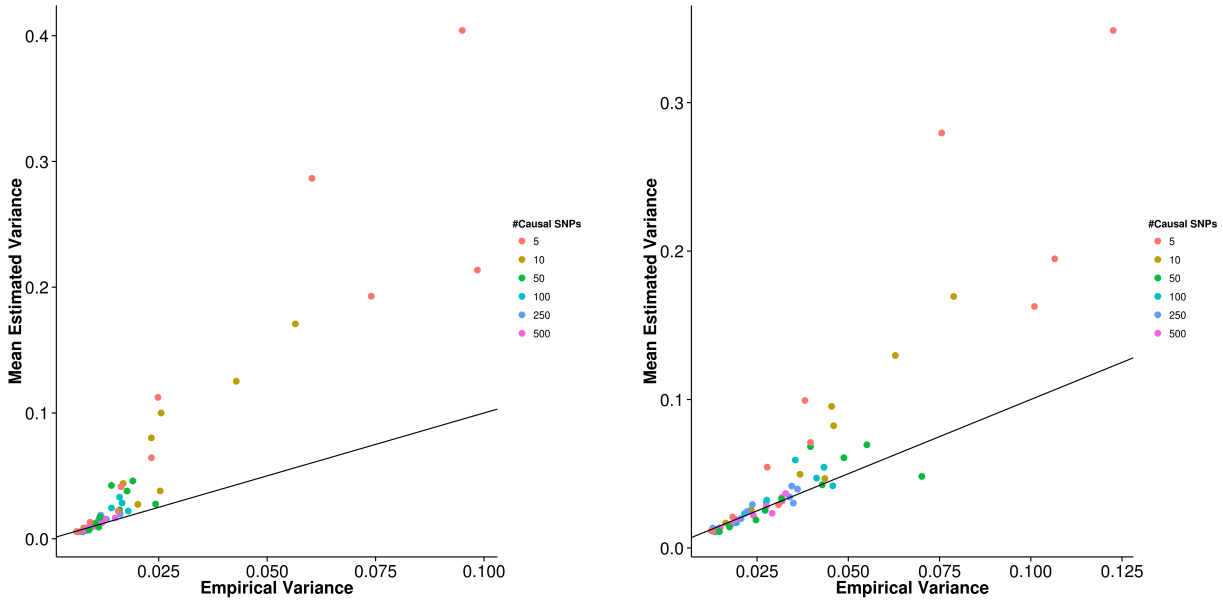
(c) LDSC with fix intercept

(d) LDSC with intercept estimation

Figure 1.7: Variance of results from quantitative trait simulation with random effect size simulation. Again, the variance of the estimate were almost the same as in simulation of equal effect size where 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 programme, with variance of the SHREK estimate being the least sensitive to change in heritability.



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



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

Figure 1.8: Estimated variance of results from quantitative trait simulation with random effect size simulation when compared to the empirical variance. Similar to the simulation with equal effect size, the estimated variance seems to be affected by the number of causal SNPs.

Number of Causal SNPs	SHREK	LDSC	LDSC-In	GCTA
5	0.167	0.308	0.526	0.177
10	0.158	0.243	0.337	0.0944
50	0.150	0.163	0.354	0.0749
100	0.154	0.161	0.304	0.0664
250	0.157	0.147	0.255	0.0659
500	0.147	0.148	0.247	0.0661

Table 1.1: MSE of quantitative trait simulation with equal effect size. It was observed that the overall MSE of GCTA is very low, follow by SHREK. As the number of causal SNPs decreases, the MSE increases for all programmes. The performance of SHREK and LDSC with fixed intercept converges as the number of causal SNPs increases.

Under this simulation condition, it was observed that the mean estimation of heritability from SHREK(fig. 1.6a) and GCTA(fig. 1.6b) were similar to what was observed in the equal effect size simulation. For LDSC with intercept estimation(fig. 1.6d), less bias was observed with only the 10 causal SNPs scenario being under estimated. On the other hand, the performance of LDSC with fixed intercept remain more or less the same, with a larger degree of fluctuation when small number of causal SNPs(5 or 10) was simulated. The fluctuation in estimate can also be observed in the empirical variance of LDSC(figs. 1.7c and 1.7d). Despite the relative stable performance of GCTA, the empirical variance of GCTA also fluctuate when the number of causal SNPs was small. Such pattern was not observed in SHREK suggesting that it might be robust against the change in number of causal SNPs.

When inspecting the relationship between the estimated and empirical variance, it was observed all programmes have a less accurate estimation of its variance when there is only 5 causal SNPs. The difference was most obvious for GCTA where the under-estimation of variance under the oligo-genic scenario(5 or 10 causal SNPs) was more severe when a random effect size was assigned to the causal SNPs(fig. 1.8b). On the other hand, the degree of bias in estimating the variance remain more or less unchanged for SHREK(fig. 1.8a) and LDSC with intercept estimation(fig. 1.8d), with roughly 0.9 and 1.25 times difference from the empirical variance respectively. However, for LDSC with fixed intercept(fig. 1.8c), the fold difference increased slightly, changed from 1.5 fold difference to 1.65 fold difference.

Overall, simulating the effect size using the exponential distribution only slightly increases the MSE of the programmes when the number of causal SNPs is small and decreases the MSE when the number of causal SNPs is larger. Taking into considering of both the bias and standard error, SHREK has the better performance over LDSC except when the trait is extremely polygenic(e.g. ≥ 500 causal SNPs).

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 1.2: MSE of quantitative trait simulation with random effect size. Again, 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.

