Understanding How Genetics and Environments Shape the Development of 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
December 31, 2015

Abstract

Schizophrenia (SCZ) is a detrimental disorder affecting approximately 1% of the population worldwide. To fully understand the disease mechanism for the development of proper treatments, it is important not only to examine how certain genetic polymorphisms can predispose individuals to the disease development, but also how environmental factors triggers the disorder in apparently healthy individuals.

Genome Wide Association Study (GWAS) is now a standard approach for investigating associations of common genetic variations (mainly Single Nucleotide Polymorphisms (SNPs)) with SCZ. A recent meta-analysis of GWAS of SCZ has identified 108 loci significantly associated with SCZ. However, due to the limitation of sample size and the moderate-to-small effect size of an unknown number of causal loci, many SNPs associated with SCZ may be left undetected and a much larger sample size of GWAS may be required. However, it is also possible that these 108 loci have already contained all or near most of the SNPs associated with the disease. So estimating the contribution of these common SNPs to SCZ has important implications for future research strategy.

In this thesis, we proposed an alternative approach for estimating the contribution of SNPs to SCZ (SNP-heritability) from GWAS summary statistics, called the SNP HeRitability Estimation Kit (SHREK). Our simulation results suggested that when compared to the existing method (LD SCore regression (LDSC)), SHREK provided a more robust estimate for oligogentic traits and in case-control designs in which no confounding variables was present. Using the summary statistics from the latest

meta-analysis of GWAS of SCZ, we estimated that SCZ has a SNP-heritability of 0.174 (SD=0.00453), which is similar to the estimate of 0.197 (SD=0.0058) by our competitor LDSC. The result indicated that common SNPs have relatively less contribution to the genetic predisposition of individuals to SCZ as measured by the heritability estimated. Also, it suggested that alternative strategies like whole genome sequencing would be more efficient for identifying additional SCZ genes, compared to GWAS.

On the other hand, prenatal infection has been identified as the single largest environmental risk factor of SCZ. It was observed that a wide variety of infections are associated with the increased SCZ risk in the offspring. This suggests that maternal immune activation (MIA) during prenatal development may have a negative impact on fetal brain functions as well as behaviors. So it is important to understand how MIA triggers the disorder by examining the molecular events that take place in the cerebellum using established animal models, such as those involving the viral RNA mimic polyriboinosinic-polyribocytidilic acid (PolyI:C).

As a result, we also performed a RNA-sequencing study for the MIA on the change in global gene expressions in the fetal cerebellum in PolyI:C-treated pregnant mice. We found that several pathways related to neural functioning and calcium ion signaling were likely to be disrupted by MIA in the cerebellum. In addition, we investigated how a n-3 polyunsaturated fatty acid (PUFA) rich diet can help to reduce the SCZ-like phenotype in mice exposed to early MIA insults. We found that *Sgk1*, a gene that regulates the glutamatergic system, is potentially affected by the n-3 PUFA rich diet in the PolyI:C exposed mice. In conclusion, our results suggested that genes related to neural function or calcium ion signaling, as well as glutamate-related genes such as *Sgk1*, are potential targets for future SCZ research.

(550 words)

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.....

Choi Shing Wan

Acknowledgements

I would like to express my deepest gratitude to Professor Pak Sham. I am eternally grateful for his trust, supervision, patience and support in the course of my study. I would also like to thanks Dr Stacey Cherny and Dr Wanling Yeung for giving me valuable advice for my projects. My special thanks go to Dr Johnny Kwan. He has provided critical advices on my projects and has taught me a great deal in the field of statistic.

The past 4 years has been a blast and I really enjoy my time in this department. This is only possible because of all the great people here. Thank you Beatrice Wu, Dr Li Qi, Tomy Hui, Vicki Lin, Nick Lin, John Wong, Dr Clara Tang, Dr Amy Butler, Dr Emily Wong, Dr Allen Gui, Dr Sylvia Lam, Yung Tse Choi, Oi Chi Chan, Pui King Wong and Dr Miaoxin Li, without you everything will be much different. I will forever cherish the time I spent with you.

Words alone cannot express my gratitude to Beatrice Wu and my family. Their support and encouragement have been my greatest source of energy and have helped me to continue on with my study.

THANK YOU!

Abbreviations

bp	base pair.
DEG	differentially expressed gene.
ECM	extracellular matrix.
EGF	epidermal growth factor.
ERCC	External RNA Controls Consortium.
FGF	fibroblast growth factor.
GD	Gestation Day.
GWAS	Genome Wide Association Study.
IL-6	Interleukin-6.
kb	kilobase.
LD	Linkage Disequilibrium.
LDSC	LD SCore regression.
LRT	likelihood ratio test.
maf	minor allele frequency.
MAPK	mitogen-activated protein kinase.
MIA	maternal immune activation.
MMP	matrix metalloproteinase.
MSigDB	Molecular Signatures Database.
NGS	next generation sequencing.
PC	Principle Component.
PCA	principle component analysis.
PET	positron emission tomography.
PGC	Psychiatric Genomics Consortium.
PI3K	phosphatidylinositol 3-kinase.
PolyI:C	polyriboinosinic-polyribocytidilic acid.
PUFA	polyunsaturated fatty acid.
QC	quality control.

RIN	RNA integrity number.
rt-PCR	real time PCR.
SCZ	schizophrenia.
SE	standard error.
SHREK	SNP HeRitability Estimation Kit.
SNP	Single Nucleotide Polymorphism.

Contents

List of Figures

List of Tables

2 Heritability Estimation

2.1 Introduction

The development of LD SCore regression (LDSC) (B. K. Bulik-Sullivan et al., 2015) has allow researchers to estimate the true contribution of common Single Nucleotide Polymorphisms (SNPs) to the variance in different diseases. However, it is unclear how different sampling strategies (e.g. extreme phenotype selection) affects the performance of LDSC. Additional simulations might therefore be required to investigate how different samplings affect the performance of LDSC.

The estimation of heritability in discontinuous trait has always been complicated as correction for ascertainment bias is necessary. Nevertheless, the correction of ascertainment bias are nontrivial and often introduce bias to the estimates. For example, Golan, Eric S Lander, and Rosset (2014) observed that Genome-wide Complex Trait Analysis (GCTA) underestimates the heritability explained by common variants for discontinuous traits. The magnitude of this bias is affected by the population prevalence of the trait, the observed prevalence, the true underlying heritability and the number of genotyped SNPs (Golan, Eric S Lander, and Rosset, 2014). Because for discontinuous traits, B. K. Bulik-Sullivan et al. (2015) only investigated the performance of LDSC for traits with heritability of 0.8 and a population prevalence of either 0.1 or 0.01, more simulates are required investigate whether if LDSC suffers from the same bias as GCTA.

Finally, as noted by B. K. Bulik-Sullivan et al. (2015), when there were few causal variants, the standard errors of the LDSC estimates will become very large, meaning that LDSC is best suited to polygenic traits. An alternative algorithm is therefore required for the estimation of SNP heritability for oligogenic traits.

Herein, SNP HeRitability Estimation Kit (SHREK), an alternative algorithm to LDSC for the estimation of SNP heritability based on Genome Wide Association Study (GWAS) summary statistics were introduced. To examine the effect of different sampling strategies and genetic architectures on the performance of LDSC and SHREK, we performed a series of extensive simulation analyses. To demonstrate that SHREK also works outside of simulated data, we also estimated the SNP heritability of schizophrenia and other psychiatric disorders using SHREK.

The work in this chapter were done in collaboration with my colleagues who have kindly provided their support and knowledges to make this piece of work possible. Dr Johnny Kwan, Dr Miaxin Li and Professor Sham have helped to lay the foundation 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 Desmond Campbell, 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

2.2.1 Heritability Estimation

Remember that the heritability (h^2)is defined as the proportion of total variance of the phenotype ($\boldsymbol{y}$) in a population explained by the variation of genetic factors ($\boldsymbol{x}$):

$$h^2 = \frac{\mathrm{Var}(\boldsymbol{y})}{\mathrm{Var}(\boldsymbol{x})}$$

In GWAS, regression are performed between the SNPs and the phenotypes:

$$\boldsymbol{y} = \boldsymbol{\beta x} + \boldsymbol{\epsilon} \tag{2.1}$$

where $\boldsymbol{y}$ and $\boldsymbol{x}$ are both standardized. $\boldsymbol{\epsilon}$ is defined as the error term, accounting for the non-genetic elements contributing to the phenotype (e.g. environment factors). By assuming $\beta\boldsymbol{x}$ to be independent of $\boldsymbol{\epsilon}$, one can transform eq. (2.1) into:

$$\begin{aligned} \mathrm{Var}(\boldsymbol{y}) &= \mathrm{Var}(\beta\boldsymbol{x}) + \mathrm{Var}(\epsilon) \\ \mathrm{Var}(\boldsymbol{y}) &= \beta^2 \mathrm{Var}(\boldsymbol{x}) \\ \beta^2 &= \frac{\mathrm{Var}(\boldsymbol{y})}{\mathrm{Var}(\boldsymbol{x})} \end{aligned} \tag{2.2}$$

As a result, β^2 represents the portion of phenotype variance explained by the variance of genotype.

A challenge in calculating the heritability from GWAS data is that usually only the summary statistic or p-value are provided. Without the raw genotypes, it is impossible to calculate $\mathrm{Var}(\boldsymbol{x})$, thus eq. (2.2) cannot be used.

In order to estimate the SNP heritability using the summary statistics of a GWAS, we first need to estimate the relative effect size of each individual SNPs. It was observed that when $\boldsymbol{x}$ and $\boldsymbol{y}$ are standardized, β^2 equals to the coefficient of determination (r^2). Thus, 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 under the null and n is the number of samples, one can then obtain the r^2 by taking the square of eq. (2.3)

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

Although t^2 follows the F-distribution under the null, it will converge into χ^2 distribution when n is large.

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

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

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

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

When there are Linkage Disequilibrium (LD) between each individual SNPs, the situation becomes more complicated as each SNPs' observed effect will be influenced by 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 such that

$$\boldsymbol{y} \sim f(0, 1)$$

$$\boldsymbol{x} \sim f(0, \boldsymbol{R})$$

Where $f(m, \boldsymbol{V})$ denotes a general distribution with mean m and variance $\boldsymbol{V}$ with

$\boldsymbol{R}$ being the LD matrix.

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 SNP 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 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. Heritability can then be expressed as

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

Now if we consider the covariance between SNP_i ($\boldsymbol{x_i}$) and $\boldsymbol{y}$, we have

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

As both $\boldsymbol{x}$ and $\boldsymbol{y}$ are standardized, the covariance equals to the correlation and we can define the correlation between SNP_i and Y as

$$\rho_i = \sum_j R_{ij}\boldsymbol{\beta}_j \tag{2.12}$$

In reality, the *observed* correlation contains errors. Therefore we define the *observed* correlation between SNP_i and the phenotype to be:

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

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

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

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

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

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

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

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

$$\begin{aligned}\chi_i^2 &= \hat{\rho_i}^2 n \\ &= n(\sum_j R_{ij}\boldsymbol{\beta}_j + \frac{\epsilon_i}{\sqrt{n}})^2\end{aligned}$$

We have

$$\begin{aligned}\mathrm{E}(\chi^2) &\approx n\boldsymbol{R_i^t H R_i} + 1 \\ &= n\sum_j R_{ij}^2 h_i^2 + 1\end{aligned}$$

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

$$\sum_i(\chi_i^2 - \mathrm{E}(\chi_i^2))^2 = \sum_i(\chi_i^2 - (n\sum_j R_{ij}^2\hat{h}_i^2 + 1))^2$$

If we define

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

we get

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

where $\boldsymbol{R_{sq}} = \boldsymbol{R} \circ \boldsymbol{R}$ and $\circ$ denotes the element-wise product (Hadamard product). By differentiating eq. (2.15) with respect to $\hat{h}$ and set to 0, we get

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

the SNP heritability is then defined as

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

where the $\mathbf{1}^t$ are multiplied to $\boldsymbol{R_{sq}^{-1}f}$ to get the sum of the vector $\hat{\boldsymbol{h}}$.

2.2.2 Calculating the Standard error

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

$$\mathrm{Var}(\hat{Heritability}) = \mathbf{1}^t \boldsymbol{R}_{\boldsymbol{sq}}^{-1} \mathrm{Var}(\boldsymbol{f}) \boldsymbol{R}_{\boldsymbol{sq}}^{-1} \mathbf{1} \tag{2.18}$$

Therefore, to obtain the variance of $\hat{Heritability}$, the variance covariance matrix of $\boldsymbol{f}$ is necessary.

If we consider the standardized genotype x_i has a standard normal mean z_i and non-centrality parameter μ_i, we have

$$\begin{aligned}
\mathrm{E}[x_i] &= \mathrm{E}[z_i + \mu_i] \\
&= 0 \\
\mathrm{Var}(x_i) &= \mathrm{E}[(z_i + \mu_i)^2] + \mathrm{E}[(z_i + \mu_i)]^2 \\
&= \mathrm{E}[z_i^2 + \mu_i^2 + 2 z_i \mu_i] + \mu_i^2 \\
&= 1 \\
\mathrm{Cov}(x_i, x_j) &= \mathrm{E}[(z_i + \mu_i)(z_j + \mu_j)] - \mathrm{E}[z_i + \mu_i]\mathrm{E}[z_j + \mu_j] \\
&= \mathrm{E}[z_i z_j + z_i \mu_j + \mu_i z_j + \mu_i \mu_j] - \mu_i \mu_j \\
&= \mathrm{E}[z_i z_j] + \mathrm{E}[z_i \mu_j] + \mathrm{E}[z_j \mu_i] + \mathrm{E}[\mu_i \mu_j] - \mu_i \mu_j \\
&= \mathrm{E}[z_i z_j]
\end{aligned}$$

As the genotypes are standardized, $\mathrm{Cov}(x_i, x_j) = \mathrm{Cor}(x_i, x_j)$, 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:

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

As $\mathrm{E}[z_i z_j] = R_{ij}$,

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

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 is represented as

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

After some tedious algebra, we can get

$$\mathrm{Var}(H) = \mathbf{1}^t \boldsymbol{R}_{\boldsymbol{sq}}^{-1} \frac{2\boldsymbol{R}_{\boldsymbol{sq}} + 4\boldsymbol{R} \circ \boldsymbol{z}\boldsymbol{z}^t}{n^2} \boldsymbol{R}_{\boldsymbol{sq}}^{-1} \mathbf{1} \tag{2.20}$$

where $\boldsymbol{z} = \sqrt{\boldsymbol{\chi}^{\mathbf{2}}}$ from eq. (2.14), with the direction of effect as its sign.

The problem with eq. (2.20) is that it requires the direction of effect. Without the direction of effect, the estimation of standard error (SE) will be inaccurate. As $n \times \boldsymbol{f} + 1$ is approximately χ^2 distributed, we might view eq. (2.16) as a decomposition of a vector of χ^2 distributions with degree of freedom of 1. Replacing the vector $\boldsymbol{f}$ with a vector of 1, we will be able to calculate the "effective number"(e) of the association (M.-X. X. Li et al., 2011). Substituting e into the variance equation

of non-central χ^2 distribution will yield

$$\mathrm{Var}(H) = \frac{2(e + 2H)}{n^2} \tag{2.21}$$

Theoretically, eq. (2.21) should provide a heuristic estimation of the SE without requiring the direction of effect. This reduces the number of input required from users.

2.2.3 Case Control Studies

The estimation of heritability in case control studies requires the correction of ascertainment bias. Therefore, estimates from eq. (2.17) must be adjusted in order to obtain an accurate estimation of SNP heritability for case control studies.

Based on the derivation of Jian Yang, Naomi R. Wray, and Peter M. Visscher (2010), the approximate ratio between the summary statistic obtained from case control studies (χ^2_{CC}) and quantitative trait studies(χ^2_{QT}) is

$$\frac{\chi^2_{CC}}{\chi^2_{QT}} = \frac{i^2 v(1-v)}{(1-K)^2} \tag{2.22}$$

where

$K =$ Population Prevalence

$v =$ Proportion of Cases

$i = \dfrac{z}{K}$

$z =$ height of standard normal curve at truncation pretained to K

Therefore we can transform the summary statistic from a case control study Using this approximation, we can directly transform the summary statistic between the case control studies and quantitative trait studies. As we are not interested

in transforming the non-centrality parameter (NCP) between two different studies, the sample size of the case control study (N_{CC}) and sample size of the quantitative trait study (N_{QT}) will be the same in eq. (2.22), therefore eq. (2.22) becomes

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

By combining eq. (2.23) and eq. (2.14), we can then have

$$f = \frac{(\chi^2_{CC} - 1)}{n} \frac{(1-K)^2}{i^2 v(1-v)} \tag{2.24}$$

where χ^2_{CC} is the test statistic from the case control association test. As eq. (2.24) is only eq. (2.14) multiply with the constant $\frac{(1-K)^2}{i^2 v(1-v)}$, the heritability estimation of case control studies can be simplified to

$$\hat{Heritability} = \frac{(1-K)^2}{i^2 v(1-v)} \mathbf{1}^t \boldsymbol{R}_{\boldsymbol{sq}}^{-1} \boldsymbol{f} \tag{2.25}$$

2.2.4 Extreme Phenotype Sampling

GWAS provides unprecedented power to perform hypothesis free association throughout the whole genome. However, with a limited budgets, studies often struggle to obtain sufficient sample size. To increase the detection power of the study, sampling strategies such as extreme phenotype selection can be applied. By performing extreme phenotype sampling, the frequency distortion between samples from the two extreme end of phenotype will be inflated, thus increases the statistical power (Guey et al., 2011). So for example, by including only the samples from the top 5% and bottom 5% of the phenotype distribution, the power of the detection is the same as a study with random sampling design that has 4 times the sample size (Pak C Sham and Shaun M Purcell, 2014). This allows studies to be conducted using a smaller sample size with the same degree of power, therefore reducing the cost.

The extreme selection design increases the summary statistic by a factor of

$\frac{V_{P'}}{V_P}$ where $V_{P'}$ is the trait variance of the selected sample and V_P is the trait variance of the general population (Pak C Sham and Shaun M Purcell, 2014). Thus, to adjust for the inflation, $\frac{V_P}{V_{P'}}$ is multiplied to eq. (2.14)

$$\hat{Heritability} = \frac{V_P}{V_{P'}} \mathbf{1}^t \boldsymbol{R}_{\boldsymbol{sq}}^{-1} \boldsymbol{f} \tag{2.26}$$

2.2.5 Inverse of the Linkage Disequilibrium matrix

The SNP heritability can be estimated from eq. (2.17) which calculates the sum of $\hat{\boldsymbol{h}^2}$ from eq. (2.16). When $\boldsymbol{R}_{\boldsymbol{sq}}$ is full rank and positive definite, eq. (2.16) can be solved using the QR decomposition or LU decomposition without explicitly calculating the inverse of $\boldsymbol{R}_{\boldsymbol{sq}}$.

However, LD matrices are usually ill-conditioned. Therefore the solution of eq. (2.16) is prone to large numerical errors (Neumaier, 1998). Therefore, in order to solve for eq. (2.16), regularization techniques such as Tikhonov Regularization (also known as Ridge Regression) and Truncated Singular Value Decomposition (tSVD) has to be performed (Neumaier, 1998). Herein, we focus on the use of tSVD in the regularization of the LD matrix.

Given the matrix equation $\boldsymbol{Ax} = \boldsymbol{B}$ where $\boldsymbol{A}$ is ill-conditioned or singular with $n \times n$ dimension. The Singular Value Decomposition (SVD) of $\boldsymbol{A}$ can be expressed as

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^t \tag{2.27}$$

where $\boldsymbol{U}$ and $\boldsymbol{V}$ are both orthogonal matrix and $\boldsymbol{\Sigma} = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n)$ is the diagonal matrix of the *singular values* (σ_i) of matrix $\boldsymbol{A}$. Based on eq. (2.27), the

inverse of $\boldsymbol{A}$ can be expressed as

$$\boldsymbol{A}^{-1} = \boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^{t} \tag{2.28}$$

Where $\boldsymbol{\Sigma}^{-1} = \text{diag}(\frac{1}{\sigma_1}, \frac{1}{\sigma_2}, \ldots, \frac{1}{\sigma_n})$.

Now consider the vector $\boldsymbol{B}$ is collected with some error $\boldsymbol{\epsilon}$ attached to it. The solution to $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{B}$ becomes:

$$\begin{aligned} \boldsymbol{x} &= \boldsymbol{A}^{-1}(\boldsymbol{B} + \boldsymbol{\epsilon}) \\ &= \boldsymbol{A}^{-1}\boldsymbol{B} + \boldsymbol{A}^{-1}\boldsymbol{\epsilon} \\ &= \boldsymbol{x}^{*} + \boldsymbol{A}^{-1}\boldsymbol{\epsilon} \end{aligned} \tag{2.29}$$

where $\boldsymbol{x}^{*}$ is the true solution. The error of the solution $\delta\boldsymbol{x}$ caused by the error in the data is therefore:

$$\begin{aligned} \delta\boldsymbol{x} &= \boldsymbol{x} - \boldsymbol{x}^{*} \\ &= \boldsymbol{A}^{-1}\boldsymbol{\epsilon} \end{aligned} \tag{2.30}$$

The ratio of relative error in the solution to the relative error in the data is then defined as:

$$\begin{aligned} \frac{||\delta\boldsymbol{x}||}{||\boldsymbol{x}||}\,\frac{||\boldsymbol{\epsilon}||}{||\boldsymbol{B}||} &= \frac{||\delta\boldsymbol{x}||}{||\boldsymbol{\epsilon}||}\frac{||\boldsymbol{B}||}{||\boldsymbol{x}||} \\ &= \frac{||\boldsymbol{A}^{-1}\boldsymbol{\epsilon}||}{||\boldsymbol{\epsilon}||}\frac{||\boldsymbol{A}\boldsymbol{x}||}{||\boldsymbol{x}||} \\ &= ||\boldsymbol{A}^{-1}||\;||\boldsymbol{A}|| \end{aligned} \tag{2.31}$$

where $||.||$ is the matrix norm. When l_2-norm is used, the condition number of matrix $\boldsymbol{A}$ ($\kappa(\boldsymbol{A})$) can then be defined as

$$\kappa(\boldsymbol{A}) = \frac{\sigma_{max}(\boldsymbol{A})}{\sigma_{min}(\boldsymbol{A})}$$

Thus, a large $\kappa(\boldsymbol{A})$ means that small errors in the data will lead to large derivation in the solution.

To obtain a meaningful solution from this ill-conditioned/singular matrix $\boldsymbol{A}$, the tSVD method can be performed to obtain a pseudo inverse of $\boldsymbol{A}$. Similar to eq. (2.27), the tSVD of $\boldsymbol{A}$ can be represented as

$$\boldsymbol{A}^{+} = \boldsymbol{U}\boldsymbol{\Sigma}_k\boldsymbol{V}^t \qquad \text{and} \qquad \boldsymbol{\Sigma}_k = \text{diag}(\sigma_1, \ldots, \sigma_k, 0, \ldots, 0) \tag{2.32}$$

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

$$\sigma_i = \begin{cases} \sigma_i & \text{for} \quad \sigma_i \geq t \\ 0 & \text{for} \quad \sigma_i < t \end{cases} \tag{2.33}$$

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

By selecting an appropriate t, tSVD can effectively regularize the ill-conditioned matrix and help to find a reasonable approximation to $\boldsymbol{x}$. 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.

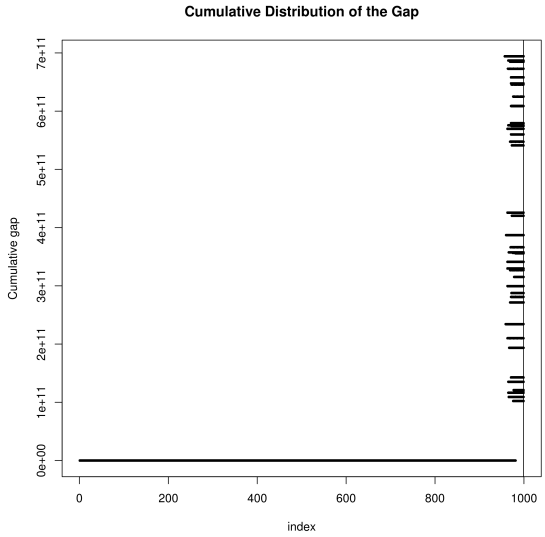
The easiest way to test if the matrix $\boldsymbol{A}$ has well-defined rank by calculating the "gap" in the singular values:

$$gap = \sigma_k/\sigma_{k+1} \tag{2.34}$$

a large gap indicates that the matrix has a well-defined rank.

Simulation was carried out to investigate whether if LD matrix has a well defined rank. 1,000 samples were randomly simulated from the HapMap (Altshuler et al., 2010) Northern Europeans from Utah (CEU) population with 1,000 SNPs

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. The results suggest that the rank of the LD matrix is well defined



randomly select from chromosome 22 using HAPGEN2 (Su, Marchini, and Donnelly, 2011). HAPGEN2 allow for the simulation of samples with LD structure similar to that observed in the reference panel. The LD matrix and its corresponding singular values were computed. 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 shown that the LD matrix has a well-defined rank with a mean maximum "gap" of 466,198,939,298. Therefore the choice of tSVD for the regularization is appropriate.

In view of this, tSVD was selected as the method for regularization for solving eq. (2.16). MATLAB, NumPy and GNU Octave defined the threshold for tSVD as $t = \epsilon \times \max(m, n) \times \max(\sigma)$ where ϵ is the machine epsilon (the smallest number a machine define as non-zero), , n is the number of rows and m is the number of columns. Here, the same threshold definition was used in our algorithm.

2.2.6 Comparing Different LD correction Algorithms

An important consideration in our algorithm is the sampling error in LD. In reality, the population LD matrix is not available. Therefore , the LD matrix must be estimated from various reference panels such as the 1000 genome project (Project et al., 2012) or the HapMap project (Altshuler et al., 2010). Given these reference panels are subsets of the whole population, this results in sampling errors in the estimated sample LD. The sample LD can then be represented as:

$$\hat{R} = R + \epsilon$$

where R is the population LD and ϵ is the sampling error which is unbiased. However, in eq. (2.17), squared LD are required. The expected value of the LD squared (R^2) is then calculated as

$$\begin{aligned} \hat{\mathrm{R}^2} &= \mathrm{E}[(R+\epsilon)^2] \\ &= \mathrm{E}(R^2 + 2R\epsilon + \epsilon^2) \\ &= \mathrm{E}(R^2) + \mathrm{E}(\epsilon^2) \end{aligned} \tag{2.35}$$

A positive bias is observed in the sample R^2.

Weir and W G Hill (1980) and Z. Wang and Thompson (2007) proposed methods for the correction of sample R^2:

$$\text{Ezekiel} : \tilde{R^2} = 1 - \frac{n-1}{n-2}(1-\hat{R^2}) \tag{2.36}$$

$$\text{Olkin-Pratt} : \tilde{R^2} = 1 - \frac{(n-3)(1-\hat{R^2})}{n-2}(1+\frac{2(1-\hat{R^2})}{n}) \tag{2.37}$$

$$\text{Pratt} : \tilde{R^2} = 1 - \frac{(n-3)(1-\hat{R^2})}{n-2}(1+\frac{2(1-\hat{R^2})}{n-3.3}) \tag{2.38}$$

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

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

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

In order to assess the performance of each individual correction methods, simulations were performed. Firstly, 5,000 SNPs with minor allele frequency (maf) ≥ 0.1 were randomly selected from chromosome 22 from the 1000 genome 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 allows for the simulation of 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 could be used to approximate the genetic architecture of adaptation. As a result, effect sizes were simulated with an exponential distribution with $\lambda = 1$:

$$\begin{aligned} \theta &= \exp(\lambda = 1) \\ \beta &= \pm\sqrt{\frac{\theta \times h^2}{\sum \theta}} \end{aligned} \tag{2.41}$$

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

Using the normalized genotype matrix of the causal SNPs of all individuals ($\boldsymbol{X}$) and the vector of effect sizes ($\boldsymbol{\beta}$), the phenotype were simulated with heritability of h^2 using

$$\begin{aligned} \epsilon_i &\sim N(0, \mathrm{Var}(\boldsymbol{X\beta})\frac{1-h^2}{h^2}) \\ \boldsymbol{\epsilon} &= (\epsilon_1, \epsilon_2, ..., \epsilon_n)^t \\ \boldsymbol{y} &= \boldsymbol{X\beta} + \boldsymbol{\epsilon} \end{aligned} \tag{2.42}$$

To simulate the whole spectrum of heritability, the h^2 were varied from 0 to 0.9 with increment of 0.1.

The association between the genotype and phenotype were then calculated using PLINK (Shaun Purcell et al., 2007). Heritability estimation were then performed based on the resulting summary statistics using different LD correction algorithms. An independent 500 samples, which corresponds 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, the raw sample genotypes were unavailable 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 the common usage of the algorithm.

The whole process were repeated 50 times such that a distribution of the estimates can be obtained. In summary, the following simulation procedure was performed:

1. Randomly select 5,000 SNPs with maf$>$ 0.1 from chromosome 22
2. Simulate 500 samples using HAPGEN2 and used as the reference panel
3. Randomly generate 100 effect sizes with eq. (2.41)
4. Randomly assign the effect sizes 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. (2.42)
6. Perform heritability estimation using our algorithm with different LD correction algorithm
7. Repeat step 5-6 50 times

2.2.7 Comparison with Other Algorithms

After identifying the optimal LD correction algorithm, it is important to compare the performance of our algorithm to the existing methods. Therefore, simulations were performed where quantitative and binary traits were simulated with different genetic architectures. The effect of different sampling strategies, such as random sampling and extreme phenotype selection, on the heritability estimation were also investigated.

Currently, the only other algorithm that is capable to estimate the SNP-heritability using summary statistics from GWAS is the LDSC (B. K. Bulik-Sullivan et al., 2015). Whereas GCTA (J Yang et al., 2011) is the most commonly used programme for the estimation of SNP-heritability from GWAS data when the raw genotypes are available. Therefore, the performance of our algorithm was compared to LDSC and GCTA. As no confounding factors were simulated, the intercept estimation function in LDSC will be penalized with a larger SE. Thus, performance of LDSC with fixed intercept (--no-intercept) were also inspected to avoid bias against LDSC.

Sample Size

The sample size is the most important parameter in determining the SE of the estimates. As sample size increases, the samples will be more representative of the true population and will provide a more accurate estimation of the parameters, therefore results in a smaller SE. Using simple text mining, the sample size distribution of GWAS was obtained from the GWAS catalog (Welter et al., 2014). The average sample size was 7,874, with a median count of 2,506 and a lower quartile at 940. We argued that if the algorithms performed well with a small sample size (e.g. 1,000 samples), their performance should improve as sample size increases. Thus, to re-

duce the computation time required for the simulation, only 1,000 samples were simulated in each simulations unless otherwise stated.

Number of SNPs in Simulation

Although the SNP heritability can be estimated in a short amount of time, the total time required to complete the whole simulation quickly become infeasible as the number of iterations and conditions increases. As a result, to reduce the simulation time, only 50,000 SNPs from chromosome 1 were used for simulation, which correspond to 200 SNPs within a 1 megabase (mb) region.

Genetic Architecture

The LD pattern, the number of causal SNPs, the effect size of the causal SNPs and the heritability of the trait are all important factors contributing to the genetic architecture of a trait.

First and foremost, in order to investigate the performance of the SNP heritability estimation algorithms, traits with different heritability have to be considered. Therefore, traits with heritability ranging from 0 to 0.9 with increment of 0.1 were simulated.

Secondly, to obtain a realistic LD pattern, genotypes were simulated using the HAPGEN2 programme (Su, Marchini, and Donnelly, 2011), using the 1000 genome CEU haplotypes as an input. As GWAS usually lack power in detecting rare variants (e.g. maf < 0.05), SNPs with maf < 0.05 were excluded.

Thirdly, to investigate the performance of the algorithms with a different number of causal SNPs (k), the number of causal SNPs were varied with $k \in \{5, 10, 50, 100, 500\}$. The effect sizes were then simulated using eq. (2.41) and the phenotype were simulated using eq. (2.42).

For GCTA, the sample genotypes were provided to calculate the genetic relationship matrix. Sample phenotypes were also provided for GCTA to estimate the SNP 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 LD matrix. The association between the genotype and phenotype were calculated using PLINK (Shaun Purcell et al., 2007). The summary statistics and the reference panel were then provided for LDSC and our algorithm to estimate the SNP heritability. This simulation procedure should provide a realistic representation of the common usage of the algorithms.

For each population, the whole process were repeated 50 times such that a distribution of the estimate can be obtained. In total, 10 independent populations were simulated. In summary, the simulation follows the following procedures:

1. Randomly select 50,000 SNPs with maf>0.05 from chromosome 1
2. Simulate 500 samples using HAPGEN2 to be served as a reference panel
3. Randomly generate k effect size with $k \in \{5, 10, 50, 100, 500\}$ following eq. (2.41), 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. (2.42)
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

Extreme Effect Size

It is possible for a trait to have SNPs that account for a larger portion of the heritability. For example, the deleterious mutations on *RET* account for $\approx 50\%$ of the familial cases of the Hirschsprung's disease yet some of the heritability was still missing. Gui et al. (2013) therefore suggested that there might be more variants with small effects that have not been identified.

To simulate extreme effect size, 100 causal SNPs were simulated where m of those account for 50% of all the effect sizes with $m \in \{1, 5, 10\}$. The effect sizes were then calculated as

$$\begin{aligned} \beta_{eL} &= \pm\sqrt{\frac{0.5h^2}{m}} \\ \beta_{eS} &= \pm\sqrt{\frac{0.5h^2}{100-m}} \\ \beta &= \{\beta_{eL}, \beta_{eS}\} \end{aligned} \tag{2.43}$$

The effect sizes were then randomly assigned to 100 causal SNPs and phenotype were calculated using eq. (2.42). The following simulation procedure were then performed:

1. Randomly select 50,000 SNPs with maf > 0.05 from chromosome 1
2. Simulate 500 samples using HAPGEN2 and used as the reference panel
3. Randomly generate 100 effect size where m has extreme effect, following eq. (2.43), 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. (2.42)
6. Perform heritability estimation using our algorithm, LDSC with fixed inter-

cept, LDSC with intercept estimation and GCTA

7. Repeat step 5-6 50 times

8. Repeat step 1-7 10 times

Case Control Studies

However, there were two additional parameters to consider: the population prevalence and the observed prevalence. To acquire sufficient samples for the simulation under the liability model Although there were only two additional parameter, it is significantly more challenging to simulate a binary trait. It is mainly because of the number of samples required to simulate under the liability threshold model. Take for example, if one would 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 (e.g. 1%). 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 significantly changing the SNP density. We acknowledge that the current simulation was relatively brief, however, it should serves as a proof of concept simulation to study the performance of the algorithms under the case control scenario.

In the case control simulation, we randomly selected 5,000 SNPs from

chromosome 22 with maf ≥ 0.05 in the CEU haplotypes as an input to HAPGEN2. We then randomly selected k SNPs where $k \in \{10, 50, 100, 500\}$, each with effect size simulated based on eq. (2.41). In order to simulate a case control samples with 1,000 cases, we then simulated $\frac{1{,}000}{p}$ samples and calculate their phenotype using eq. (2.42). 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> 0.05 from chromosome 22
2. Simulate 500 samples using HAPGEN2 and used as a reference panel
3. Randomly generate k effect size following eq. (2.41) where $k \in \{10, 50, 100, 500\}$
4. Randomly assign the effect size to k SNPs
5. Simulate $\frac{1{,}000}{p}$ samples using HAPGEN2 and calculate their phenotype according to eq. (2.42)
6. Define case control status using the liability threshold and randomly select the same number of case and controls for statistic analysis
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

Extreme Phenotype Sampling

With a limited budget, it is usually difficult to obtain adequate sample size for a GWAS, leading to studies with insufficient power. One possible approach is to perform the extreme phenotype sampling that only select samples with phenotypes on the extreme end of the distribution. By performing extreme phenotype sampling, a smaller sample size is required to achieve the same detection power (Pak C Sham and Shaun M Purcell, 2014), therefore reducing the cost of the study. The extreme phenotype sampling can lead to inflation in the summary statistics (Guey et al., 2011) and might therefore affect the heritability estimation of LDSC and our algorithm.

Herein, simulation was performed to investigate the effect of extreme phenotype sampling on the performance of the algorithms. 50,000 SNPs with maf > 0.05 were selected from chromosome 1 and were used as an input for HAPGEN2. Again, 500 samples were first simulated to serve as the reference panel for LDSC and our algorithm.

From the 50,000 SNPs, 100 SNPs were randomly selected as the causal SNPs and their effect sizes were simulated using eq. (2.41). $\frac{1000}{K\times 2}$ samples were then simulated where K is the portion of samples selected from the most extreme end of the distribution (e.g. 0.1 or 0.2). Phenotype of the individuals were then simulated using eq. (2.42) and were standardized. 500 samples were selected at both end of the phenotype distribution. To compare the effect of extreme phenotype sampling and random sampling strategies on the performance of the algorithms, 1,000 samples were randomly drawn from the $\frac{1000}{K\times 2}$ samples. Sample genotypes and phenotypes were provided to GCTA for the estimation of the SNP heritability, whereas for LDSC and our algorithm, the summery statistic generated calculated by PLINK (S Purcell, Cherny, and P C Sham, 2003) and the reference panel were provided.

It was noted that the extreme phenotype sampling were not supported by the LDSC and GCTA. To allow for a fair comparison, extreme phenotype adjustment from Pak C Sham and Shaun M Purcell (2014) were applied to the estimates from LDSC and GCTA. Finally, the heritability estimated based on different sampling strategies were compared. For each population, the whole process were repeated 50 times. In total, 10 independent populations were simulated. In summary, the following simulation procedures were used:

1. Randomly select 50,000 SNPs with maf$>$ 0.05 from chromosome 1
2. Simulate 500 samples using HAPGEN2 and used as the reference panel
3. Randomly generate 100 effect size following eq. (2.41), with heritability ranging from 0 to 0.9 (increment of 0.1)
4. Randomly assign the effect sizes to 100 SNPs
5. Simulate $\frac{1{,}000}{K\times 2}$ samples using HAPGEN2 where K is the portion of samples selected from the extreme end of the distribution with $K \in \{0.1, 0.2\}$
6. Phenotype of the samples were calculated according to eq. (2.42) and were standardized
7. Top 500 and bottom 500 samples (ranked by phenotype) were selected, representing the extreme phenotype sample selection strategy
8. 1,000 samples were also randomly selected to represent the general random sampling strategy
9. Perform heritability estimation using our algorithm, GCTA, LDSC with fixed intercept and LDSC with intercept estimation.
10. Adjust the estimation from LDSC and GCTA by the extreme phenotype adjustment factor as proposed by Pak C Sham and Shaun M Purcell (2014)

11. Repeat step 5-10 50 times

12. Repeat step 1-11 10 times

2.2.8 Application to Real Data

To demonstrate our algorithm also works outside of simulated data, we also estimated the heritability of schizophrenia and other psychiatric disorders using the Psychiatric Genomics Consortium (PGC) datasets (Stephan Ripke, B. M. Neale, et al., 2014; Psychiatric GWAS Consortium Bipolar Disorder Working Group, 2011; Stephan Ripke, Naomi R Wray, et al., 2013). LDSC were also used alongside our algorithm to serves as a baseline comparison.

The reference genome were downloaded from 1000 genome (hg19) (Project et al., 2012) and were converted to PLINK binaries using the PLINK --vcf function. The European super population was extracted which contains a total of 503 samples. Singleton and multi-allelic SNPs were filtered out from the reference panel. Cryptic relatedness between samples can inflate the LD due to increased allele sharing amongst relatives. It is therefore important to filter out related samples. Genotypes were first pruned, then the identity by descent (IBD) between samples were calculate using the PLINK option --genome. Sample pairs with relatedness ≥ 0.125 ($\approx$ third degree relatedness) were removed. In total, 446 samples remained after quality control. The LD score was calculated based on the 446 samples using a 1 mb window size. SNPs with maf < 0.1 were filtered out by default.

The summary statistics were obtained from the PGC website. As SNPs in the bipolar and major depression data follows the old genomic annotations (hg18), liftover (Hinrichs et al., 2006) were performed to convert the genomic coordinates to genome version hg19. Due to difference in composition of the sex chromosome in male and female (e.g. XY in male, XX in female) and the lack of information

on the male to female ratio, it is difficult to estimate the SNPs heritability on the sex chromosomes. Therefore, the SNP heritability were only estimated using the autosomal SNPs.

As the datasets contain binary traits, the population prevalence of the trait has to be provided in order for the adjustment of the ascertainment bias. Based on B. K. Bulik-Sullivan et al. (2015) a population prevalence of 0.15 were selected for major depression disorder and 0.01 were selected for schizophrenia and bipolar disorder.

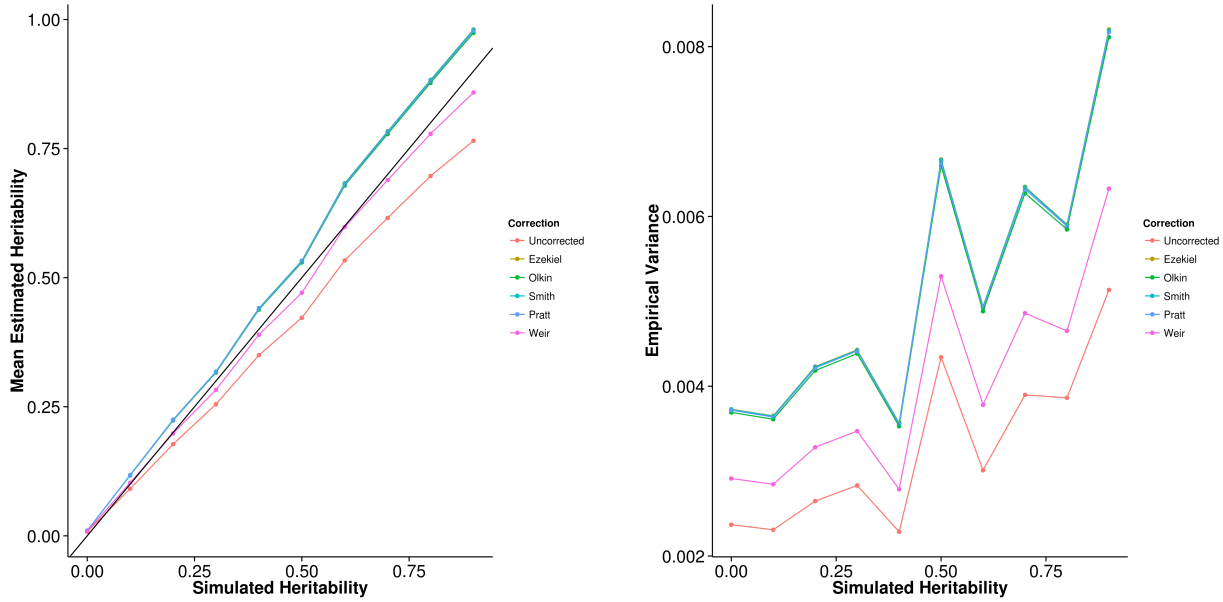
Unfortunately, because of the high SNP density of the PGC schizophrenia GWAS, the computational resources required to complete the SNP heritability estimation exceeds the current available resources. To facilitate the analysis, the distance between each bin was reduced to 50,000 base pair (bp) for our algorithm. This will results in an inflation in the final estimates. Therefore estimates from our algorithm can only serve as an upper bound of the true SNP heritability of schizophrenia.