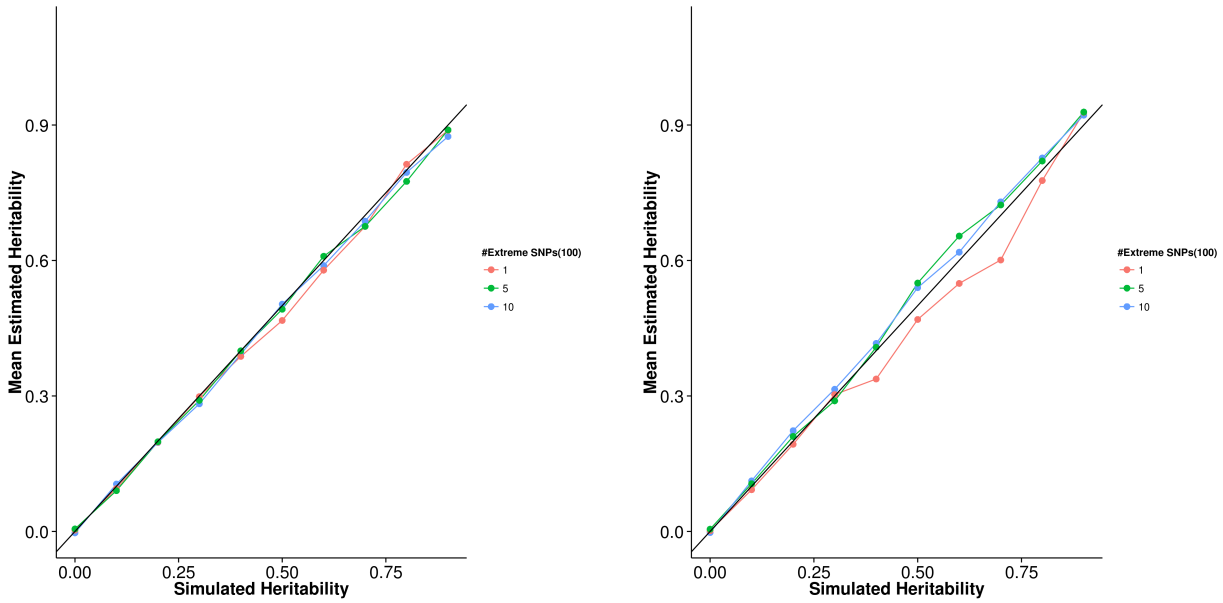
1.5.5 Quantitative Trait Simulation with Extreme Effect Size

Another condition that we were interested in was in the case where a small portion of SNPs has a much larger effect than other SNPs. In this simulation, we simulated either 100 or 250 causal SNPs with 1, 5 or 10 SNPs having a much larger effect.