2.3 Result

The heritability estimation were implemented in SHREK and is available on `https://github.com/choishingwan/shrek`.

2.3.1 LD Correction

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



(a) Mean Estimation **(b)** Empirical Variance

Figure 2.2: Effect of LD correction to Heritability Estimation. We compared the performance of our algorithm when different R^2 bias correction algorithm was used. When no bias correction was carried out, a downward bias was observed. After the application of the bias correction algorithms, the mean estimations of all except in the case of Weir eq. (2.40) algorithms leads to an overestimation of heritability. On the other hand, the corrections all lead to increase in variance of the estimation.

Different bias correction algorithms were applied and their performance was compared (fig. 2.2a). From the graph, it was observed that when no bias correction was applied, the mean estimation were in general downwardly biased. This was consistent with our expectation of a general upward bias in sample R^2 which will downwardly penalize the resulting heritability estimation. On the other hand, the bias correction algorithms all worked as expected where they increases the mean estimation of heritability because removal of the upward bias in the sample R^2 should increase the heritability estimation. However for most algorithms except for Weir's formula (eq. (2.40)) an over adjustment were observed, leading to a general upward bias in the estimation. Taking into account of the variance of estimation (fig. 2.2b), Weir's formula was the most suitable for SHREK where not only it reduces the bias in the final heritability estimation, it introduced the smallest amount of additional

variance to the estimates. As a result of that, we selected the Weir's formula as our default LD correction algorithm.

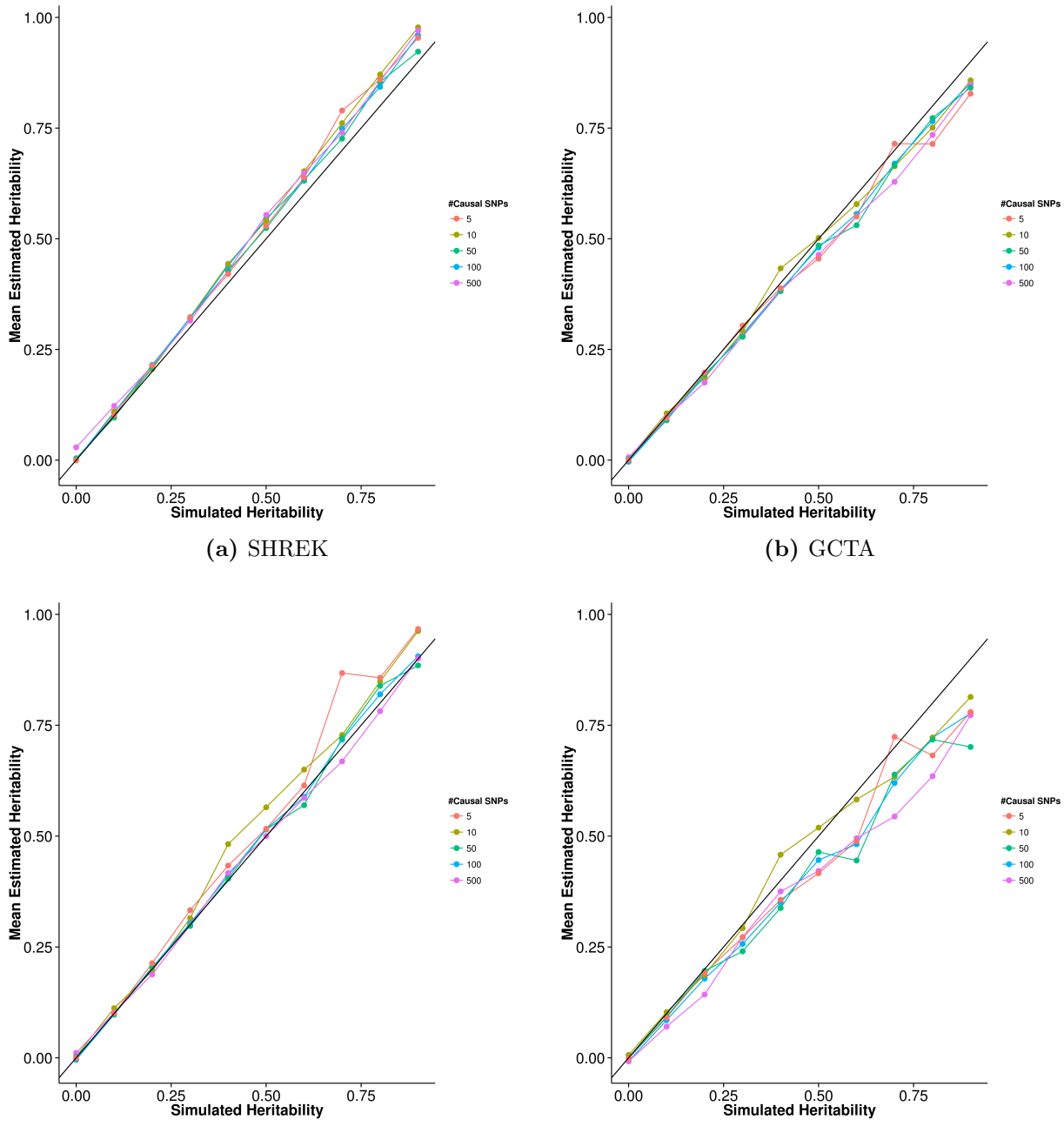
2.3.2 Comparing with Other Algorithms

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

Quantitative Trait Simulation

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

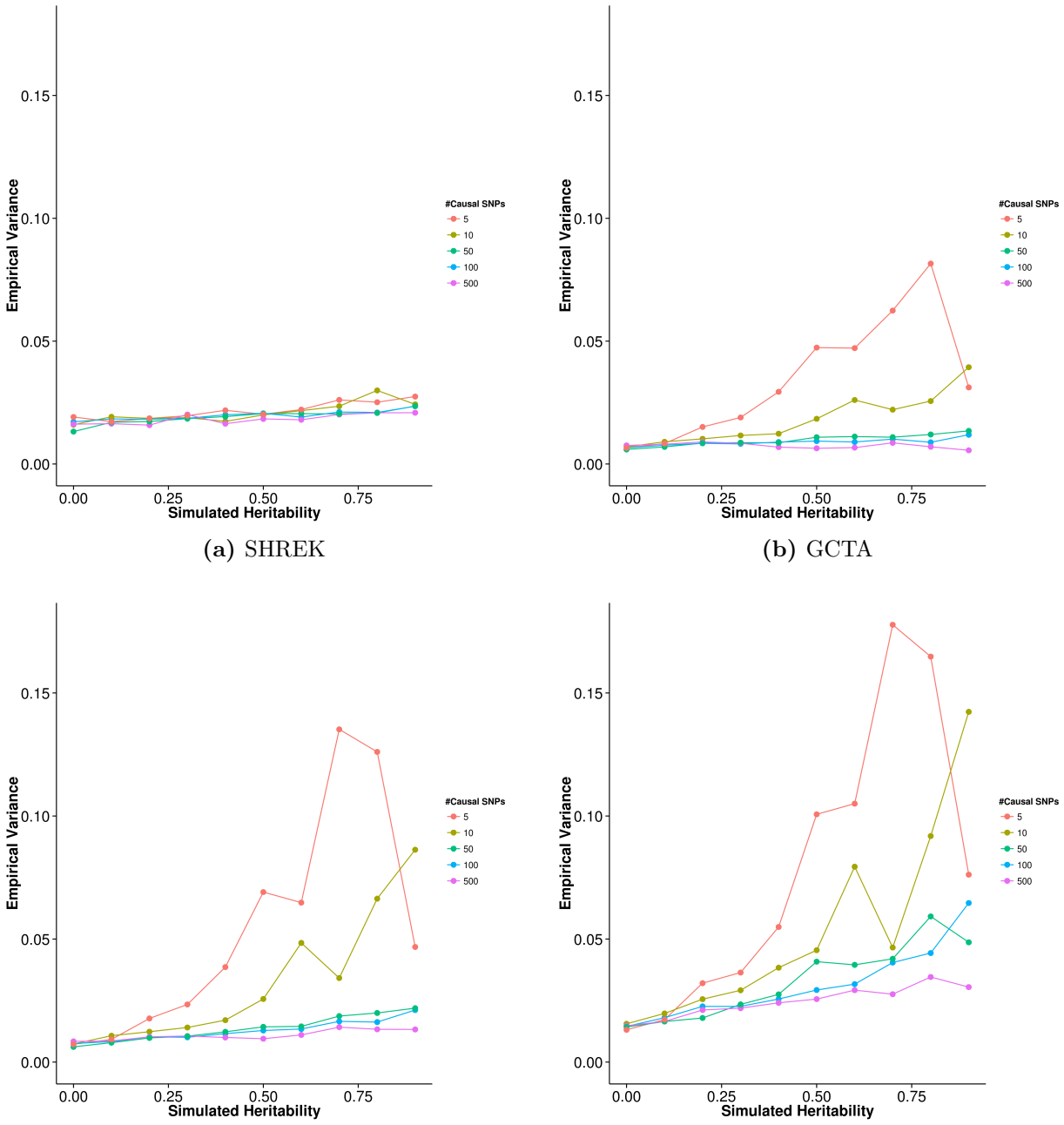
First, the mean heritability estimation were compared to the simulated heritability in order to identify the bias in estimation for each algorithms. From the graph (fig. 2.3), it was observed that the mean estimations of SHREK has a small upward bias (fig. 2.3a). However, the bias was insensitive to the change in number of causal SNPs suggesting that SHREK is relatively robust to trait complexity. On the other hand, estimations form GCTA were moderately biased downward (fig. 2.3b), similar to the estimations from LDSC with intercept estimation (fig. 2.3d), but with a smaller variability. Finally, when the intercept is fixed, LDSC has the smallest bias when the trait is polygenic but an upward bias is also observed when the number of causal SNPs is small.



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

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

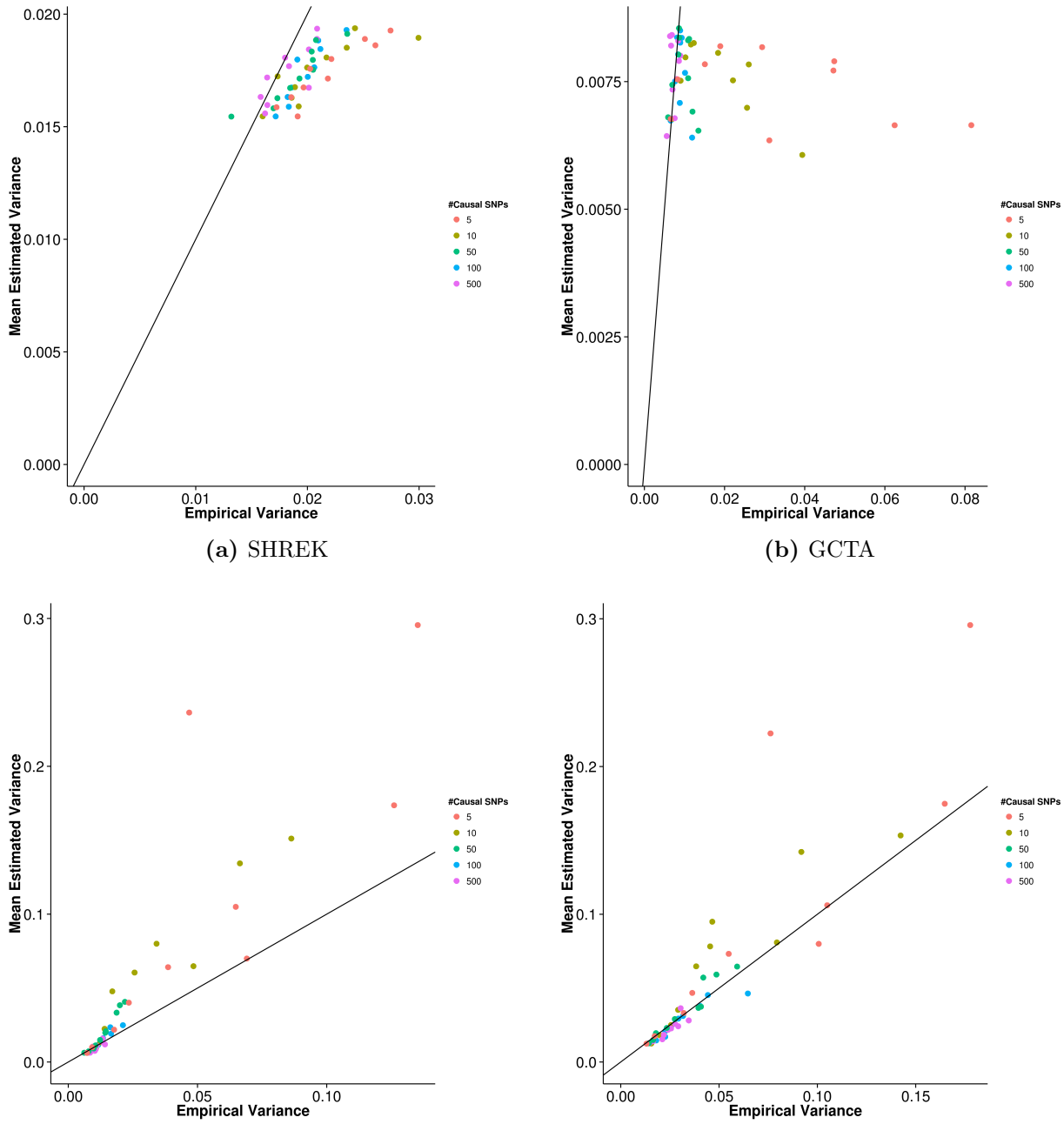
Figure 2.3: Mean of results from quantitative trait simulation with random effect size simulation. Estimations form SHREK were slightly biased upwards whereas GCTA and LDSC with intercept estimations both biased downwards. On the other hand, LDSC with fixed intercept provides least biased estimates under polygenic conditions. However, when the number of causal SNPs is small (e.g. 5 or 10), an upward bias was observed.



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

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

Figure 2.4: Variance of results from quantitative trait simulation with random effect size simulation. Under the polygenic conditions, GCTA has the smallest variance, follow by LDSC. However, it was observed when the number of causal SNPs decreases, the variance of the estimation increases for all algorithm, with variance of the SHREK estimate being the least affected. In fact, under oligogenic conditions, SHREK has a lower empirical variance when compared to LDSC.



(a) SHREK

(b) GCTA

(c) LDSC with fix intercept

(d) LDSC with intercept estimation

Figure 2.5: Estimated variance of results from quantitative trait simulation with random effect size simulation when compared to the empirical variance. GCTA has the best estimate of its empirical variance under the polygenic conditions whereas SHREK tends to under-estimate its empirical variance. On the other hand, LDSC tends to over-estimate the variance especially when the number of causal SNPs is small.

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

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

Furthermore, while comparing the empirical variance of the estimates (fig. 2.4), variance of the estimates from LDSC are sensitive to the number of causal SNPs. As the number of causal SNPs decreases (figs. 2.4c and 2.4d), the variance of LDSC estimates increases, similar to what was reported by B. K. Bulik-Sullivan et al. (2015). The variance are also higher when intercept estimation was performed. On the other hand, although the variance of SHREK is relatively higher when compared to LDSC when the intercept was fixed, the variation of its estimates is insensitive to the number of causal SNPs. When the number of causal SNPs was small, the variance of estimates from SHREK can even be lower than LDSC (fig. 2.4a). Finally, of all the algorithms, the estimates from GCTA has the lowest variation when compared to other algorithm (fig. 2.4b), except when only 5 causal SNPs were simulated where it has a slightly higher variance in comparison to SHREK when the simulated heritability was high (e.g. ≥ 0.8).

Another important factor to consider was the estimation of the SE. Of all the algorithms, GCTA (fig. 2.5b) has the best estimate, follow by SHREK (fig. 2.5a). However, it was noted that a consistent underestimation of variance is observed with SHREK whereas GCTA only underestimate the variance when the number of

causal SNPs is small. On the other hand, when the intercept was fixed (fig. 2.5c), LDSC cannot accurately estimate its variance and tends to overestimate, especially when the number of causal SNPs simulated were small. When intercept estimations was performed (fig. 2.5d), the estimation of variance is relatively better yet the overestimation are still observed when the number of causal SNPs is small.

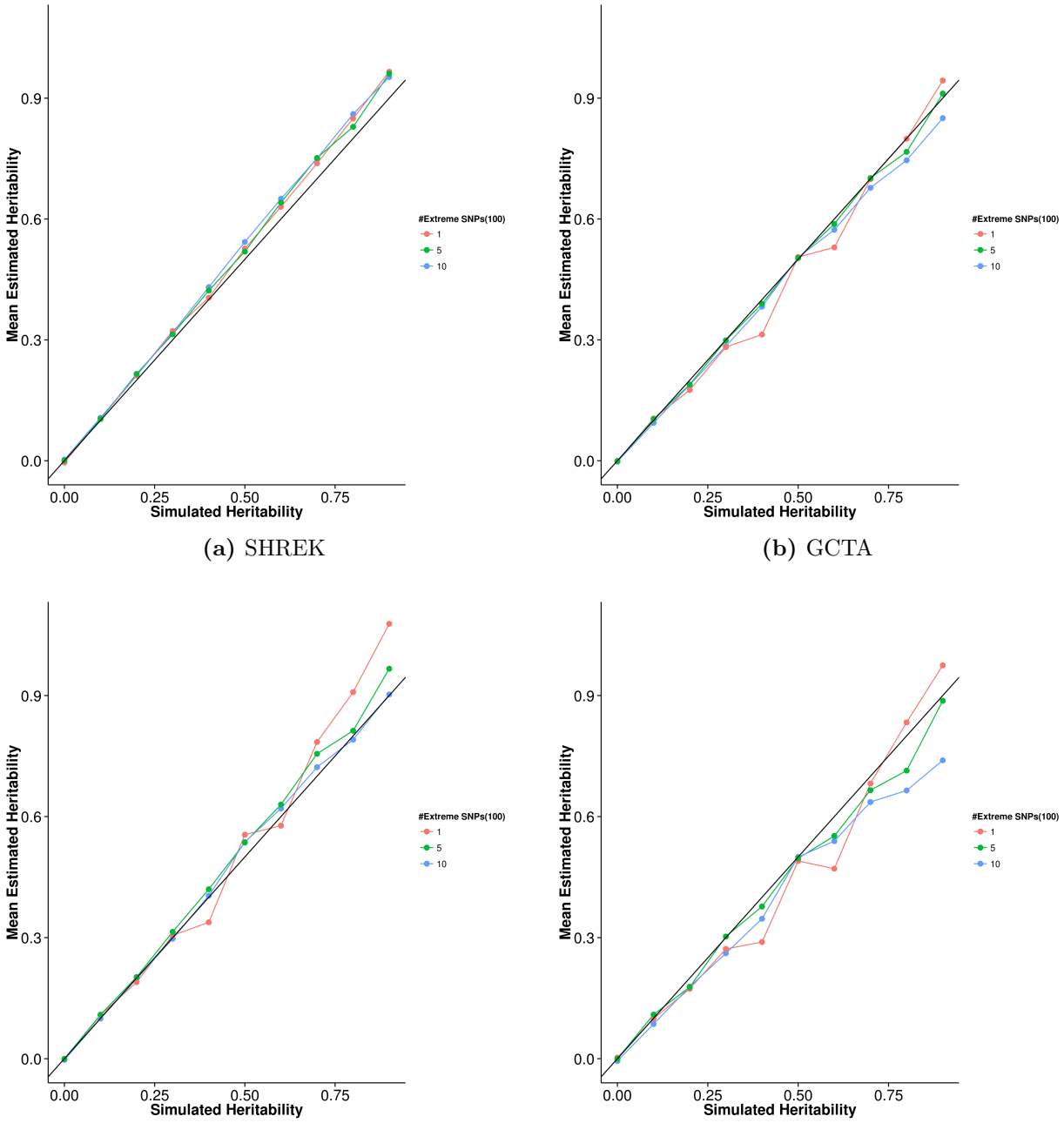
By taking into consideration of both the bias and variance of the estimates, GCTA has the best overall performance. Under the oligogenic condition (e.g. number of causal SNPs ≤ 10), SHREK has relatively better performance when compared to LDSC. Whereas under the polygenic condition, LDSC has better performance.

Quantitative Trait Simulation with Extreme Effect Size

For some diseases such as Hirschsprung's disease, a small number of SNPs can account for majority of the effect with a large number of SNPs with small effect size. Therefore we were interested to test the performance of SNP heritability estimation in such scenario. We performaned the quantitative trait simulation with 100 causal SNPs where 1,5 or 10 of those SNP(s) has a large effect.

When assessing the mean estimation of heritability (fig. 2.6), the performance of the algorithms are similar to that in the quantitative trait simulation. The only exception is when 1 SNP with large effect was simulated, the mean estimation of LDSC and GCTA fluctuates (figs. 2.6b to 2.6d). The same fluctuation is not observed in SHREK (fig. 2.6a). Similarly, the empirical variance of the estimation (fig. 2.7) from GCTA and LDSC increases and fluctuates when only 1 SNP with large effect was simulated. It is most obvious in the case of LDSC where the variance increased drastically as the heritability is high (fig. 2.7c). However, SHREK does not seems to be affected and are robust to the number of SNPs with large effect.

The estimated variance are also affected by the number of SNPs with large



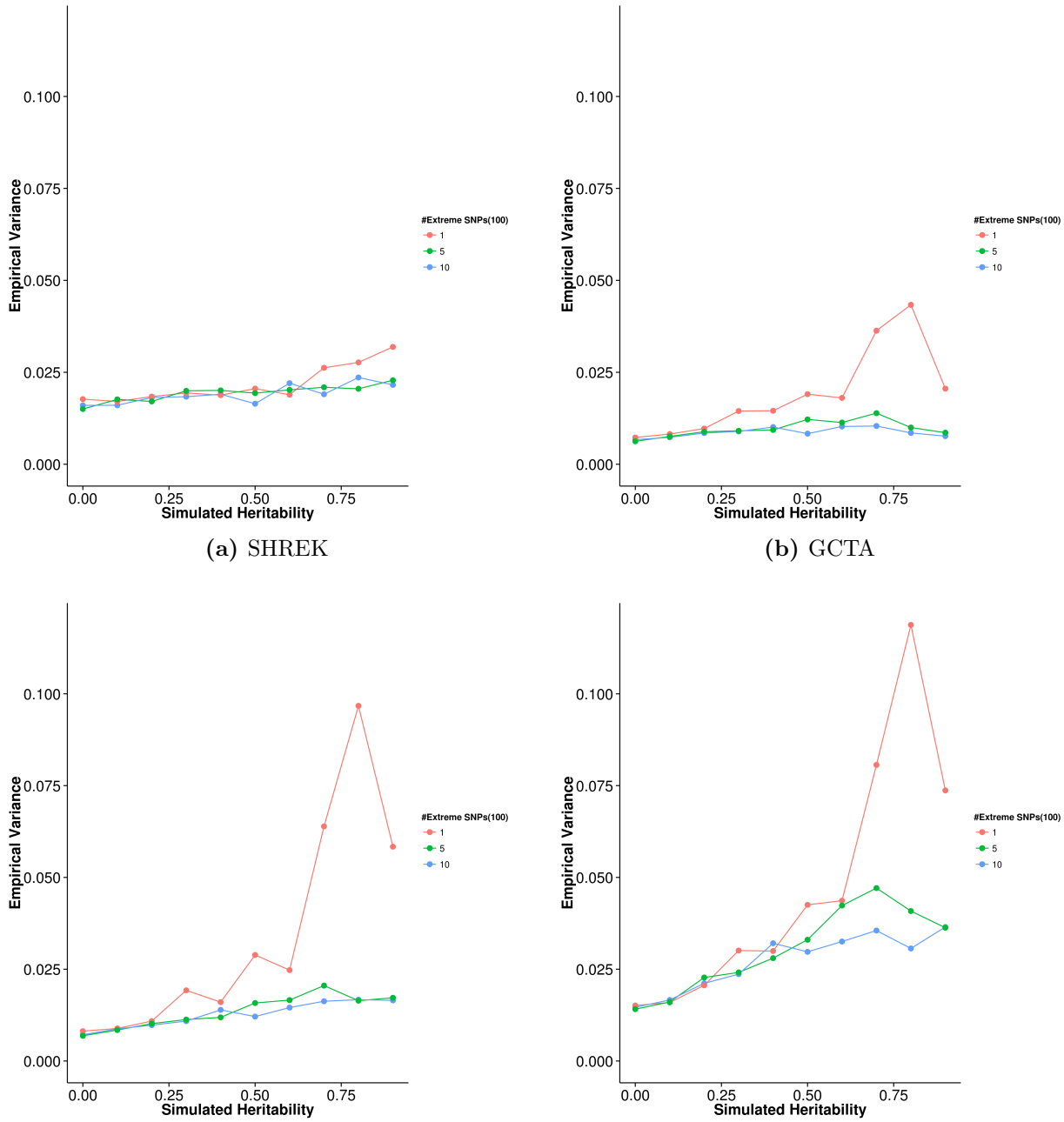
(a) SHREK

(b) GCTA

(c) LDSC with fix intercept

(d) LDSC with intercept estimation

Figure 2.6: Mean of results from quantitative trait simulation with extreme effect size simulation. It is observed that the mean estimation of heritability of SHREK is not affected by the number of SNP(s) with large effect but with slight upward bias. On the other hand, the mean estimation of LDSC and GCTA seems to fluctuate with respect to the simulated heritability.



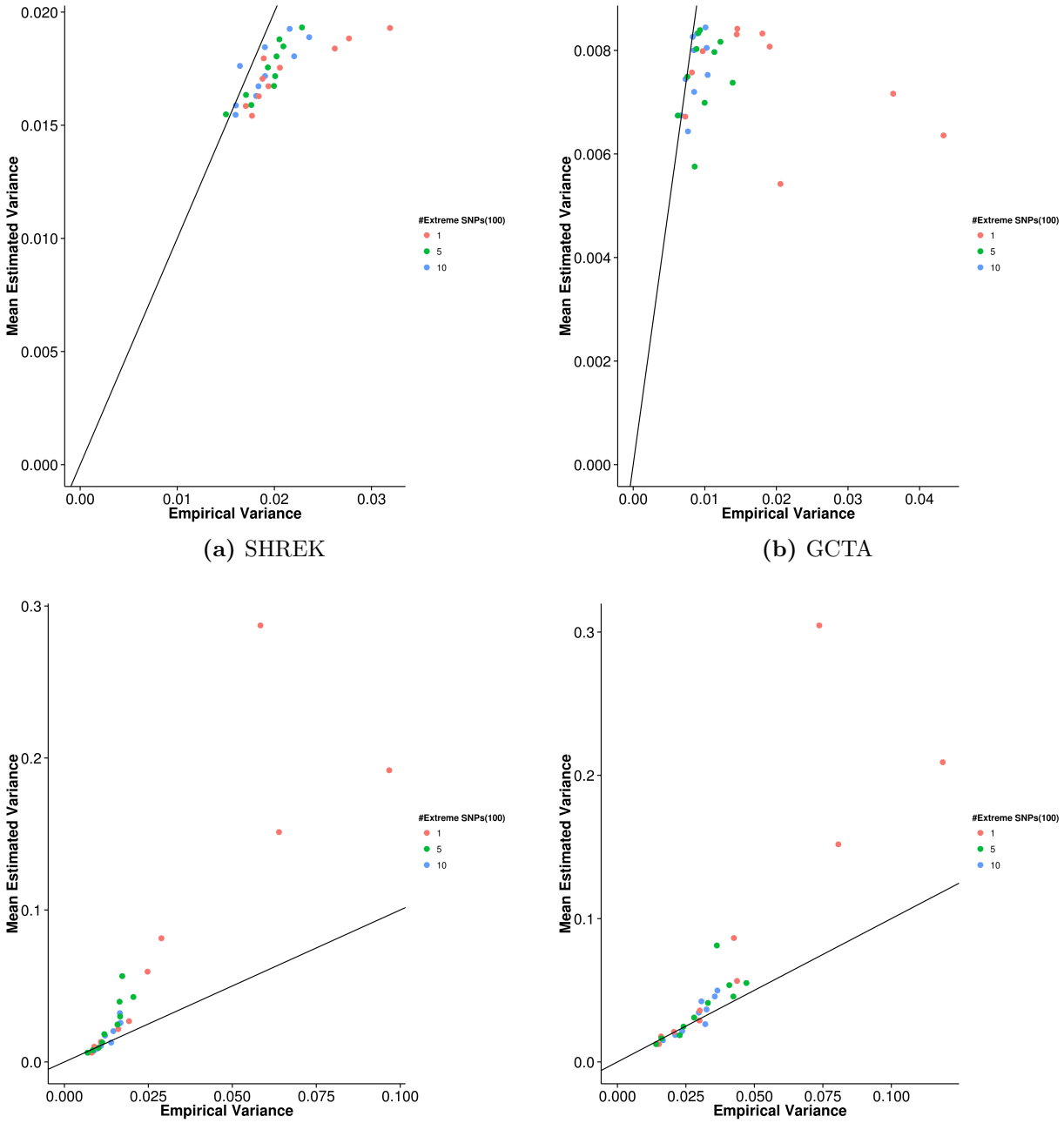
(a) SHREK

(b) GCTA

(c) LDSC with fix intercept

(d) LDSC with intercept estimation

Figure 2.7: Variance of results from quantitative trait simulation with extreme effect size simulation. 100 causal SNPs were simulated. When only 1 SNP with extreme effect was simulated, the empirical variance of GCTA and LDSC increases and a large fluctuation was observed. Whereas the empirical variance of SHREK only increases slightly when the simulated heritability is large and with only 1 SNP with extreme effect. This suggests that SHREK is more robust to the change in number of extreme SNP(s).



(a) SHREK

(b) GCTA

(c) LDSC with fix intercept

(d) LDSC with intercept estimation

Figure 2.8: 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 and GCTA generally under-estimate the variance with the magnitude of bias being the highest when there is only 1 SNP with extreme effect. On the other hand, LDSC tends to over-estimate the variance and it can overestimate the variance by more than 3 folds when there is only 1 SNP with extreme effect.

Number of Extreme SNPs	SHREK	LDSC	LDSC-In	GCTA
1	0.0227	0.0393	0.0508	0.0206
5	0.0203	0.0145	0.0316	0.00985
10	0.0205	0.0129	0.0329	0.00939

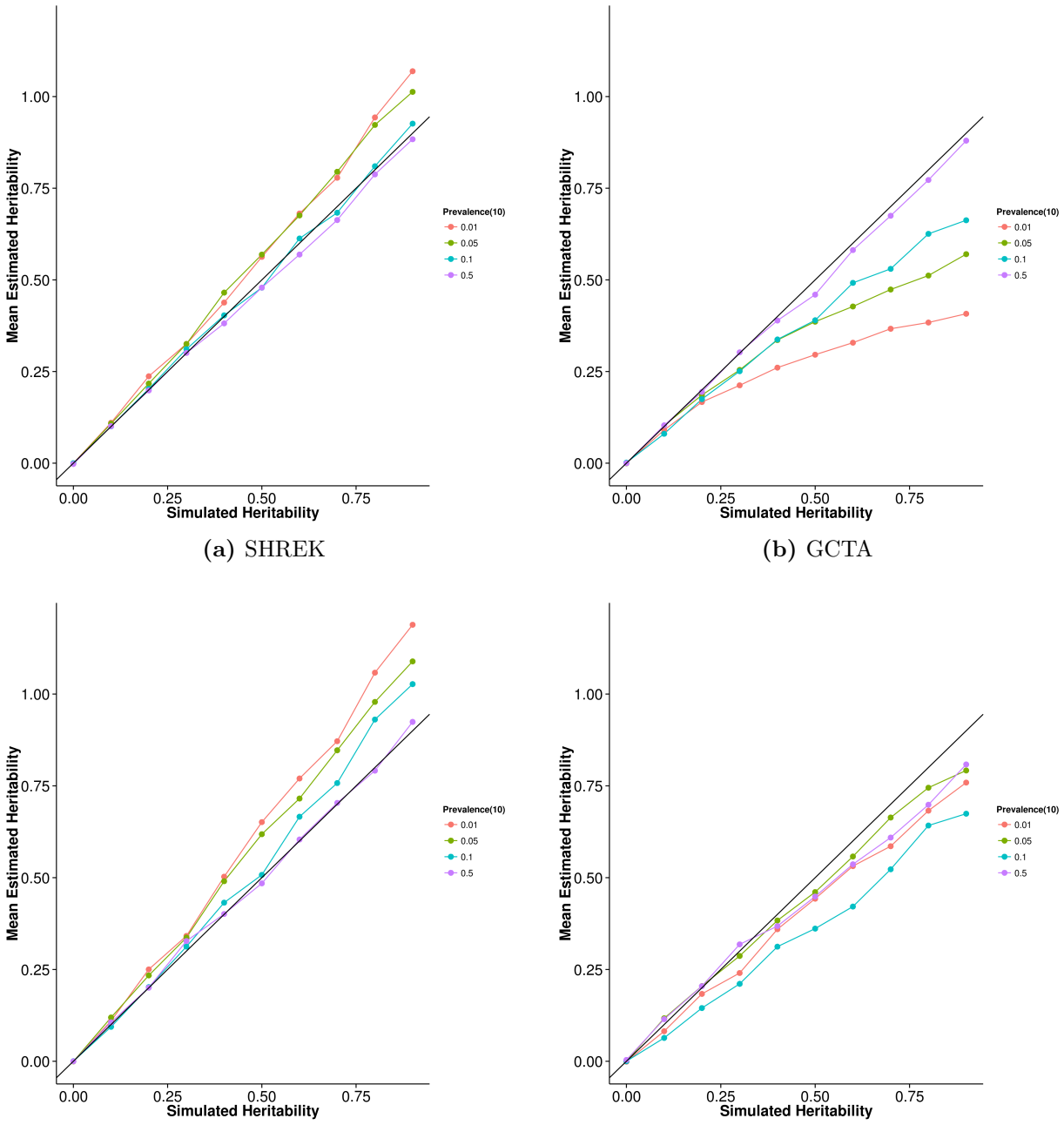
Table 2.2: MSE of quantitative trait simulation with extreme effect size. Of all the algorithms, GCTA has the lowest MSE in all situations. When comparing the performance of SHREK and LDSC, SHREK only has a better performance when there is one SNP with large effect. For other scenarios, LDSC with fixed intercept has better performance. However, we can observe that the performance of SHREK is very consistent and robust to the change in number of SNPs with extreme effect size.

effect where the larges discrepancy between the estimated and empirical variance is observed when only 1 SNP with large effect was simulated. It is observed that both SHREK and GCTA tends to underestimates their empirical variance whereas LDSC tends to overestimates the empirical variance. The difference between the estimated and empirical variance for LDSC with fixed effect can be as much as 3 fold.

To conclude, the performance of GCTA is superior to other algorithm(table 2.2). However, if we only consider the algorithms using summary statistic for heritability estimation, the performance of LDSC is better than SHREK when there are more than 1 SNP with large effect. Again, as no confounding factors were simulated, LDSC with fixed intercept outperforms LDSC with intercept estimation. It is interesting to note that the MSE of SHREK was least affected by the number of SNP(s) with large effect.

Case Control Simulation

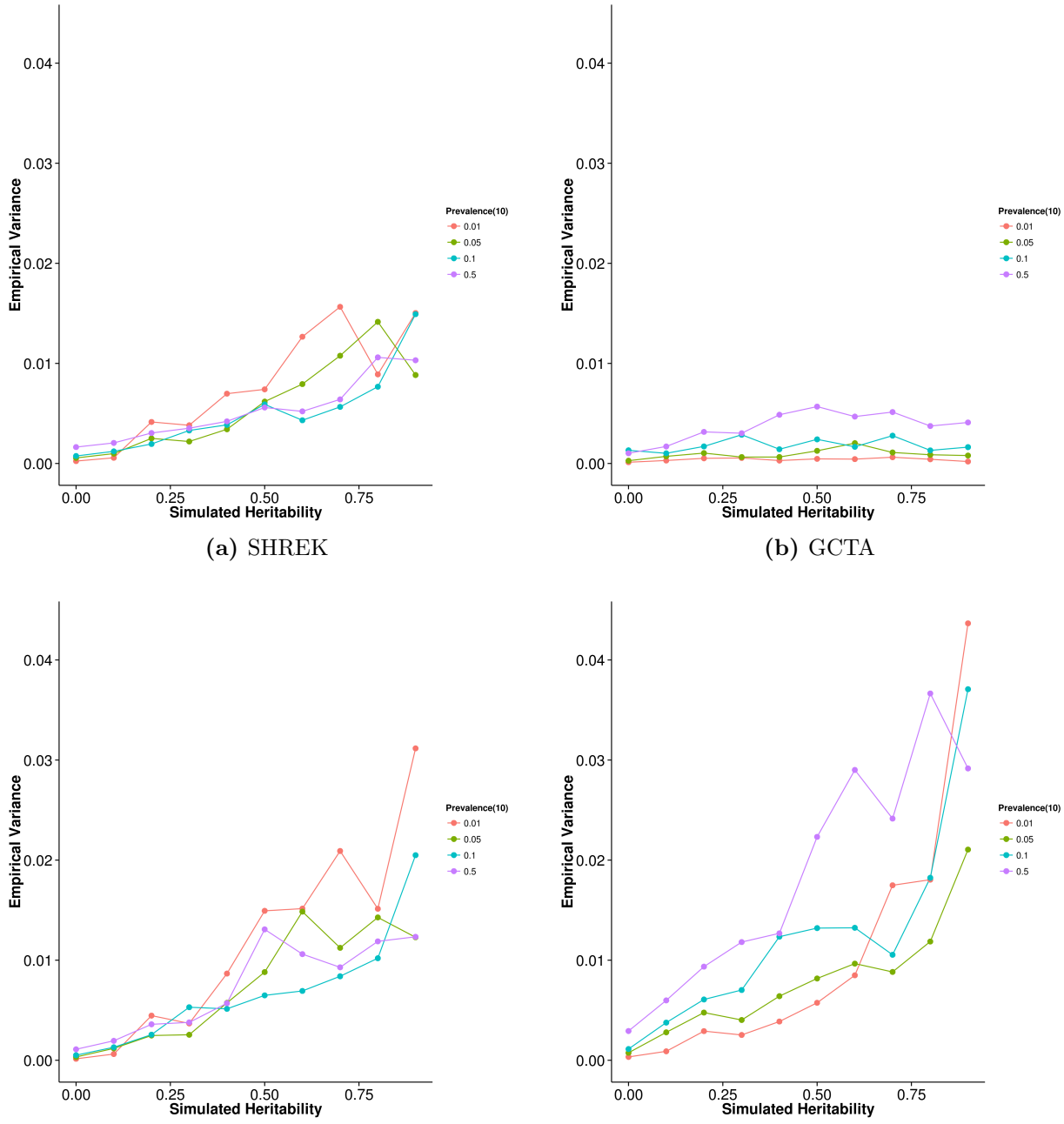
Nowadays, most of the GWAS are Case Control studies, thus it is important to test the performance of the algorithms when dealing with case control samples. In the case control simulation, we varied the population prevalence and the trait heritability. We also varied the number of causal SNPs to assess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



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

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

Figure 2.9: Mean of results from case control simulation with random effect size simulation with 10 causal SNPs. The performance of GCTA was as suggested by Golan, Eric S Lander, and Rosset (2014) where there was an underestimation as prevalence decreases. On the other hand, the upward bias of both LDSC with fixed intercept and SHREK increases as the prevalence decreases whereas LDSC with intercept estimation seems relatively robust to the change in prevalence.



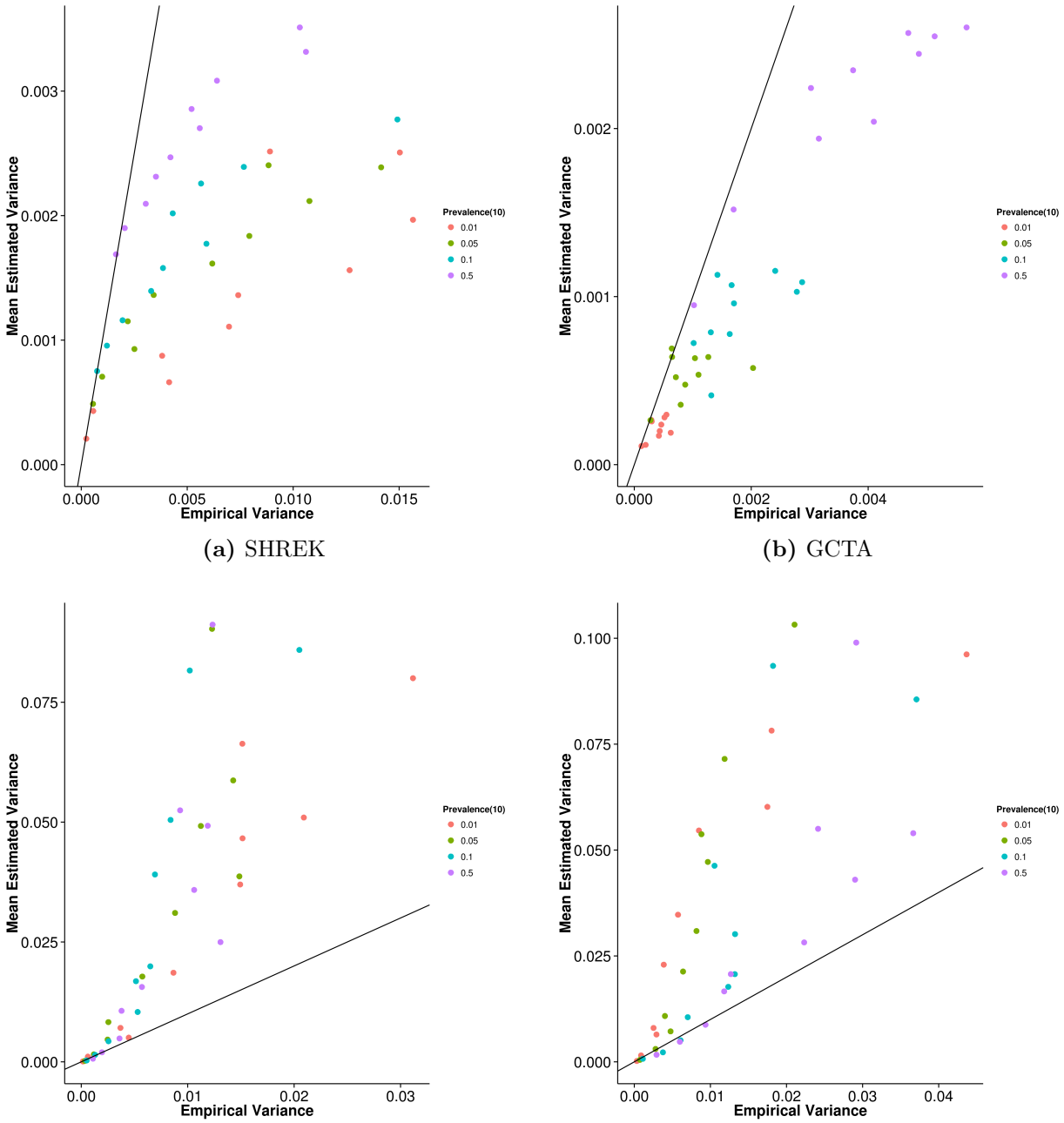
(a) SHREK

(b) GCTA

(c) LDSC with fix intercept

(d) LDSC with intercept estimation

Figure 2.10: Variance of results from case control simulation with random effect size simulation with 10 causal SNPs. There were no clear pattern as to how the prevalence affect the empirical variance of estimates from SHREK and LDSC. For GCTA, it seems like a larger prevalence tends to result in a larger empirical variance. Again, GCTA has the lowest variance, follow by SHREK and LDSC with fixed intercept. Nonetheless, it was important to remember that in case control simulation, a much smaller amount of SNPs was used, thus the results was not directly comparable to results from the quantitative simulation.