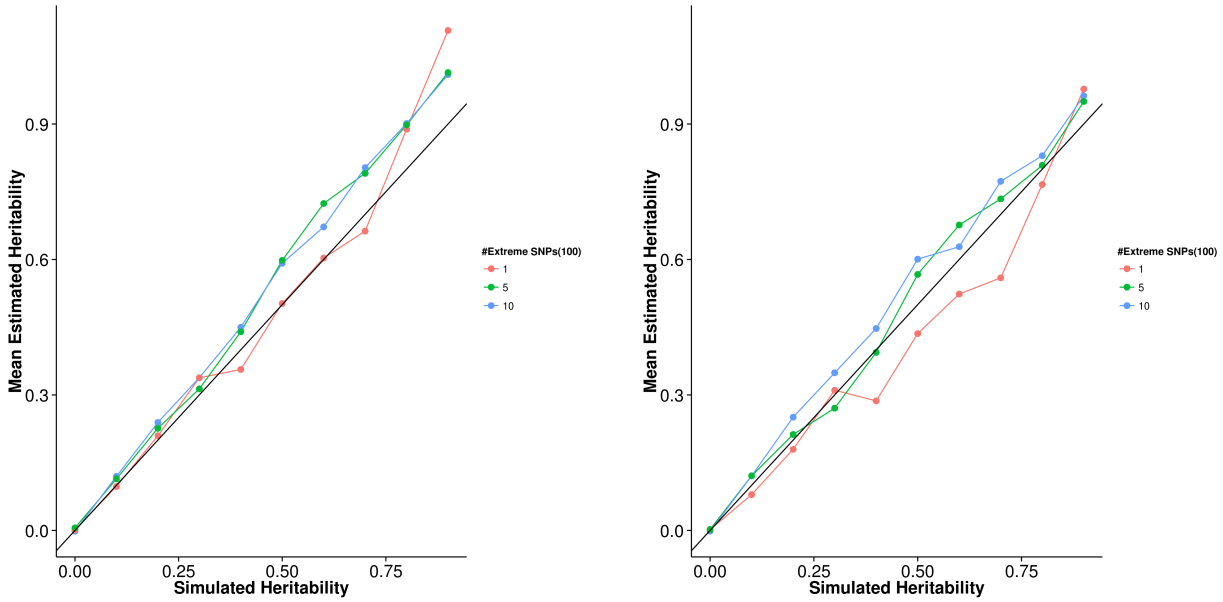
1.5.6 Case Control Simulation

1.6 Discussion

1.7 Supplementary place holder

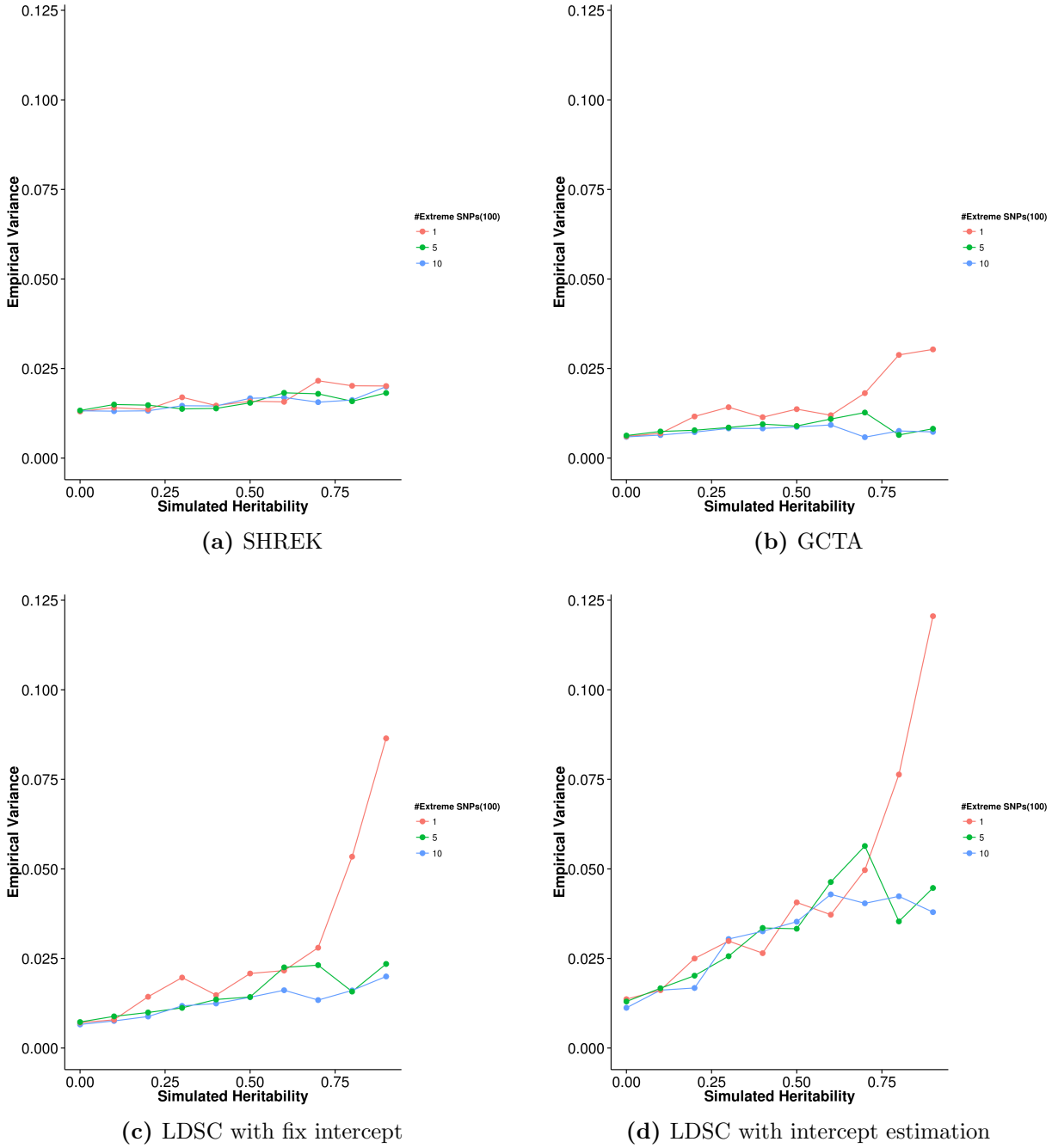


(a) SHREK **(b)** GCTA



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

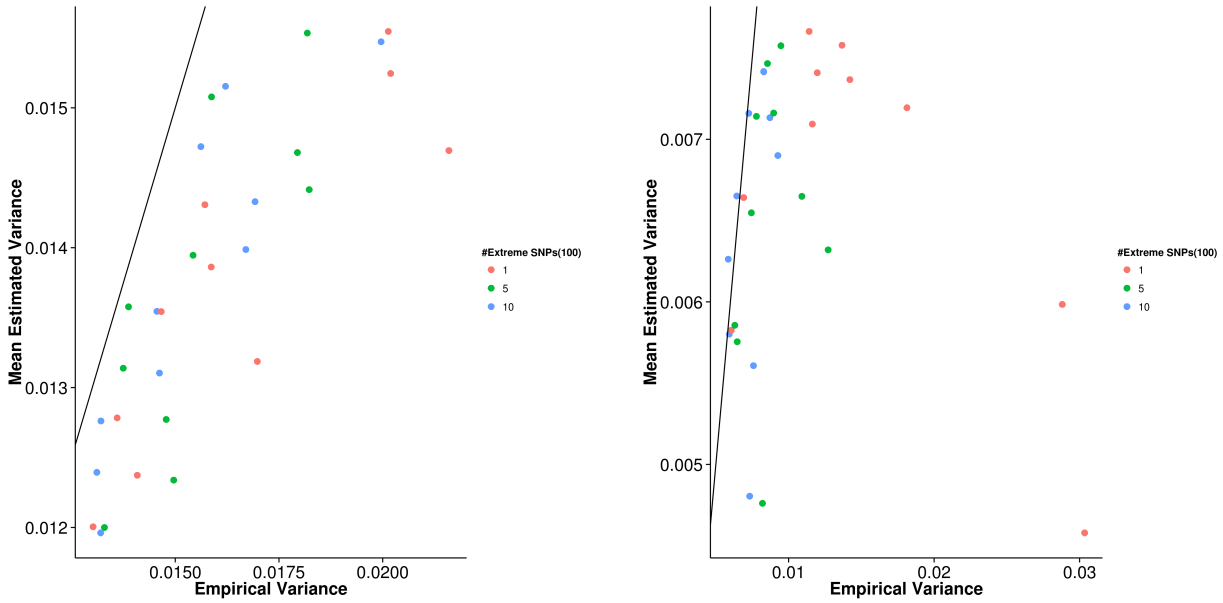
Figure 1.9: Mean of results from quantitative trait simulation with extreme effect size simulation. 100 causal SNPs were simulated. It was observed that the mean estimation of heritability of all the tools were relatively unaffected by the number of SNPs representing a large portion of effect where SHREK has the least amount of bias.



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

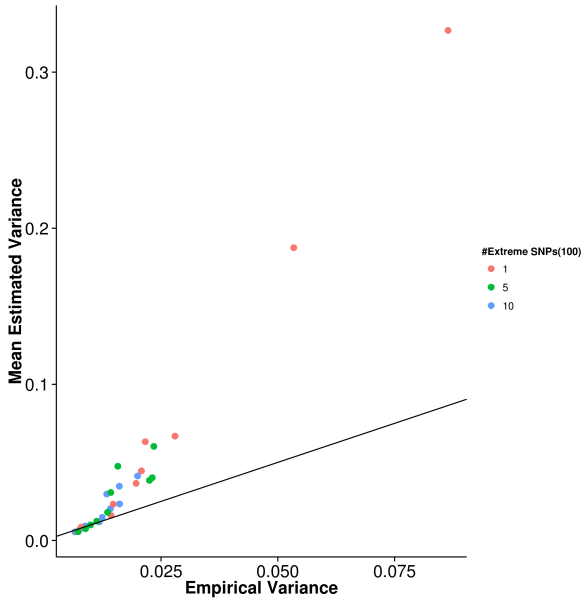
(c) LDSC with fix intercept **(d)** LDSC with intercept estimation

Figure 1.10: 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.

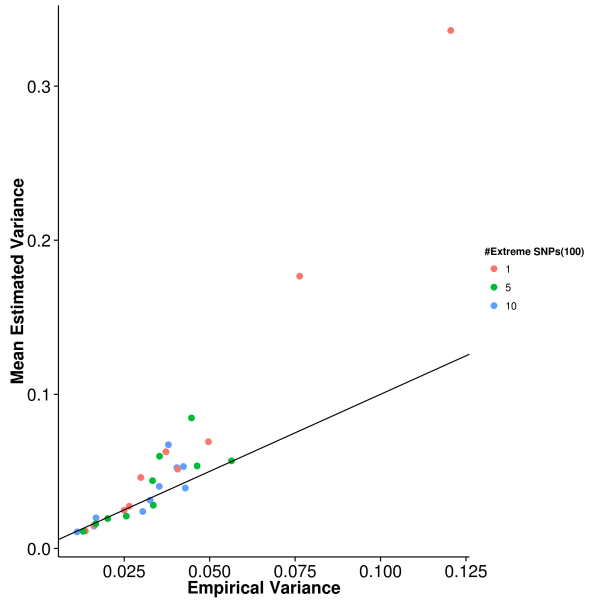


(a) SHREK

(b) GCTA

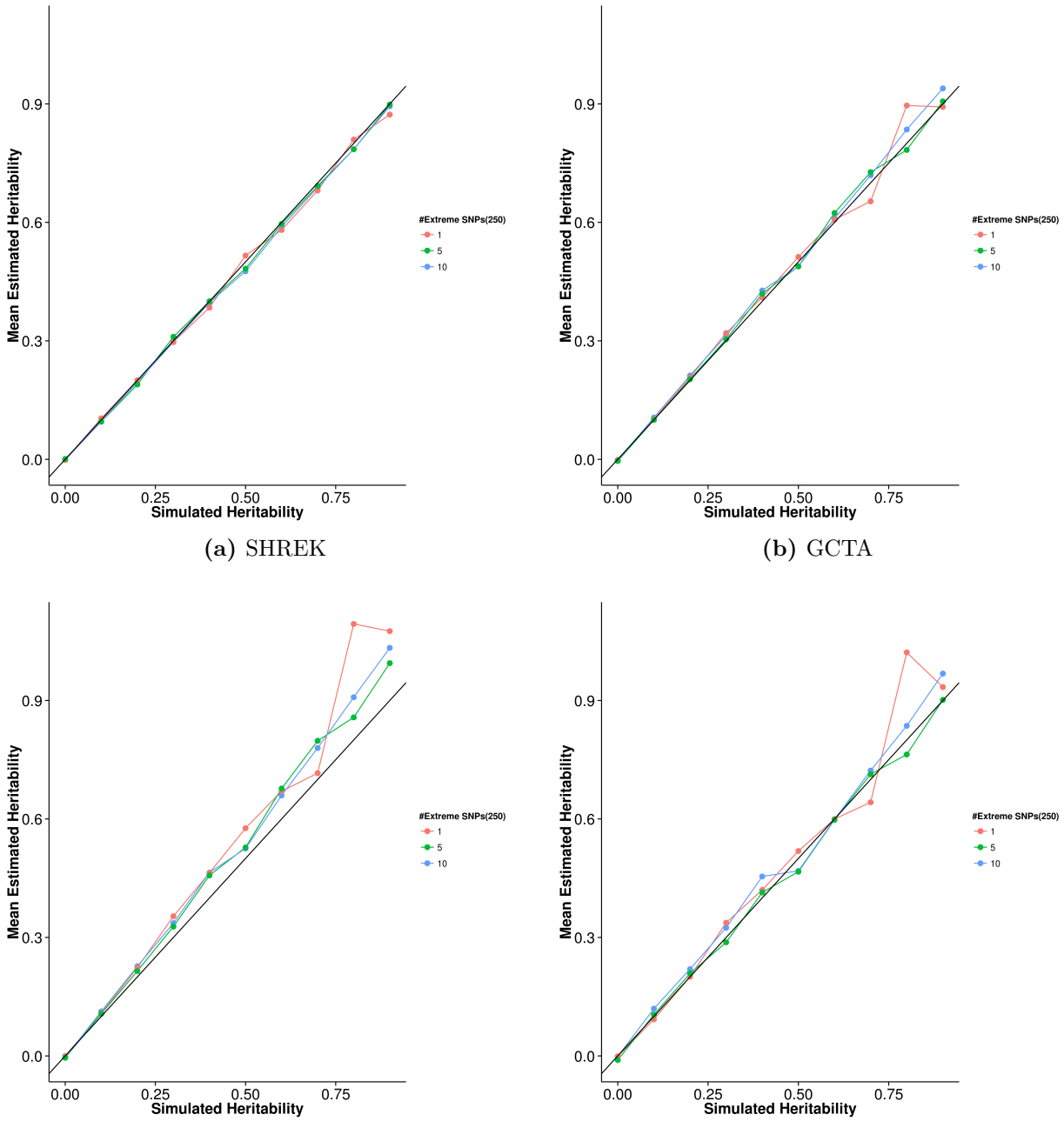


(c) LDSC with fix intercept



(d) LDSC with intercept estimation

Figure 1.11: 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.