(a) SHREK

(b) GCTA

(c) LDSC with fix intercept

(d) LDSC with intercept estimation

Figure 2.11: Estimated variance of results from case control simulation with random effect size simulation when compared to empirical variance when 10 causal SNPs was simulated. A general underestimation was observed for SHREK and GCTA whereas a larger upward bias was observed for LDSC.

of these parameters to the performance of the algorithms.

First, we simulated traits with 10 causal SNPs. From the graph, it is clear that the population prevalence has a significant impact to the performance of the algorithms (fig. 2.9). The performance of GCTA is as suggested by Golan, Eric S Lander, and Rosset (2014) where the degree of underestimation increases as the prevalence decreases. On the other hand, the opposite effect is observed for SHREK and LDSC with fixed intercept. Interestingly, when allow the estimate the intercept, the heritability estimated from LDSC becomes underestimated. The magnitude of the bias also decreases, suggesting that the intercept estimation might have corrected for part of the bias of LDSC. The same pattern are also observed when the number of causal SNPs increases (figs. 2.17, 2.20 and 2.23), suggesting that the effect of number of causal SNPs are not the main contributor to the difference in bias.

As one inspect the empirical variance of the algorithms, GCTA clearly has the smallest average empirical variance among the algorithms (fig. 2.10b) where LDSC with intercept estimation has the largest empirical variance (fig. 2.10d). Unlike the quantitative trait simulation, the empirical variance of the estimates from SHREK (fig. 2.10a) are very close to that of LDSC with fixed intercept (fig. 2.10c). When the heritability of the trait is high, the empirical variance of SHREK is even lower than that of LDSC with fixed intercept. As one increases the number of causal SNPs, the empirical variance of all algorithms decreases (figs. 2.18, 2.21 and 2.24) agreeing with the results from the quantitative trait simulation.

On the other hand, both SHREK (fig. 2.11a) and GCTA (fig. 2.11b) underestimates their empirical variance whereas LDSC overestimates its empirical variance no matter if the intercept estimation was performed (fig. 2.11). As the number of causal SNPs increases (figs. 2.19, 2.22 and 2.25), the bias of variance estimation remain unchanged for SHREK. However, for LDSC, the magnitude of bias of vari-

ance estimation reduces as the number of causal SNPs increases and are able to provide a relatively accurate estimation of its empirical variance when 500 causal SNPs were simulated (fig. 2.25c).

Taking into account of the bias and variance of the estimations (table 2.3), SHREK has the best average performance of all the algorithm tested. Interestingly, the performance of LDSC with intercept estimation were better than LDSC with fixed intercept when the prevalence is small even-though we did not simulate any confounding factors. In such scenario, one would expect the intercept estimation to be unnecessary and will only increase the SE of the heritability estimation without improving the estimates yet from the simulation results, it was suggested that the intercept estimation might helps correct for some of the bias in the estimates when the prevalence is small.

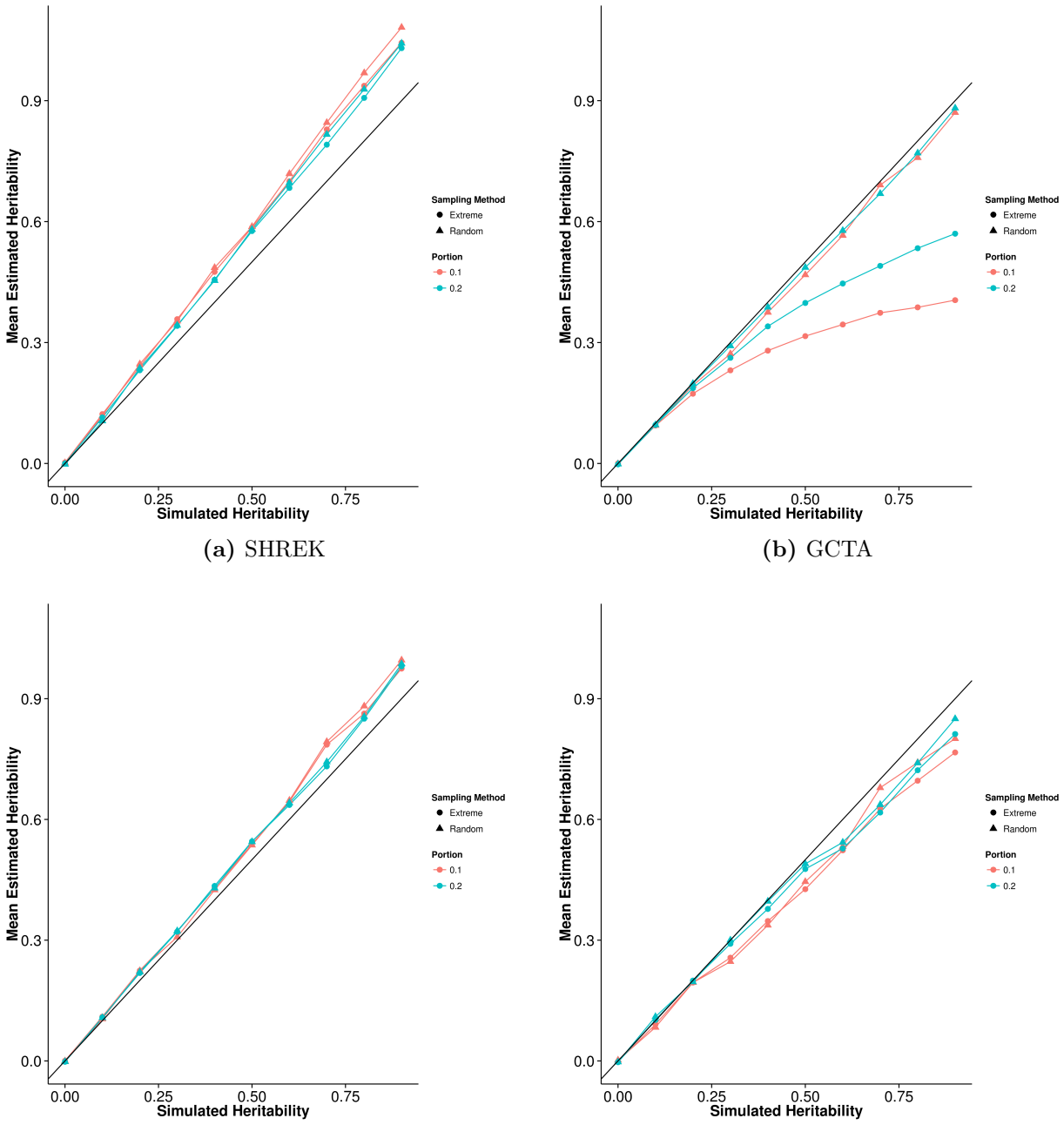
In general, the effects of the number of causal SNPs in the case control simulation agrees with what was observed in the quantitative trait simulations. As the number of causal SNPs increases the MSE tends to decrease for all algorithms, with SHREK least sensitive. Finally, it is important to note that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

Extreme Phenotype Simulation

Sometimes, when budget is limited, it is not possible to include all samples in the final GWAS. By using appropriate sampling strategy, such as that of extreme phenotype sampling (Peloso et al., 2015), one can increase the power of the association study. Here we perform simulations using extreme phenotype sampling and study

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

Table 2.3: MSE of Case Control simulation. Algorithm with the best performance under each condition were bold-ed. When the population prevalence is 0.5, GCTA has the best performance, followed by SHREK. For most other conditions, SHREK has the best performance. Of all the algorithms, SHREK has the lowest average MSE. Also, as the number of causal SNPs increases, the MSE tends to decrease for all algorithms, similar to what was observed in the quantitative simulation.



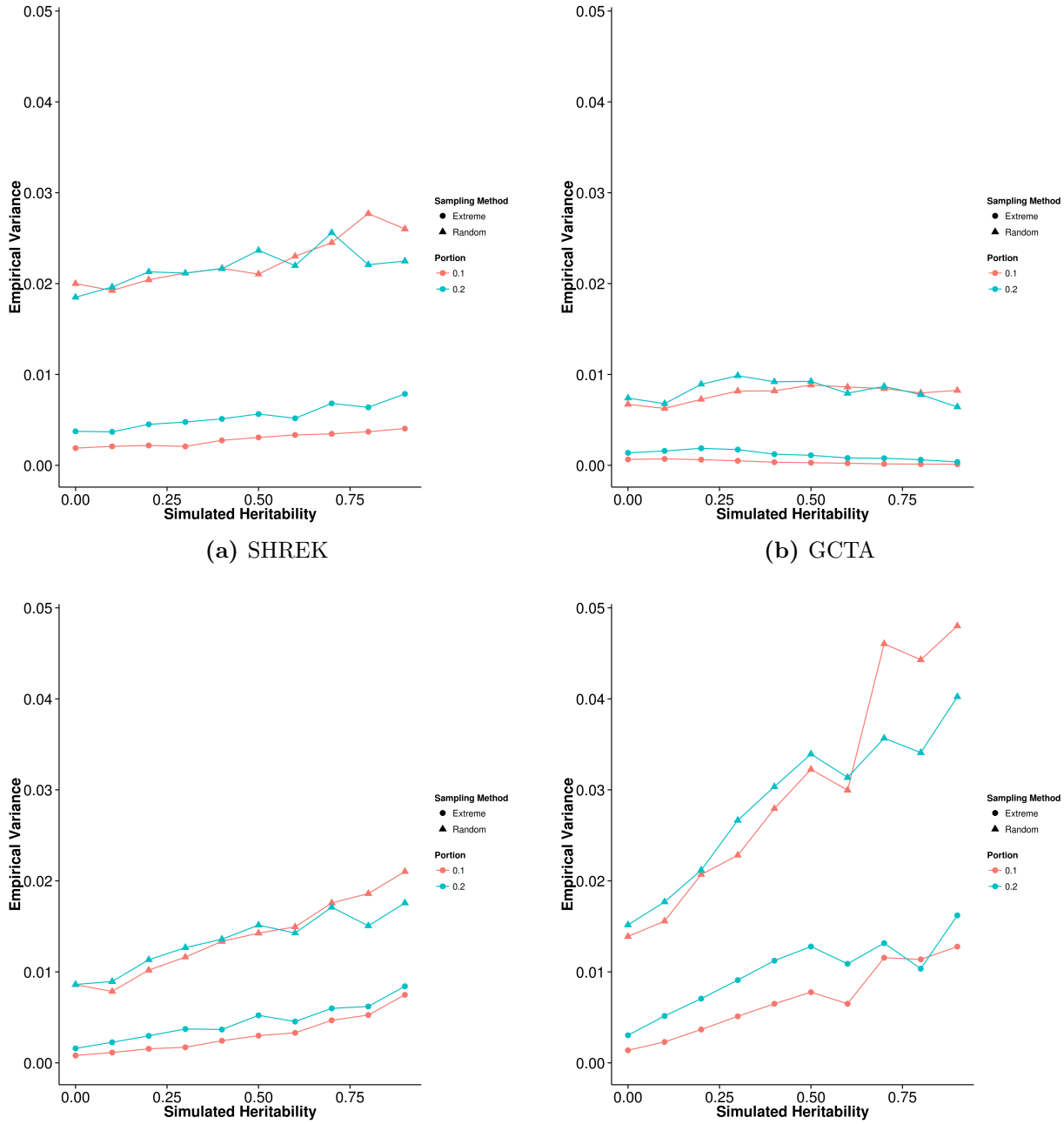
(a) SHREK

(b) GCTA

(c) LDSC with fix intercept

(d) LDSC with intercept estimation

Figure 2.12: Mean of results from extreme phenotype simulation. The performance of the algorithms when random sampling was performed were similar to what was observed in the quantitative trait simulation. However, when extreme phenotype was performed, a larger under estimation was observed for GCTA and it gets worst when the portion of sample selected decreases. On the other hand, the performance of SHREK and LDSC under the extreme phenotype selection was similar to that from the random samplings.



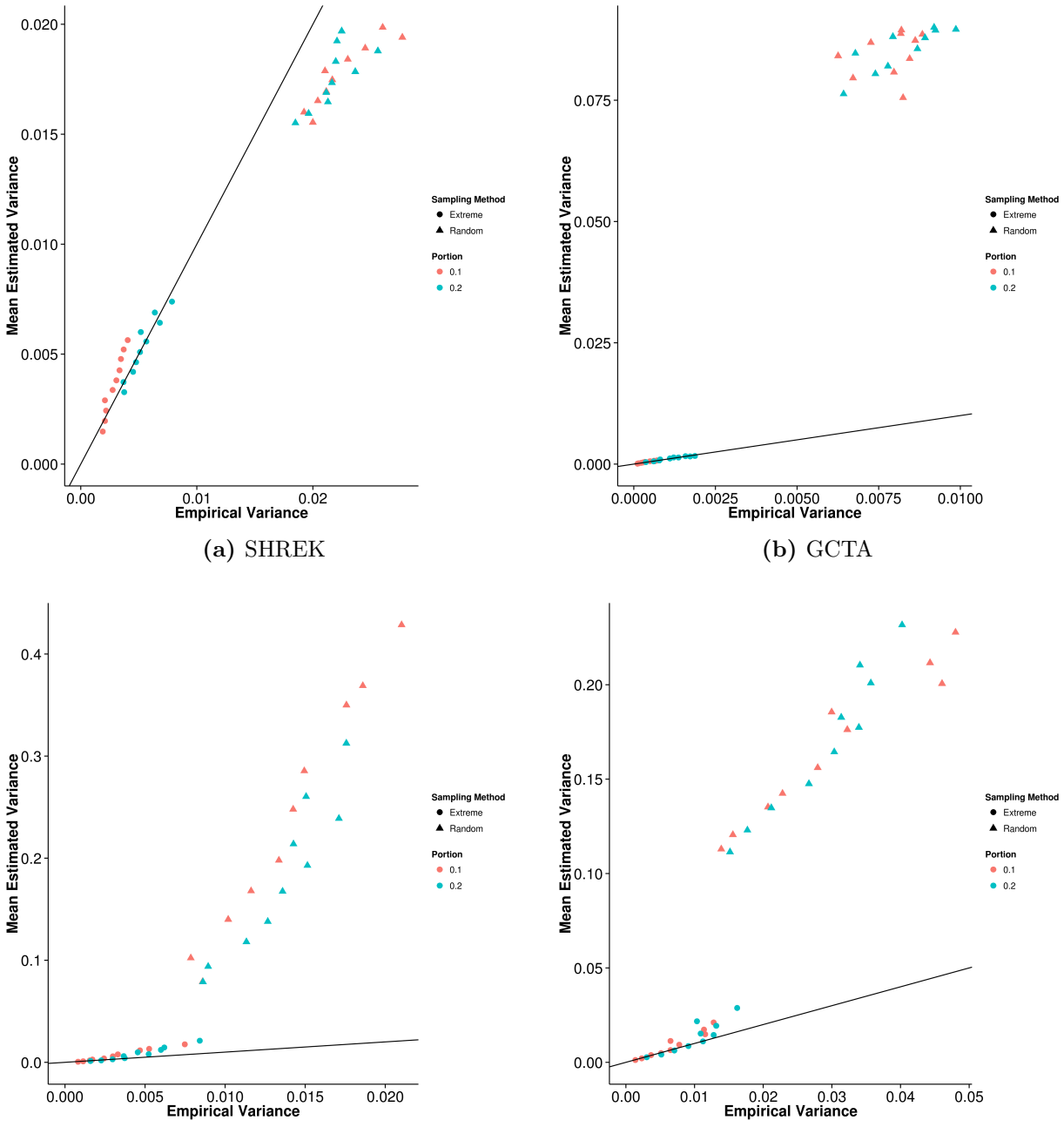
(a) SHREK

(b) GCTA

(c) LDSC with fix intercept

(d) LDSC with intercept estimation

Figure 2.13: Variance of results from extreme phenotype simulation. It is obvious that when the extreme phenotype selection was performed, the empirical variance of all the algorithm decreases and is much smaller than the empirical variance of the estimation when random sampling was performed. We also compared the empirical variance of random sampling with those from quantitative trait simulation with 100 causal SNPs and they are highly similar.



(a) SHREK

(b) GCTA

(c) LDSC with fix intercept

(d) LDSC with intercept estimation

Figure 2.14: Estimated variance of results from extreme phenotype selection when compared to empirical variance. Surprisingly, except for SHREK, the estimated variance from LDSC and GCTA under the random sampling condition was much higher than the empirical variance. It is much different from the estimated variance from the quantitative trait simulation and further investigations are required to understand this discrepancy.

the effect of this selection on the performance of heritability estimations. The random sampling procedure were also performed in our simulations such that a clear comparison can be made between the power of extreme phenotype sampling and the traditional random sampling.

From the graph (fig. 2.12), it is observed that performance of SHREK and LDSC are similar to what was observed in the quantitative trait simulation. Moreover, when the random sampling strategy were used, a larger upward bias are usually observed. Interestingly, GCTA performs poorly when extreme phenotype sampling was performed. As the portion of sample sampled decreases, the bias of the estimates from GCTA increases (fig. 2.12b).

When comparing the empirical variance, the random sampling strategy consistently results in larger variance when compared to extreme phenotype sampling strategy (table 2.4). The MSE from extreme phenotype sampling can be as much as 4 fold smaller for SHREK and LDSC when compared to random sampling.

Peculiarly, although the empirical variance under the random sampling strategy is the same as what was observed in the quantitative trait simulation, there is a large discrepancy in the estimated variance where a tenfold overestimation is observed for LDSC and GCTA (fig. 2.14). More surprisingly, SHREK is unaffected. We are uncertain of the origin of such problem and further investigations are required.

2.3.3 Application to Real Data

We applied our method and LDSC to the PGC schizophrenia (SCZ), major depression disorder and bipolar data sets. To adjust for the confounding factors, intercept estimation were performed for LDSC.

It is estimated that the heritability for major depression disorder is around

Portion	Shrek		LDSC		LDSC-In		GCTA	
	Extreme	Rand	Extreme	Rand	Extreme	Rand	Extreme	Rand
0.1	0.0113	0.0341	0.00537	0.0167	0.0119	0.0329	0.0644	0.00849
0.2	0.0109	0.0290	0.00599	0.0152	0.0126	0.0299	0.0274	0.00852

Table 2.4: Here, we compared the MSE of random sampling (Rand) against the MSE of Extreme phenotype sampling (Extreme). With the exception of GCTA, the extreme phenotype selection generally produce a smaller MSE when compared to random sampling. However, for GCTA, because of the large bias introduced by extreme phenotype sampling (fig. 2.12b), the MSE is much higher when extreme phenotype sampling was performed.

	Major Depression Disorder	Bipolar	Schizophrenia
SHREK	0.256 (0.0273)	0.312 (0.0168)	0.174 (0.00453)
LDSC	0.161 (0.0317)	0.185 (0.0211)	0.133 (0.0071)

Table 2.5: Heritability estimated for PGC data sets. The heritability estimation form SHREK tends to be higher than that from LDSC. One major difference between LDSC and SHREK is that LDSC can remove confounding factors such as population stratifications from their estimation using the intercept estimation function. If there is any confounding factors, they can possibly inflate the estimates from SHREK

0.256 by SHREK and 0.161 by LDSC whereas the heritability of bipolar is estimated to be around 0.312 by SHREK and 0.185 by LDSC (table 2.5). As for schizophrenia, the heritability is estimated to be around 0.133 by LDSC and 0.174 by SHREK. The estimated intercept from LDSC for bipolar and major depression is 1.06 and 1.026 respectively suggesting there is little confounding factors. On the other hand, the estimated intercept is around 1.21 for schizophrenia, suggesting there might be small amount of confounding effect in the estimation. Indeed, in PGC schizophrenia study (Stephan Ripke, B. M. Neale, et al., 2014), a small amount of Asian samples were included. As SHREK does not adjust for the population stratification, caution must be paid when interpreting the results.

2.4 Discussion

In order to study complex disorders such as schizophrenia, large amount of samples are required and often it is not possible for one single group of researchers to collect sufficient samples. Therefore, collaboration and large scale consortium becomes vital and allow for sufficient sample size to be collected. However, due to privacy concerns, the raw genotypes of the participants were usually not shared amount groups or that the genotype is only provided through a tedious and lengthy application process (e.g. dbGaP). Thus these large scale studies relies on the meta analysis and only the summary statistics of the final analysis were provided to the public.

Traditional SNP heritability estimation algorithms for GWAS such as GCTA and Phenotype correlation - genotype correlation regression (PCGC) relies on the genetic relationship matrix which can only be calculated based on the genotypes of the subjects. Not until the development of LDSC and SHREK is there a way to estimates the SNP heritability without the raw genotypes. By being able to estimate the SNP heritability from only the summary statistic from a GWAS, one can now compare the difference between the heritability estimated from twin studies and the SNP heritability estimated from GWAS to estimate the relative contribution of SNPs to the disease variance without requiring the raw data. The relative contribution of SNPs will allow researchers to plan subsequent studies accordingly. For example, if the SNP heritability is much smaller than the heritability of the disease, alternative strategies like whole genome sequencing would be more efficient for identifying additional genes associated with the disease, compared to GWAS.

Despite the promise of LDSC and SHREK, their developments were far from completion. For example, a big issue observed in our simulation was the influence of the sampling bias of the LD which is one of the key element required for LDSC and SHREK.

2.4.1 LD Correction

It was known that the LD contains sampling bias and the sample R^2 is usually bigger than the true R^2. Therefore it is important for one to adjust for the sampling bias before applying them in the estimation of heritability.

When comparing impact of different bias correction algorithm on the performance of SHREK, it is observed that majority of the algorithms, except that of eq. (2.40), inflates the heritability estimated, suggesting that there is an overestimation, whereas when the sampling bias left uncorrected, the estimates are biased downward, as one would expect. The superior performance of eq. (2.40) lead us to use it as our default LD sampling bias correction algorithm.

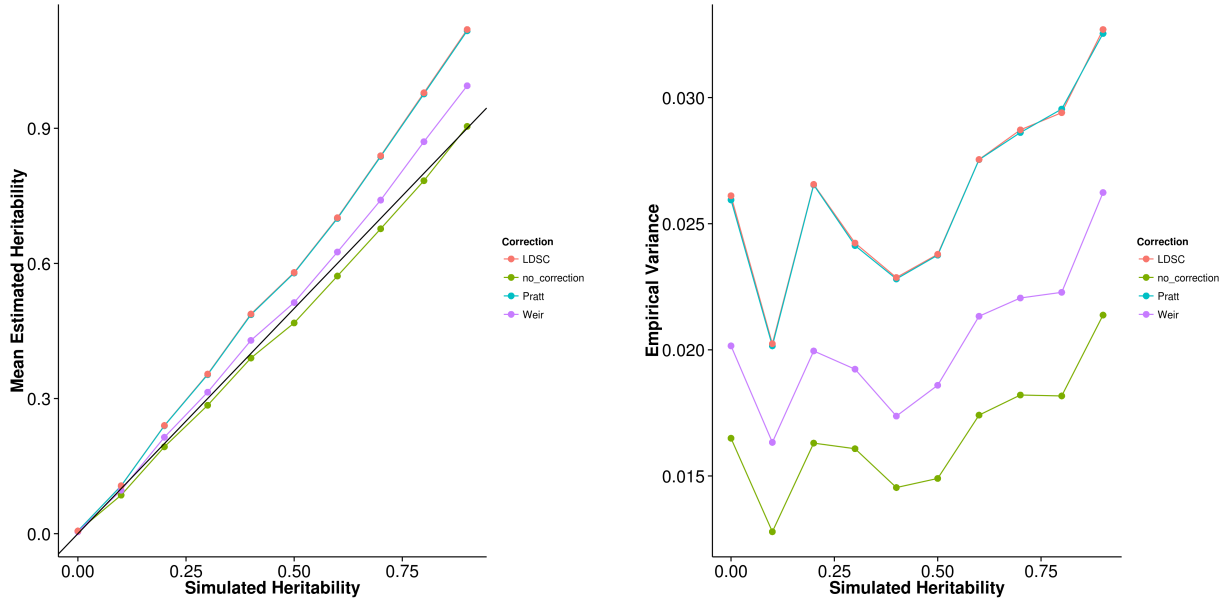
What was surprising is that in the quantitative trait simulation, an overestimation of heritability is observed despite using eq. (2.40) for LD correction. This overestimation is similar to what was observed in the previous LD correction simulations where 5,000 SNPs on chromosome 22 were simulated. It is possible that despite the superior performance of eq. (2.40), small imprecisions were introduced to the LD matrix during the bias correction. When the number of SNPs increases, these imprecisions cumulates, thus leads to bias in the final heritability estimates.

Intriguingly the same overestimation is not observed in LDSC. When inspecting the algorithm of LDSC, it is observed that LDSC also correct for the sampling bias in R^2 using:

$$\text{LDSC} : \tilde{R}^2 = \hat{R}^2 - \frac{1 - \hat{R}^2}{n - 2} \tag{2.44}$$

which was not tested in our previous LD correction simulation.

An interesting analysis will be to test the performance of the LD correction algorithm when the number of SNPs is higher (e.g 50,000 SNPs on chromosome 1) and whether if eq. (2.44) produce a better results. We therefore repeated the LD



(a) Mean Estimation **(b)** Empirical Variance

Figure 2.15: Effect of LD correction to Heritability Estimation when 50,000 SNPs were simulated. As an overestimation is observed in the quantitative trait simulation, we performed a short simulation to assess the impact of LD correction to the heritability of SHREK when the there is a larger number of SNPs. From the graph, it is observed that all LD correction algorithms inflate the heritability estimation when large number of SNPs were simulated. In fact, the bias is the smallest when no LD correction was performed.

correction simulation by increasing the number of simulated SNPs to 50,000 on chromosome 1. To reduce the run time of the simulation, we only compared the performance of SHREK when eq. (2.44), eq. (2.40) and eq. (2.38) were used for the LD correction.

From the results (fig. 2.15), it is clear that all LD correction algorithms inflates the heritability estimation from SHREK in oppose to the underestimation observed when no LD correction was performed. The underestimation is as expected because the positive sampling bias in R^2 will lead to an "over correction" of the collinearity, thus results in lower estimates. As mentioned, it is possible that eq. (2.40) does introduce small imprecision (e.g. overcorrection) to the LD matrix which accumulates as the number of SNPs increases, leading to overestimation

of the heritability. Our simulation results do support that as one of the possible explanation. What is interesting though is that the MSE is the lowest when no LD correction was performed, suggesting that when the number of SNPs increases, these LD correction algorithms actually has a negative impact to the performance of SHREK.

It is noted that most LD correction algorithm assumes the correlation was calculated on normally distributed data. However, genomic data follows a binomial distribution, which might violates the assumption, leading to a biased correction. This will be an important area for further research. Without a good bias correction algorithm, the estimates from SHREK will most likely be biased downward, especially when the reference panel is small. Meanwhile, we allow users the freedom to disable the LD correction in SHREK.

Another important observation is the overestimation observed when we use the LD correction algorithm from LDSC on SHREK. Using the same algorithm, the estimates from SHREK are biased upward whereas the same bias was not observed in LDSC. This observation suggests that SHREK might be more sensitive to the errors in the LD matrix when compared to LDSC. Indeed, SHREK requires the inverse of the LD matrix and considering the large condition number of the LD matrix, any errors can be multiplied during the inversion. On the other hand, LDSC does not compute the inverse of LD matrix. Instead, they only require the *sum* of R^2 for the regression model. By avoiding the inverse of the matrix, the algorithm will then be less sensitive to the imprecision in the LD, thus result in a better estimates. However, it will still be interesting to see whether if the application of a better LD correction algorithm can help to improve the estimates from LDSC.

2.4.2 Simulation Results

To understand how the performance of the heritability estimation algorithm was influenced by different genetic architectures, we performed a series of simulations.

Quantitative Trait Simulation

In the quantitative trait simulation, it is clear that for most situation, GCTA has the best performance. By using the genetic relationship matrix, the estimation from GCTA are more accurate when compared to LDSC and SHREK. However, when the sample genotypes are unavailable, it is not possible to calculate the genetic relationship matrix required by GCTA. Thus one can only rely on LDSC and SHREK.

When the trait is polygenic, it is observed that the estimates of LDSC with fixed intercept are more accurate than the estimates from SHREK. However, under the oligogenic condition (e.g with only 5 or 10 causal SNPs), the variance of LDSC increases, thus increasing the MSE. On the other hand, the estimates of SHREK are relatively insensitive to the number of causal SNPs. As a result of that, under the oligogenic condition, SHREK has a better performance when compared to LDSC.

An important factor to remember is that in our simulation, we did not simulate any confounding factors, therefore the intercept estimation in LDSC was expected to only increase the variance without any gain in estimation power. The results from the simulation agrees with the hypothesis and demonstrated that the intercept estimation does increase the variance of the estimates, leading to a higher MSE.

It will be interesting to assess the performance of these algorithms when there is confounding effects such that one can test the importance of the intercept estimation function in the correction of confounding effects. However, the simula-

tion of population and, especially cryptic relationship, is nontrivial. For example, although one can provide haplotype from different population to HAPGEN2, there is a lot of uncertainties in the simulation of the individual phenotypes: Should one standardize the genotype of the two population independently in the calculation of phenotype? Should the two population have the same causal SNPs? If not, should we limit the causal SNPs within the same biological pathway / function?

Moreover, heritability is dependent on the environment and genotype frequency. Theoretically, it is possible for different population to have a different heritability for a particular trait. The possible combinations and the complexity of the problem is beyond the scope of this thesis but we do acknowledge that it is an important subject and further research is required.

Overall, when compared to LDSC, the only advantage of SHREK is its relative robustness to change in genetic architecture of the trait. Under extreme scenarios such the oligogenic condition, or when there is one SNP with extreme effect size, the performance of SHREK remains relatively unaffected when compared to LDSC which usually result in a larger variance under the extremes. Whereas under polygenic condition LDSC outperforms SHREK. It is important to note that the bias of SHREK is mainly due to the LD correction algorithm, if LD correction was not performed, the MSE of the estimates form SHREK will be reduced (e.g. from 0.0217 to 0.0166 in the LD correction simulation), reducing the difference in performance between LDSC. Nonetheless, the sensitive to errors in the LD matrix remains to be one of the biggest weakness of SHREK.

Case Control Simulation

More often than not, researchers are interested in case control studies where "affected" and "normal" samples were compared. This is particular useful for the studies of disease traits such as schizophrenia. However, the heritability estima-

tion is not as straight forward and requires the adaptation of the liability threshold model. It was known that GCTA, the most widely adopted algorithm for heritability estimation in GWAS is unable to provide accurate estimates in case control scenarios and its estimates are affected by the population prevalence and sample size of the studies (Golan, Eric S Lander, and Rosset, 2014). Our simulation results agree with the observation of Golan, Eric S Lander, and Rosset (2014), suggesting that as the population prevalence decreases, the magnitude of bias in the estimates of GCTA increases.

According to Golan, Eric S Lander, and Rosset (2014), in case control studies there is an oversampling of the cases relative to their prevalence in the population. The case control sampling induced a positive correlation between the genetic and environmental effects for the samples in the study even when there is no true genetic and environmental interaction in the population (Golan, Eric S Lander, and Rosset, 2014). This leads to heritability estimates from GCTA to be strongly downward biased where the magnitude of bias increases as the population prevalence decreases, heritability increases and when the proportion of cases is closer to half.

The question then is whether if this artificial correlation will affect the performance of SHREK and LDSC. First, it is observed that as the population prevalence decreases, the magnitude of bias for both LDSC with fixed intercept and SHREK increases suggesting that the population prevalence and the sampling bias might indeed be influential to the estimates of LDSC and GCTA. However, the direction of bias is opposed to what was observed in GCTA where a smaller population prevalence leads to a larger *overestimation* in the heritability. Considering that for SHREK, we adjusted the estimates by multiplying eq. (2.23) to the estimates, an overestimation might suggest that we have an under correction of the bias. Of course the bias introduced by the LD correction is another factor to be considered, but considering that only 5,000 SNPs were simulated, the bias introduced by LD

sampling bias should be relatively small as suggested by our LD correction simulation. To understand the effect of LD correction in case control scenario, we will need to increase the number of SNPs simulated yet that is only possible when additional computation resources are made available.

What is most surprising in the case control simulation is the performance of LDSC with intercept estimation. As we did not simulate any confounding factors, we expect the performance of LDSC with intercept estimation would be worst compared to LDSC with fixed intercept because of the unnecessary additional degree of freedom in the estimation. However, it is observed that unlike SHREK and LDSC with fixed intercept, the bias of LDSC with intercept estimation is robust to the change in population prevalence (figs. 2.9, 2.17, 2.20 and 2.23), thus when the population prevalence is small, the bias of LDSC with intercept estimation is relatively smaller when compared to LDSC with fixed intercept.

Taking into consideration of the empirical variance and the bias of the estimates, SHREK has better average performance when compared to LDSC. It is important to remember that the case control simulation is not comparable to the results from the quantitative trait simulation, not only because the addition of the liability model, but also that in the case control simulation, we only simulated 5,000 SNPs on chromosome 22. Based on the LD correction simulation, it is observed that bias from the LD correction algorithms is smaller when less SNPs were simulated. On top of that, the total amount of samples included in the simulation was doubled that from the quantitative trait simulation, with 1,000 cases and 1,000 controls whereas we only simulated 1,000 total samples in the quantitative trait scenario. Nonetheless, our case control simulation does highlights the effect of population prevalence on the performance of the heritability estimation algorithms. It will be an important topic to develop better algorithm for adjusting the attenuation bias introduced by case control sampling when the population prevalence of the disease

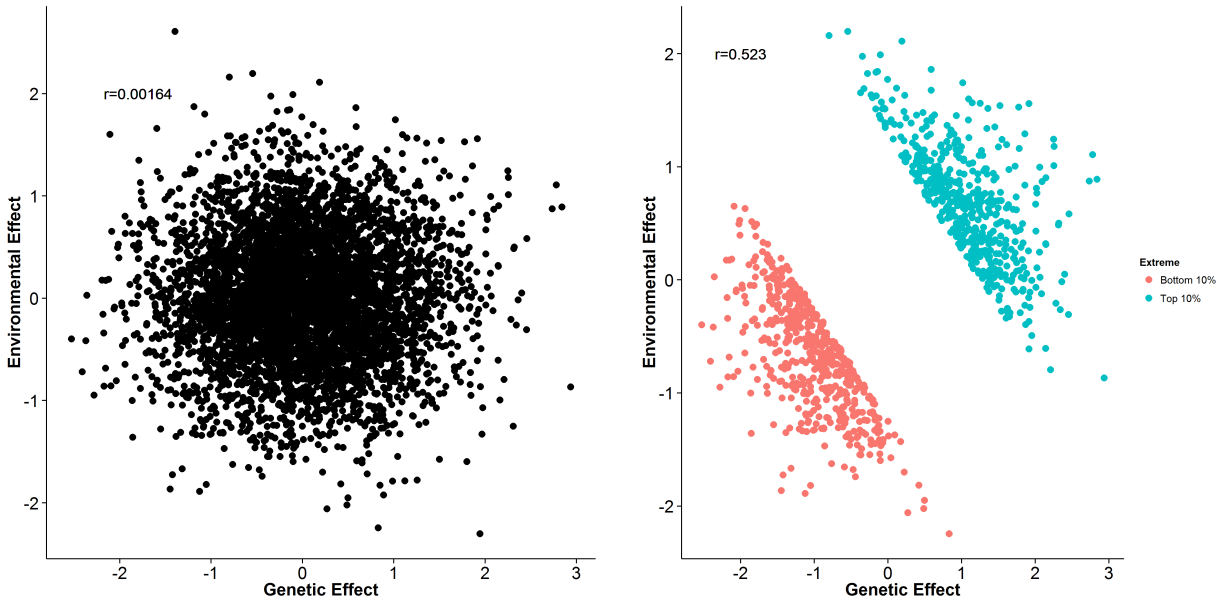
is small.

Finally, it is noted that in order to provide an accurate estimation of the heritability, one needs to know the population prevalence of the disease beforehand. Without the information of the population prevalence, it will be difficult for one to estimates the heritability from GWAS with case control design. Therefore once should always be cautious with the heritability estimations from a case control designs when the population prevalence is unknown.

Extreme Phenotype Sampling

Other than the case control study design, extreme phenotype sampling is another common experimental design for it can helps to increase the power of an association studies given the same amount samples. Compared with the same number of randomly selected individuals, the extreme selection design can increase the power by a factor of $\frac{V'}{V}$ where V' is variance of the trait of the selected sample and V is the trait variance of the general population. So for example, if one only include the samples from the top 5% and bottom 5% of the phenotype distribution, one can achieve the same power as a study with random sampling design that has 4 times the sample size (Pak C Sham and Shaun M Purcell, 2014).

Herein, we simulated the situation where an extreme selection design was performed to assess the performance of the heritability estimation algorithms. We were also interested in comparing the performance between extreme phenotype sampling and random sampling strategy. First, it is observed that when extreme phenotype sampling was performed, the estimates from GCTA are biased downward. This observation is similar to what was observed in the case control simulation. It is noted that although we were simulating independent environmental and genetic effects, the extreme phenotype sampling strategy does introduced an artificial correlation between the two effects, similar to what was observed in case control



(a) No Extreme sampling **(b)** Extreme Sampling

Figure 2.16: Effect of extreme sampling design. Here we simulated the genetic and environmental effect independently. When no extreme sampling was performed, there is no correlation between the environmental effect and genetic effect as expected. However, when extreme sampling was performed, an artificial correlation is observed. This might be the main reason why the estimates from GCTA are downward biased.

scenario (fig. 2.16). This might therefore affect the performance of GCTA where as the portion of sample selected decreases, the magnitude of bias increases, similar to the change of population prevalence in case control studies.

On the other hand, an upward bias is observed in the estimates from SHREK and LDSC. Although the same bias can be observed in the random sampling scenario, the bias is slightly high when a smaller portion of samples were selected. This level of bias concurs with the biased observed when a trait has a smaller population prevalence suggesting that the sampling method might introduce bias in the SNP heritability estimate. Studies are therefore required to identify a better algorithm for the correction of the attenuation bias. Overall, the performance of SHREK and LDSC are more than 3 fold better when extreme selection was performed, suggesting that the extreme selection does help to improve the power in estimation even though the same amount of samples were used.

However, although the empirical variance observed in the random sampling for all the algorithm are the same as what was observed in the quantitative trait simulation with 100 causal SNPs, the estimated variance for GCTA and LDSC are much worst. A larger upward bias is observed in the estimates from LDSC with fixed intercept, suggesting there might be some difference between the simulation of random sampling and the simulation of quantitative trait, even though most of the parameter for simulation are the same. The only difference in the two simulation was the standardization of genotype when calculating the phenotype. For the quantitative trait simulation, the genotype was standardized based on the genotype of 1,000 samples of which all were included in the analysis. However, in the simulation of the random sampling design, 5,000 samples were used to standardize the genotype, of which only 1,000 out of 5,000 were included in the final analysis. It is uncertain how this affects the performance of the algorithm and further analysis might be required.

Nonetheless, in this simulation, we first simulated the individuals and their phenotype *then* we perform the sampling. The only difference between the two sets of data is the sampling performed. Thus it is safe to conclude that the extreme phenotype sample does provide more power than the random sampling in heritability estimation.

Finally, we only tested the performance of the algorithms when the trait is polygenic (e.g. 100 causal SNPs). Further simulation should be performed to test the effect of extreme phenotype selection on traits with different genetic architecture.

2.4.3 Application to Real Data

Our main question of interest is to understand what is the true contribution of common genetic variants, such as SNP, to the variance of schizophrenia. Although B. Bulik-Sullivan (2015) estimated that the SNP heritability of schizophrenia is around 0.555, it is still interesting to see if the same results can be calculated when different method was used. In order to make sure our analysis is correct and that the concordance between estimates from different tools were not merely by chance, we also estimated the heritability for bipolar disorder and major depression disorder as a reference point.

What is most surprisingly os that the LDSC estimated heritability is much smaller than the estimates from the supplementary materials of B. K. Bulik-Sullivan et al. (2015) (e.g. for schizophrenia, 0.555 compared to 0.133). From B. K. Bulik-Sullivan et al. (2015), the formula of LDSC is

$$\mathrm{E}[\chi^2|l_j] = Nl_j\frac{h^2}{M} + Na + 1 \tag{2.45}$$

where l_j is the LD score of variant j, N is the sample size, a is the contribution of confounding biases, h^2 is the heritability and M is the number of SNPs. When

	Major Depression Disorder	Bipolar	Schizophrenia
SHREK	0.256 (0.0273)	0.312 (0.0168)	0.174 (0.00453)
LDSC	0.235 (0.0241)	0.267 (0.0147)	0.197 (0.0058)

Table 2.6: Heritability estimated for PGC data sets without Intercept Estimation. Indeed, when the intercept estimation was not performed, the estimates from LDSC was very close to that of SHREK.

contact the author about the discrepancy of the estimation between our run of LDSC and the estimates shown in the supplementary table, B. Bulik-Sullivan (2015) replied that the estimated from the supplementary table define M as the total number of SNPs in the reference panel used to estimated LD score whereas the current version of LDSC defines M as the number of SNPs with maf$>$ 5% in the reference panel used to estimate LD score which they deem more appropriate based on new data they observed after their original paper was published. Based on the caption of their supplementary, they stated that "... if the average rare SNP explains less phenotypic variance than the average common SNP, then a smaller value of M would be more appropriate, and the estimates in the supplementary table will be biased upwards." (B. K. Bulik-Sullivan et al., 2015). This explain the smaller estimates from our run.

Another interesting observation from the estimates in real data is that SHREK consistently return a higher estimates when compared to LDSC. Considering the fact that SHREK cannot account for confounding effects such as cryptic relationship and population stratification, it is likely that the estimates are inflated by these confounding factors. A straight forward test is to perform LDSC without the intercept estimation and compare the estimates with that from SHREK such that it is clear whether if the difference of the estimates was due to the ability of estimating the intercept by LDSC. Indeed, when the intercept estimation was not performed, the estimates from the two algorithms converges (table 2.6). Therefore, it is likely that the difference in table 2.5 is a direct result of the estimation of the intercept.

However, it is very important for one to remember that it is difficult to tell which estimates is the "correct" estimate. For example, in the case control simulation, it was observed that SHREK and LDSC with fixed intercept will *overestimates* the heritability when the prevalence is less than 0.5 whereas within the same range of population prevalence, LDSC with intercept estimation will *underestimate* the heritability. The problem of our simulation is that no confounding factors were simulated, thus it is uncertain whether if the same pattern can be observed when there is confounding factors. Nonetheless, as the confounding effects most likely will inflate the summary statistic of the association, the estimates of the heritability will likely to be biased upward. Moreover applying SHREK to the real data, we performed the LD correction. As there is a large amount of SNPs in the real data, the LD correction will inflate the estimates thus ensuring all biases were in the same direction (e.g. inflates our estimates). Because of the uni-directional bias, we can safely hypothesize that the estimates from SHREK in tables 2.5 and 2.6 are the upper-bound for the true SNP heritability in the current GWAS studies.