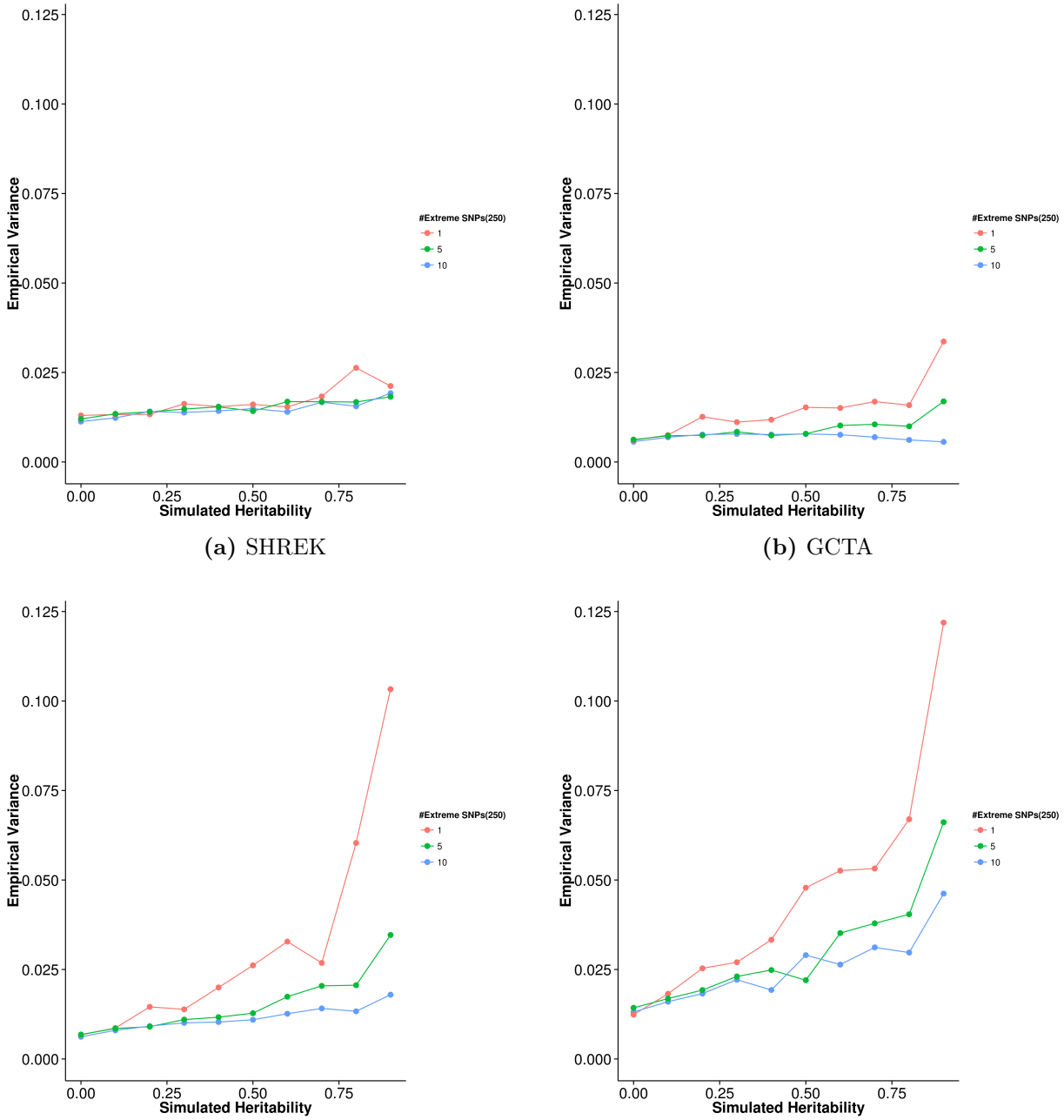
(a) SHREK

(b) GCTA

(c) LDSC with fix intercept

(d) LDSC with intercept estimation

Figure 1.12: Mean of results from quantitative trait simulation with extreme effect size simulation. 250 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, similar to what observed when 100 causal SNPs were simulated. However, there seems to be an upward bias when LDSC was performed with fixed intercept.