Based on our estimation, the PGC schizophrenia GWAS can at most account for $\sim 20\%$ of the heritability of SCZ despite the amount of samples included. When compared to the heritability estimated from twin studies, there are around $40\% \sim 60\%$ of missing heritability unaccounted for. This suggested that rare variants or other factors (e.g. copy number variation (CNV)) other than common SNPs can account for the remaining heritability of schizophrenia.

If one would like to estimates the contribution of rare variants to SCZ, new algorithms might be required. This is mainly because the LD estimates form the rare variants usually have a large variability and might not be reliable, thus leading to unreliable estimates from SHREK and LDSC. Therefore special cares are required if one would like to include the rare variants in the estimation process. Also, it is noted that we only performed the estimation on the autosomal chromosomes. The

main reason behind was that there are large difference between male and female on the sex chromosomes (e.g. 2 X for female and XY for male). The proportion of male and female are usually not provided. This leads to difficulties in eq. (2.14) where the sample size is an important factor. Special consideration might therefore be required if one would like to estimates the contribution of variants on the sex chromosome to schizophrenia.

Moreover, considering that the risk of having schizophrenia of individual with a schizophrenic mother or schizophrenic father differs, it is possible that epigenetic or the mitochondrion which were mainly contributed by the mother also have their role in the heritability of schizophrenia. Therefore epigenetic might also have an important role in the etiology of schizophrenia.

Overall, the development of SHREK and LDSC marks a new era in SNP heritability estimation and hopefully, with the continuous advancement of the methodology, problems in LD correction and the liability adjustment can all be solved in the near future.

2.4.4 Limitations and Improvements

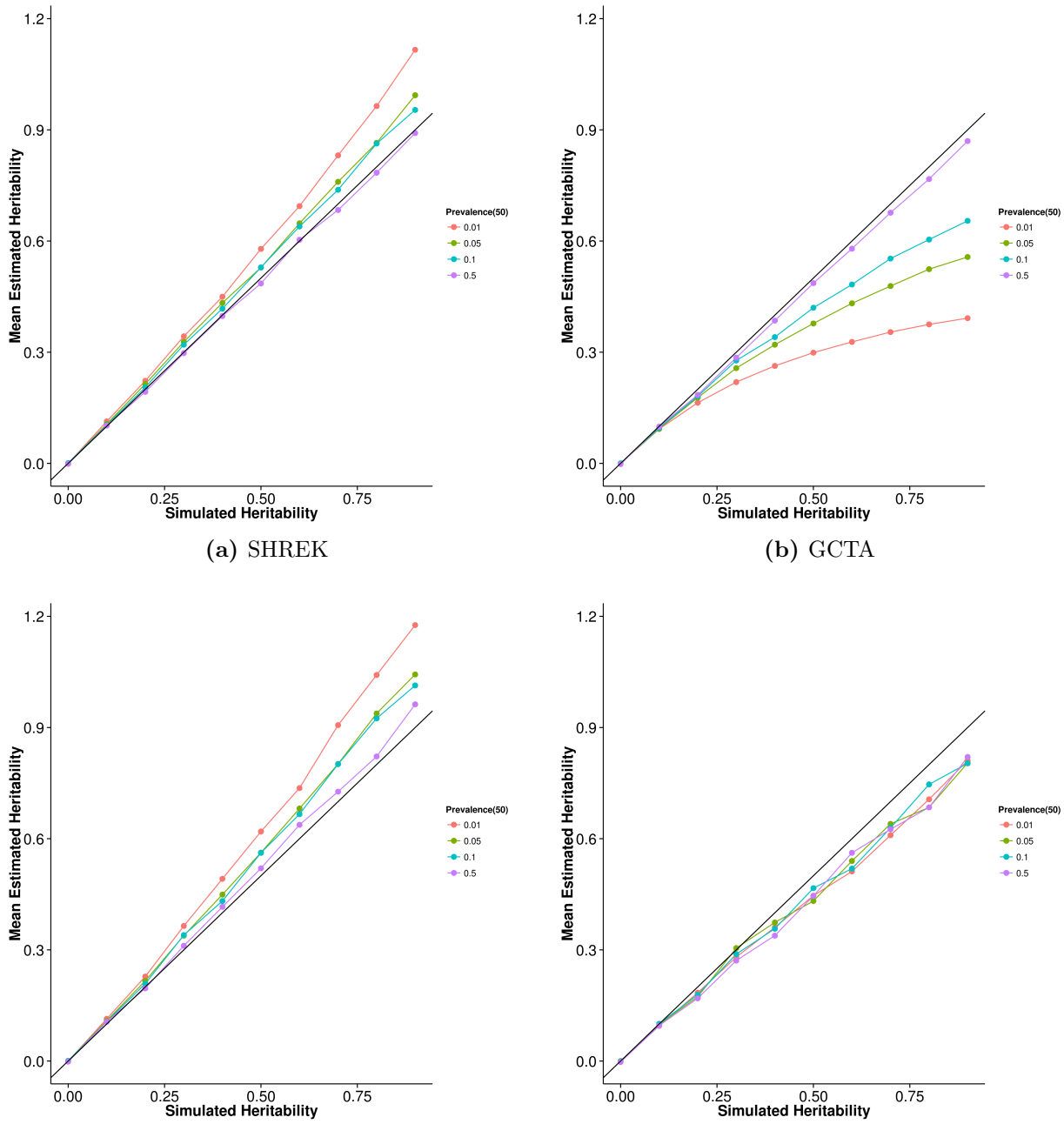
One of the biggest disadvantage of SHREK is its speed when compared to LDSC. To estimate the heritability, SHREK requires the calculation of the inverse of the LD matrix, which is an $O(n^3)$ operation. Although the use of sliding window has significantly reduce the time requirement, the run time will still increase substantially as the density of the SNPs increases. For example, it can take more than 2 days to process one chromosome of the PGC schizophrenia data set, where there can be more than 5,000 SNPs per window. When applying SHREK to the real data, the computation resources required to estimates heritability of the PGC SCZ GWAS were too high, forcing us to reduce the window size for the analysis. To make the

use of SHREK feasible, further development are required to improve the speed of SHREK. An obvious choice might be to use the Armadillo library (Sanderson, 2010) together with the OpenBLAS library which can be more than 3 times faster when compared to the EIGEN C++ library (Ho, 2011).

On the other hand, the inverse of the LD matrix proves to be one of the biggest challenge for SHREK not only because of the time required to invert the matrix, but also the accuracy of the inverse. Due to the inherently high collinearity of the LD matrix, the condition number of the matrix is very high, meaning that small imprecisions in the matrix can be amplified during the analysis. This makes SHREK very sensitive to errors in the LD matrix. The use of tSVD does help to alleviate some of this problem yet it is still possible for it to break. A possible method to reduce the problem of the LD matrix is to remove any SNPs in perfect LD with each other and we are going to implement this feature in SHREK in future release and hopefully an improve in performance can be obtained.

Finally, we do acknowledge that we have not exhaust all possible combinations of genetic architectures in our simulation. For example, one can also test the performance of the algorithms when the observed prevalence was different (e.g. not 50%). It is also possible for one to investigate the effect of number of causal SNPs on the performance of the algorithms when extreme phenotype sampling was performed. However, we do argue that we have performed a substantial amount of simulations and should be able to provide a general concept as to how the performance of SHREK, LDSC and GCTA are affected in the general scenarios.

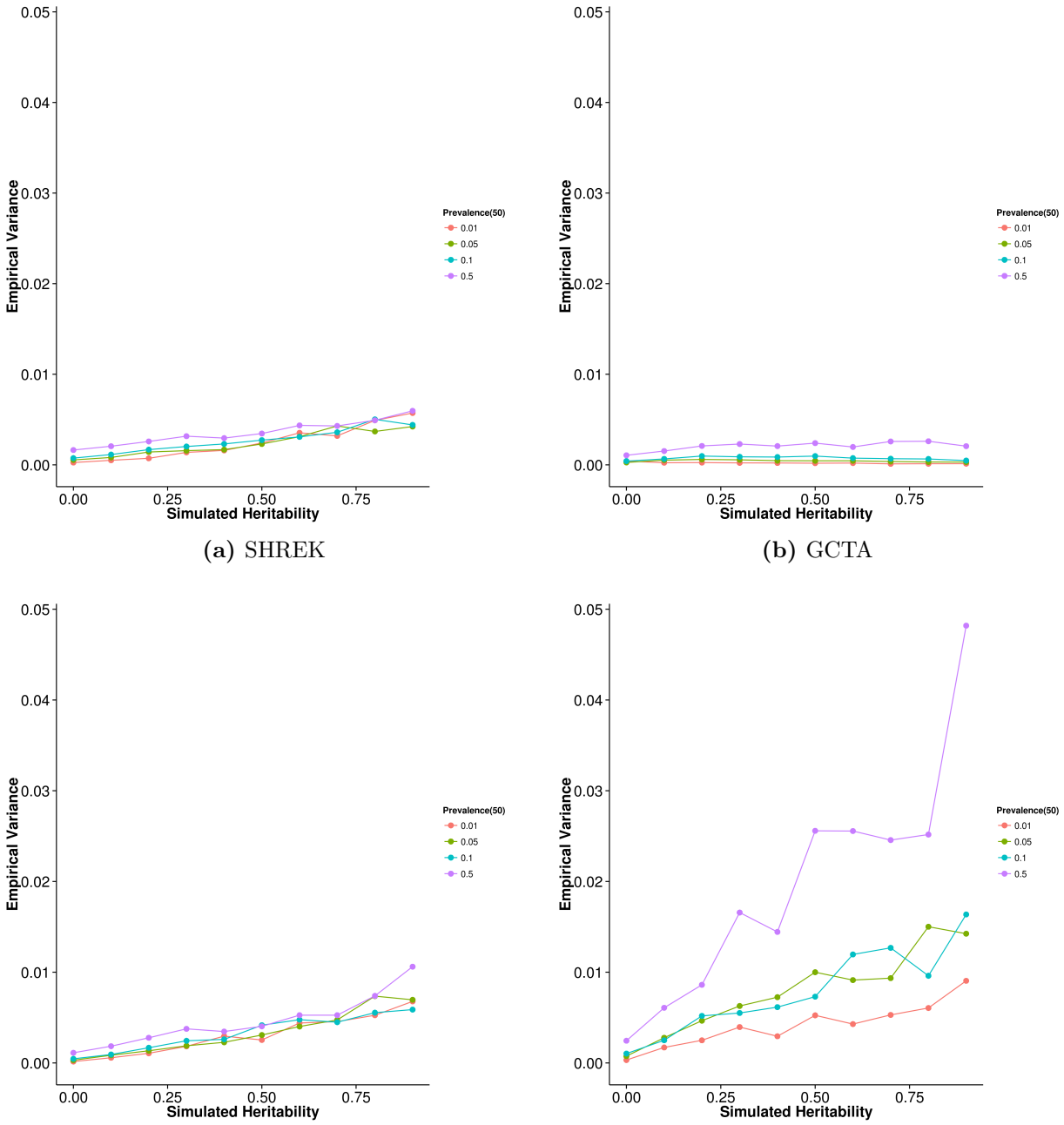
2.5 Supplementary



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

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

Figure 2.17: Mean of results from case control simulation with random effect size simulation with 50 causal SNPs. In general, the results were similar to the scenario with 10 causal SNPs with the only exception that the estimates from LDSC with intercept estimates seems to be less affected by the change in prevalence of the trait.



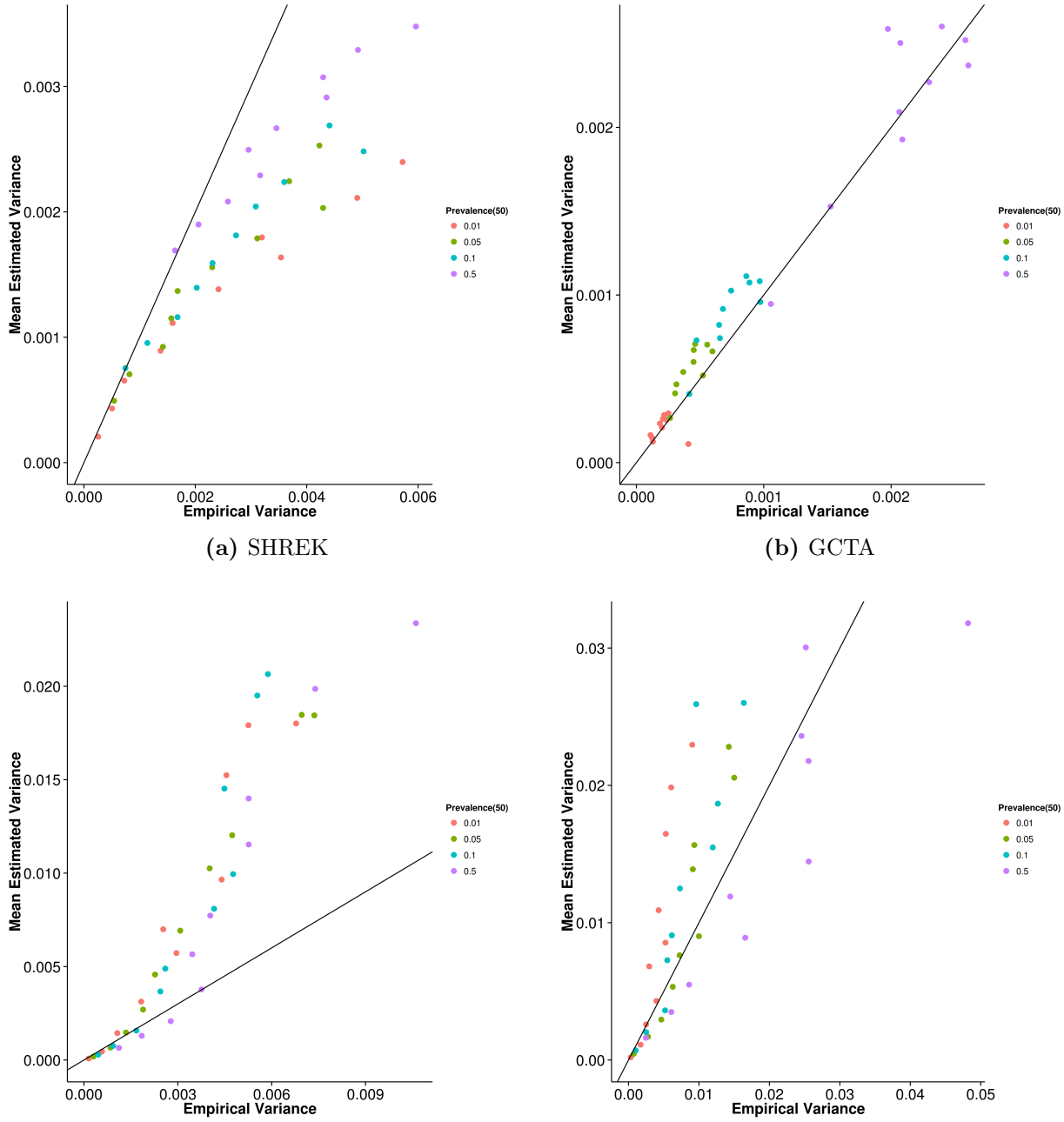
(a) SHREK

(b) GCTA

(c) LDSC with fix intercept

(d) LDSC with intercept estimation

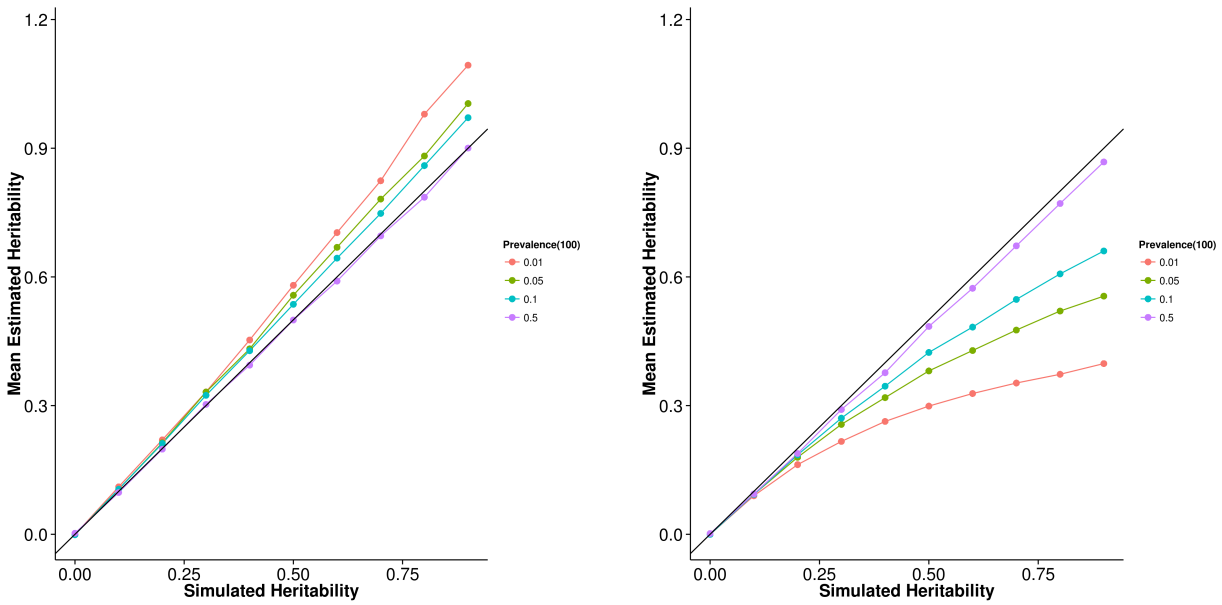
Figure 2.18: Variance of results from case control simulation with random effect size simulation with 50 causal SNPs. For most algorithm except that of LDSC with fixed intercept, the empirical variance of the estimates increases as the population prevalence of the trait increases, with the estimations from LDSC with intercept estimation display the largest variance.



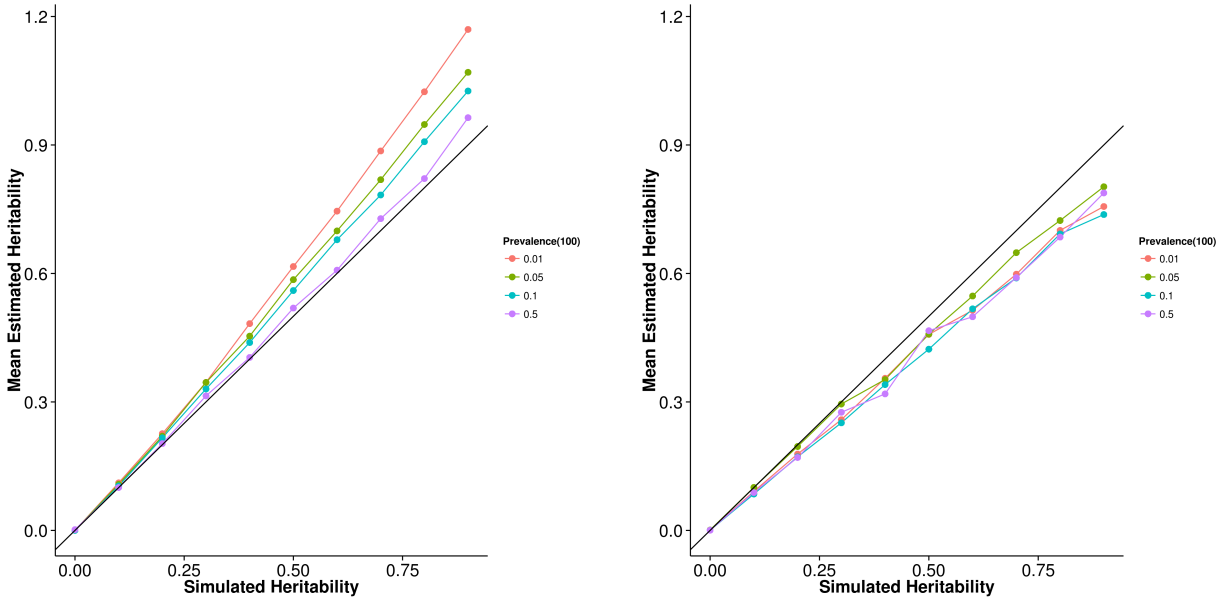
(a) SHREK **(b)** GCTA

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

Figure 2.19: Estimated variance of results from case control simulation with random effect size simulation when compared to empirical variance when 50 causal SNPs was simulated. Again, the estimation of variance from SHREK tends to be downwardly biased and LDSC with fixed intercept tends to be upwardly biased. However, when intercept estimation was performed, the estimation of variance of LDSC improved.