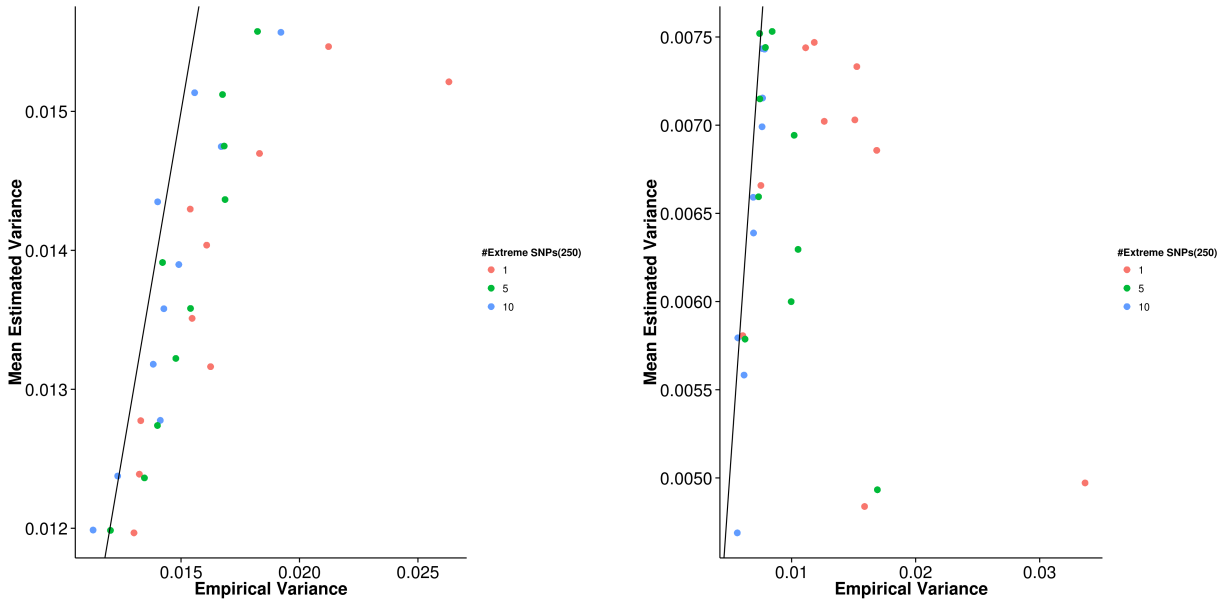
(a) SHREK

(b) GCTA

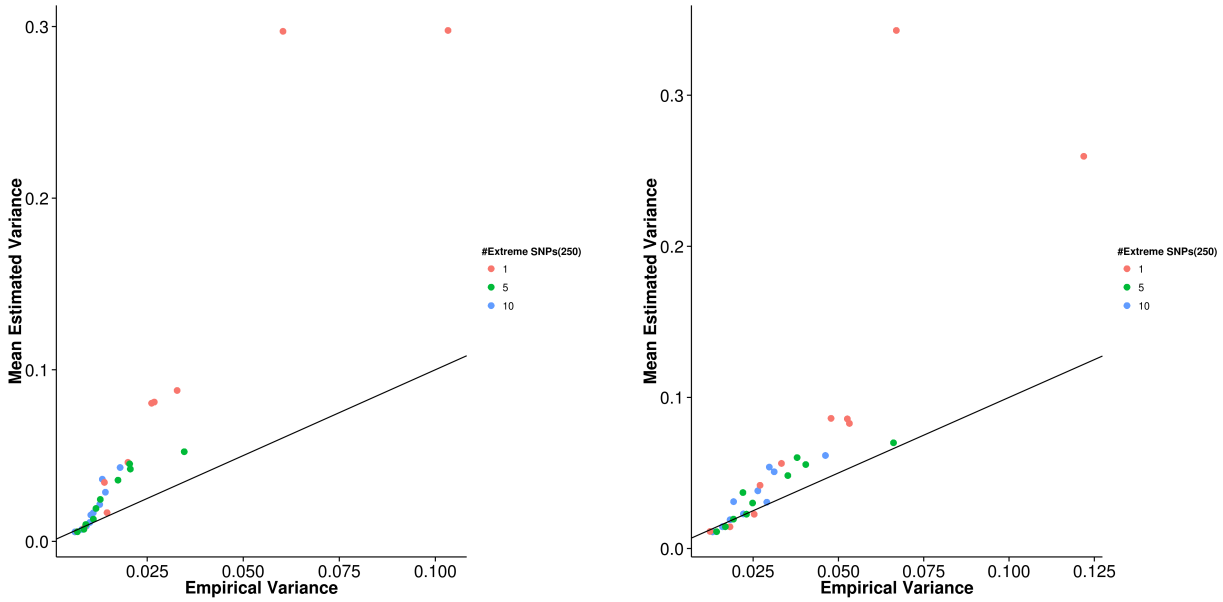
(c) LDSC with fix intercept

(d) LDSC with intercept estimation

Figure 1.13: Variance of results from quantitative trait simulation with extreme effect size simulation. 250 causal SNPs were simulated. Compared to the case where 100 causal SNPs were simulated, most tools, except SHREK seems to be more sensitive to the number of SNP(s) explaining large portion of effect, where a smaller number can lead to a higher variance.

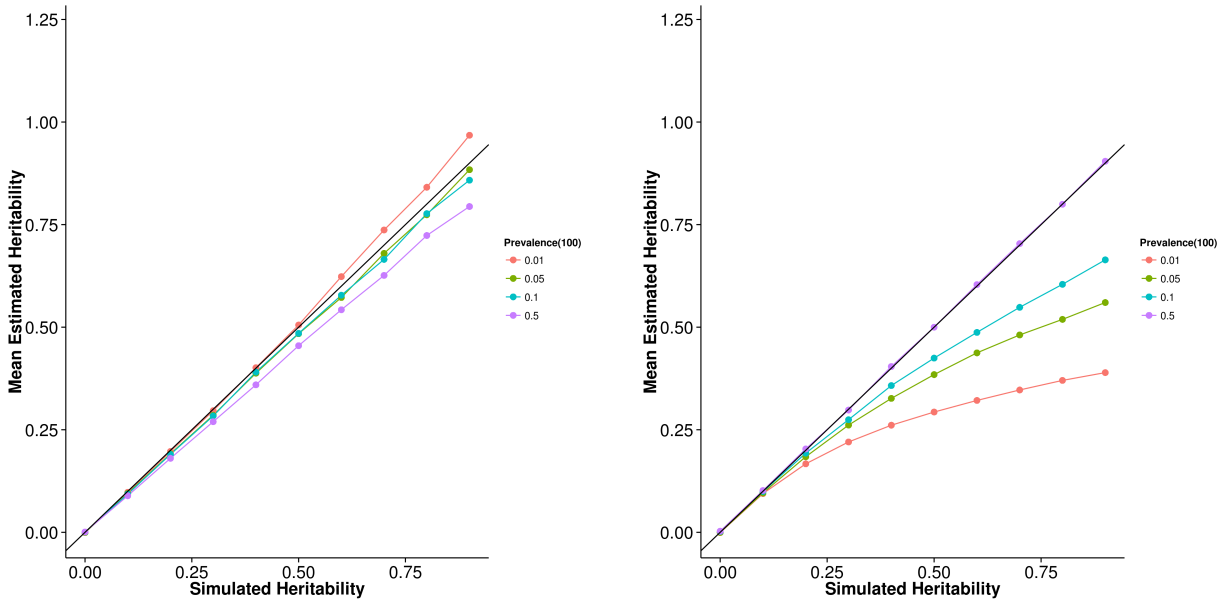


(a) SHREK **(b)** GCTA



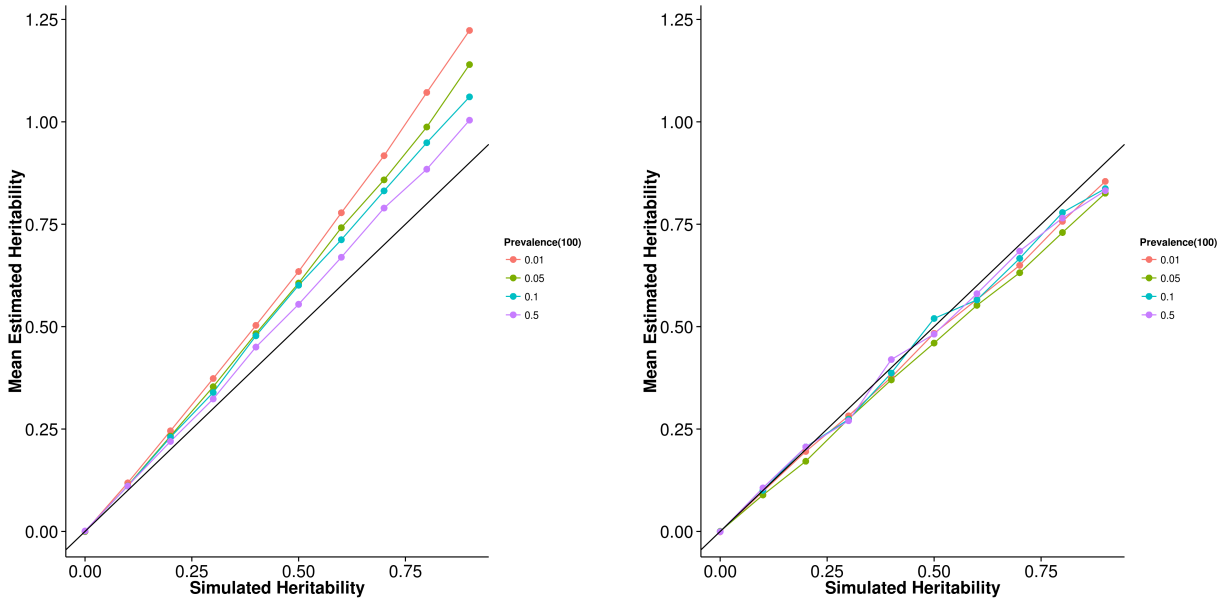
(c) LDSC with fix intercept **(d)** LDSC with intercept estimation

Figure 1.14: Estimated variance of results from quantitative trait simulation with extreme effect size simulation when compared to the empirical variance. 250 causal SNPs were simulated. The result of simulation were the same as the previous extreme effect simulation with 100 causal SNPs.



(a) SHREK

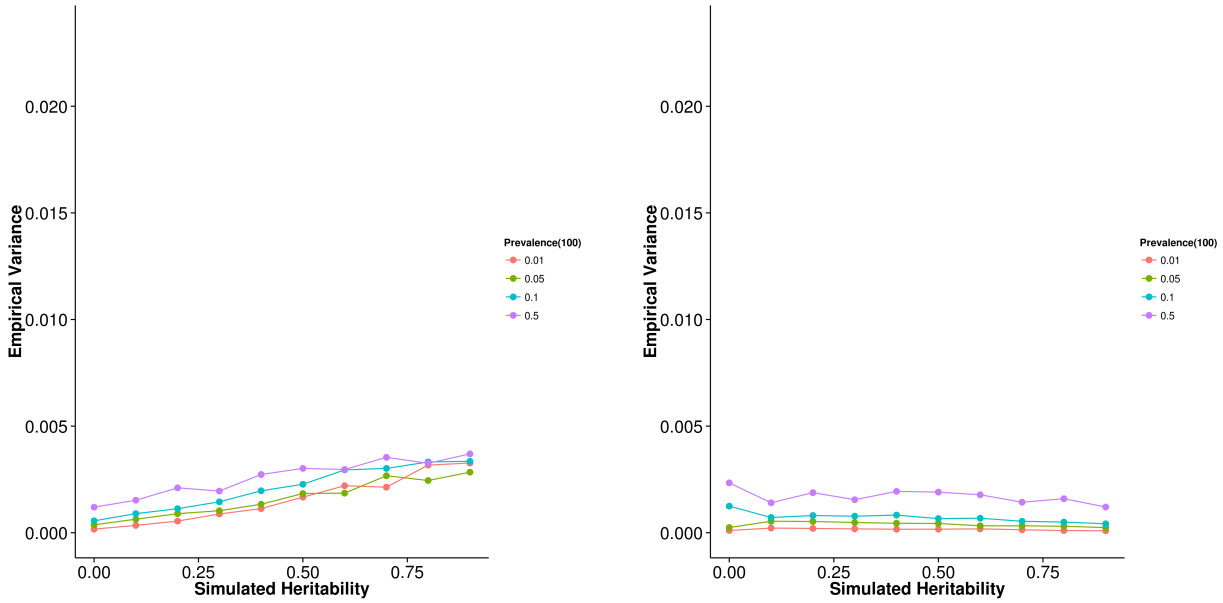
(b) GCTA



(c) LDSC with fix intercept

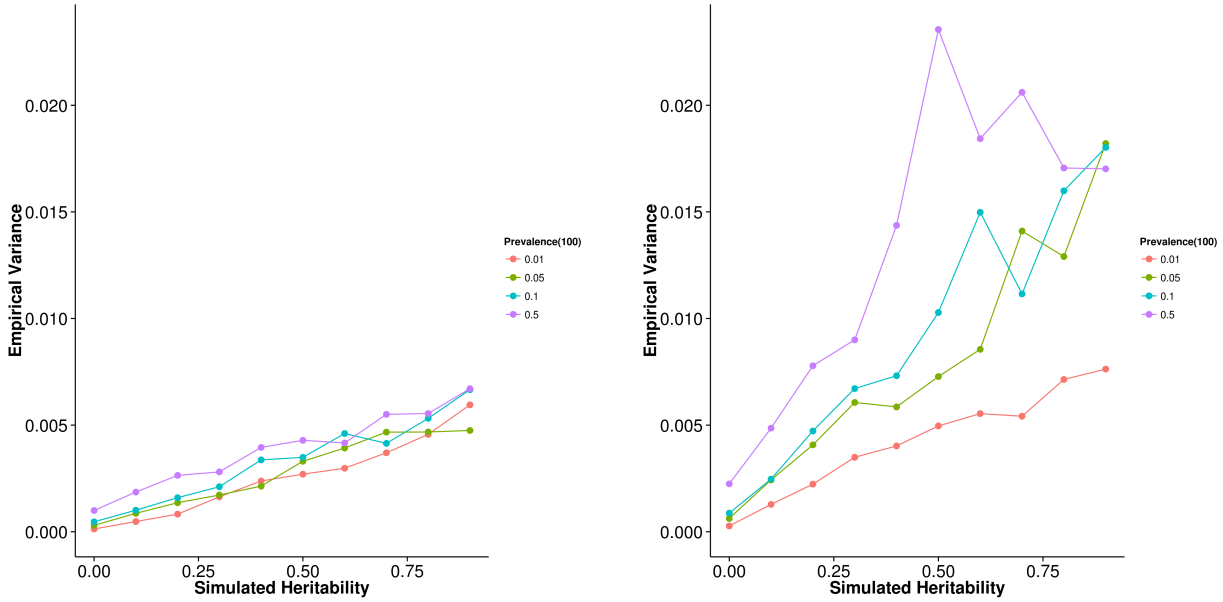
(d) LDSC with intercept estimation

Figure 1.15: 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

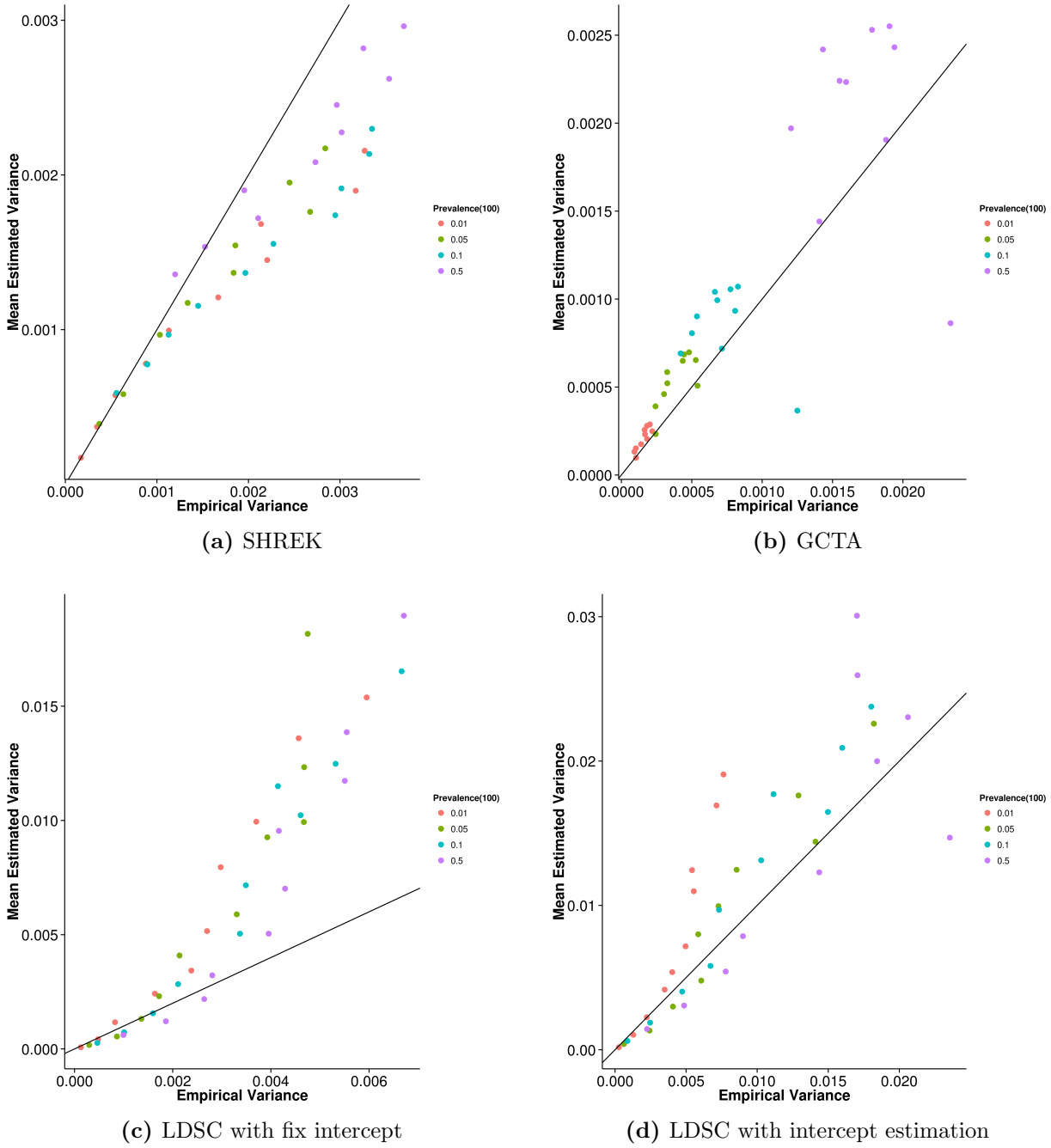
(b) GCTA



(c) LDSC with fix intercept

(d) LDSC with intercept estimation

Figure 1.16: 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 1.17: 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. 1.8), 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 2

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. 10, 13).

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. 1, 19).

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. 1, 40).

Guennebaud, Gaël, Beno\^it Jacob, et al. (2010). *Eigen v3*. http://eigen.tuxfamily.org (cit. on p. 11).

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. 12, 13).

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. 9).

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. 17).

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. 11).

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 pp. 17, 28).

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. 10).

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. 18).

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. 10, 22).

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] (cit. on p. 10).

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 p. 17).

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. 15).

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. 9).

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. 19).

Supplementary Materials

Appendix