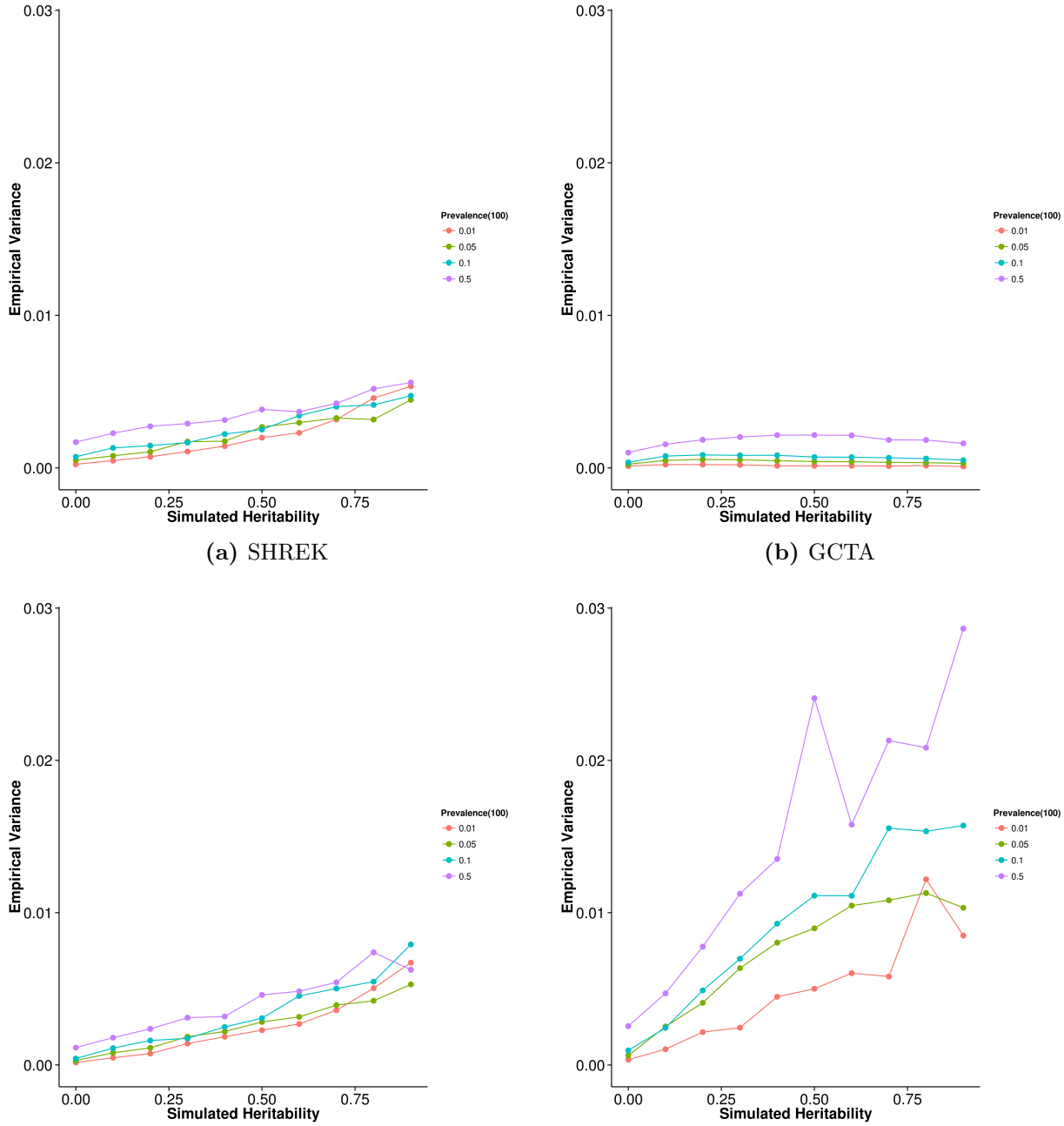


(a) SHREK **(b)** GCTA



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

Figure 2.20: Mean of results from case control simulation with random effect size simulation with 100 causal SNPs. The bias seems to be unaffected by the number of causal SNPs and were the same as what was observed when there were 10 or 50 causal SNPs.



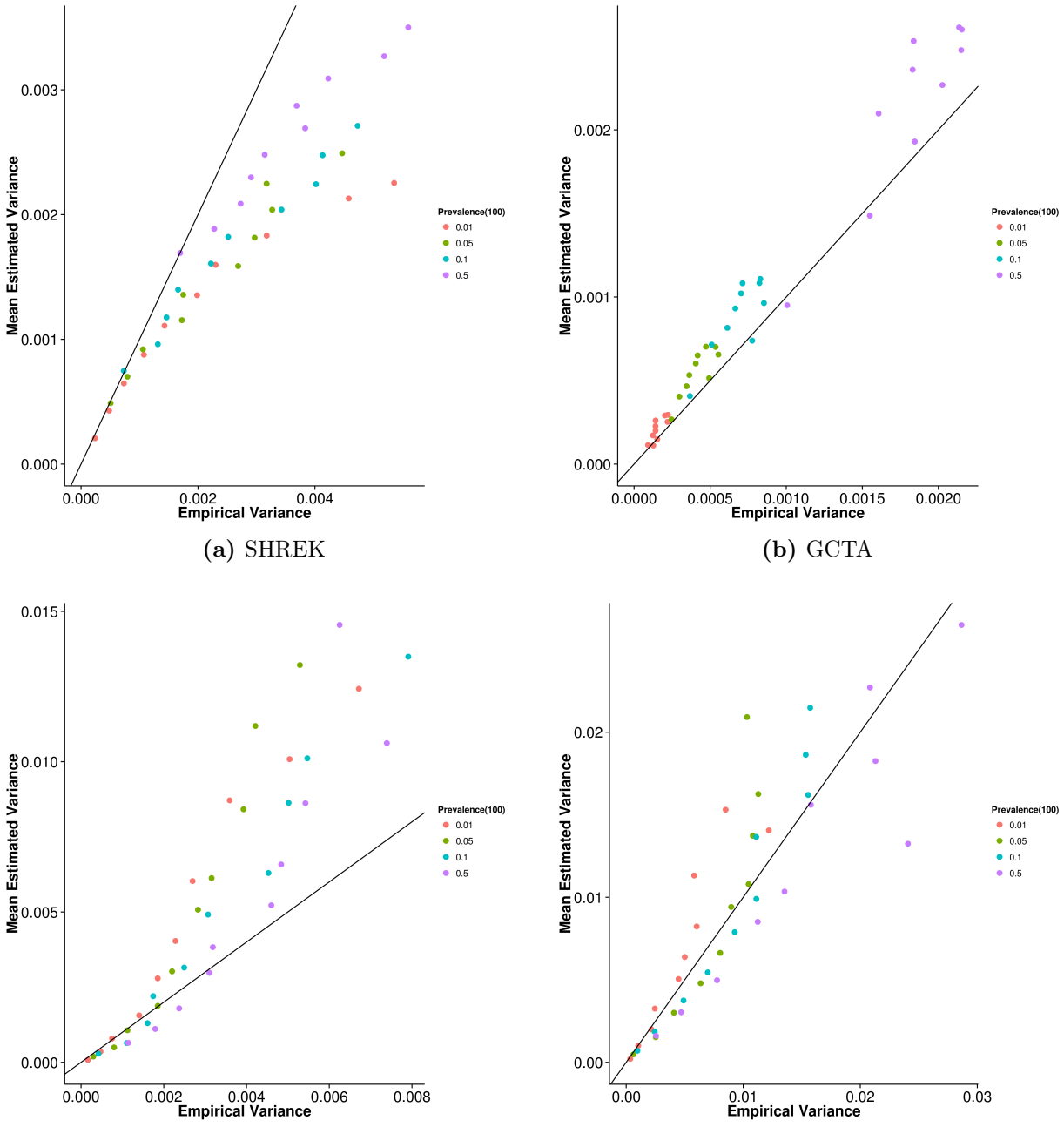
(a) SHREK

(b) GCTA

(c) LDSC with fix intercept

(d) LDSC with intercept estimation

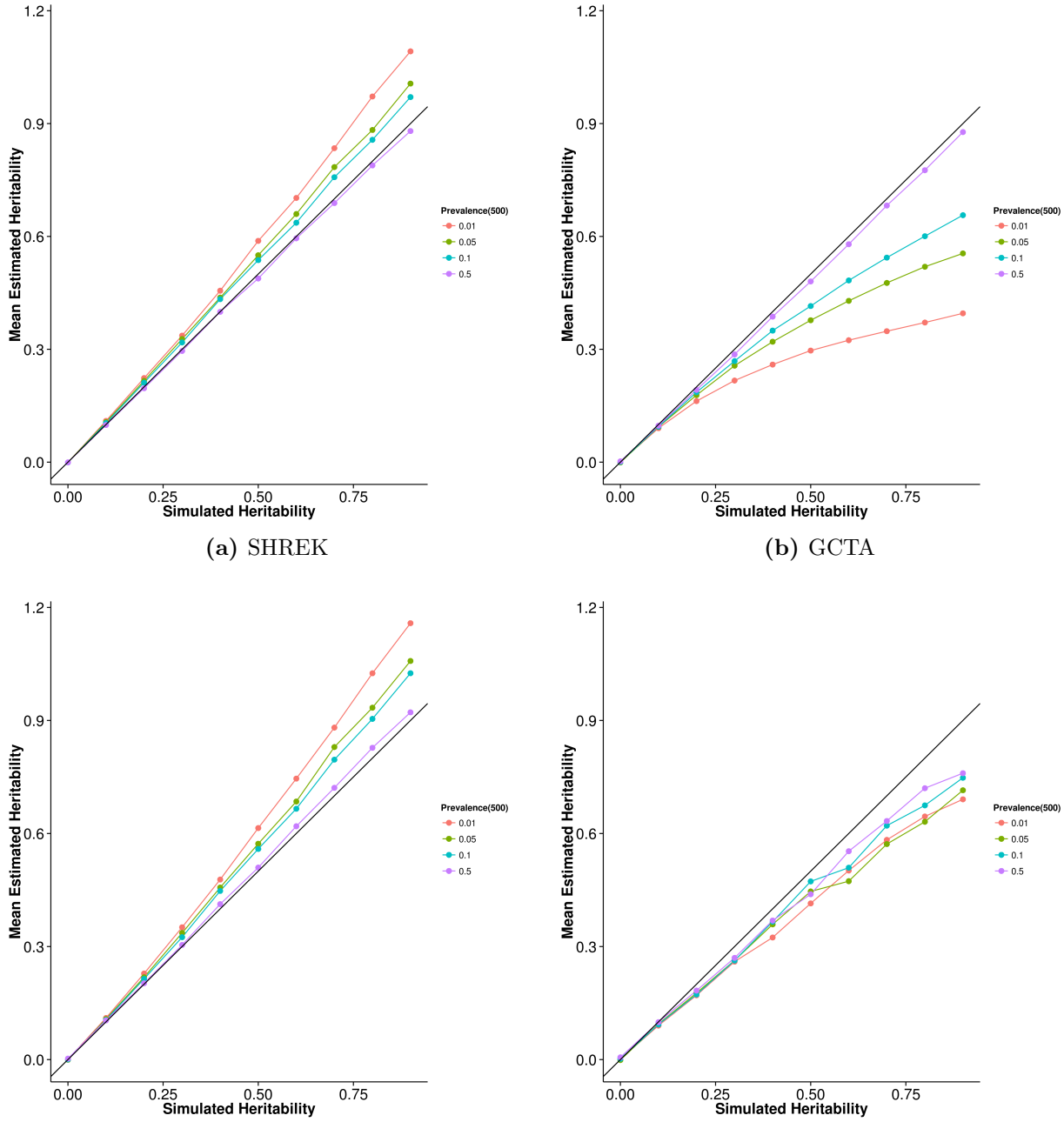
Figure 2.21: Variance of results from case control simulation with random effect size simulation with 100 causal SNPs. As the number of causal SNPs increased to 100, the relationship between the population prevalence and the empirical variance of the algorithms become clear where as the population prevalence increases, the empirical variance of all algorithm increases. Again, LDSC with intercept estimation has the largest variation of all the algorithms and the empirical variance of LDSC with fix intercept is only slightly higher than that of SHREK.



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

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

Figure 2.22: Estimated variance of results from case control simulation with random effect size simulation when compared to empirical variance when 100 causal SNPs was simulated. Once again, SHREK underestimated its empirical variance and LDSC with fixed intercept overestimates its empirical variance. However, the magnitude of overestimation of LDSC with fixed intercept decreased when compared to previous conditions.



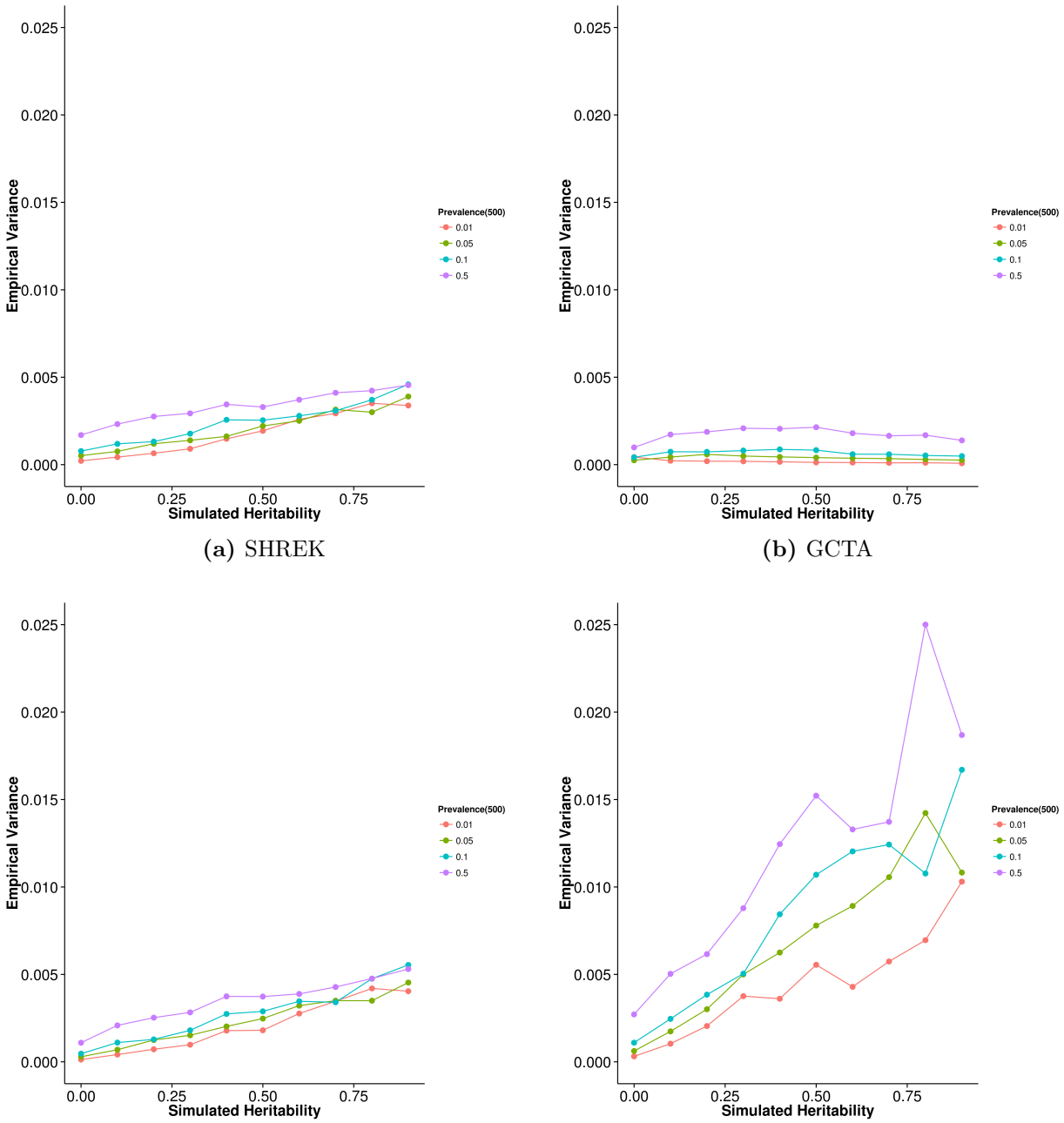
(a) SHREK

(b) GCTA

(c) LDSC with fix intercept

(d) LDSC with intercept estimation

Figure 2.23: Mean of results from case control simulation with random effect size simulation with 500 causal SNPs. Again, a clear pattern of underestimation was observed for GCTA and LDSC with intercept estimation whereas estimations from SHREK and LDSC with fixed intercepts tends to be upwardly biased, with the magnitude of bias increases as the population prevalence decreases.



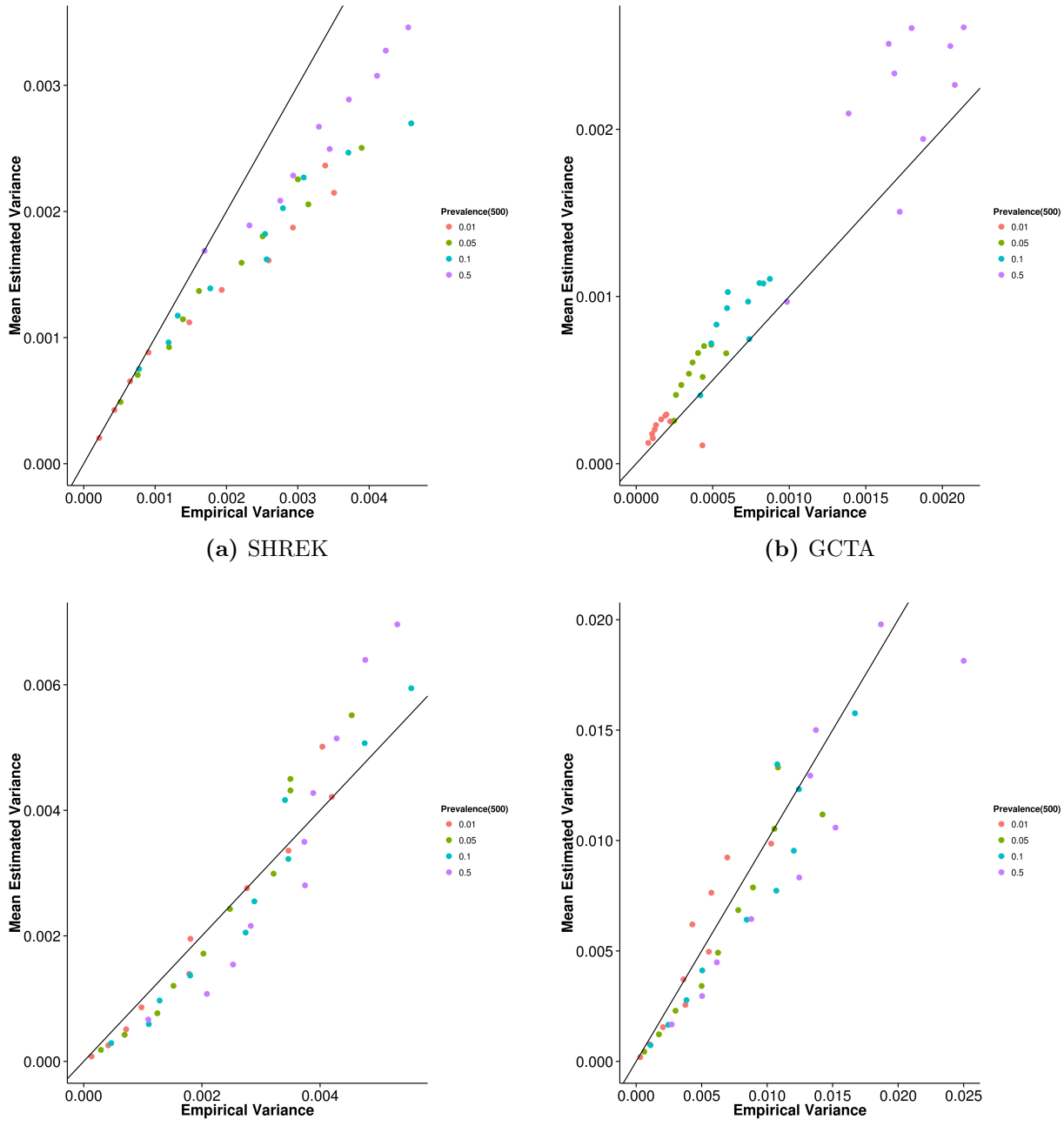
(a) SHREK

(b) GCTA

(c) LDSC with fix intercept

(d) LDSC with intercept estimation

Figure 2.24: Variance of results from case control simulation with random effect size simulation with 500 causal SNPs. As the number of causal SNPs increased to 500, the empirical variance of SHREK and LDSC with fixed intercept converges. However, the empirical variance of LDSC with intercept estimations remains high.



(a) SHREK

(b) GCTA

(c) LDSC with fix intercept

(d) LDSC with intercept estimation

Figure 2.25: Estimated variance of results from case control simulation with random effect size simulation when compared to empirical variance when 500 causal SNPs was simulated. When the trait contains 500 causal SNPs, LDSC begins to provide a good estimation of its own empirical variance both with and without intercept estimation. On the other hand, SHREK's estimation of its own empirical variance remains consistently lower than the true empirical variance.

4 Conclusion

In this thesis, we presented SNP HeRitability Estimation Kit (SHREK), an robust algorithm for the estimation of Single Nucleotide Polymorphism (SNP) heritability using summary statistics from Genome Wide Association Study (GWAS), an alternative to LD SCore regression (LDSC). Through simulations, it was suggested that when compared to LDSC, SHREK can provide a more robust estimate for oligogentic traits and in case-control designs where no confounding variables was present. Using the latest GWAS summary statistics released by the Psychiatric Genomics Consortium (PGC), we estimated that schizophrenia has a SNP-heritability of 0.174 (SD=0.00453), which is similar to the estimate of 0.197 (SD=0.0058) by LDSC.

When compared to the heritability estimated from twin studies (81%) (Sullivan, Kendler, and M. C. Neale, 2003) and large scale population based study (64%) (Lichtenstein et al., 2009), the SNP heritability is much lower, suggesting that factors other than common SNPs might have accounted for the remaining heritability.

On the other hand, we also performed an RNA sequencing on the polyriboinosinic-polyribocytidilic acid (PolyI:C) maternal immune activation (MIA) mouse model to investigate if differential gene expression induced by MIA and genetic variations observed in schizophrenia were acting on the same functional pathway in the development of schizophrenia. We were able to identify a total of 12 pathways that might be perturbed by early MIA events in the cerebellum of the mouse, including calcium ion signaling and pathways related to neural or synaptic functioning. Using

LDSC, it was found that of the 12 significant pathways, 4 pathways related to the extracellular matrix (ECM), mitogen-activated protein kinase (MAPK) signaling, neuronal system and calcium signaling were all contributed disproportionately to the SNP heritability of schizophrenia, suggesting that the differential expression induced by early MIA and the genetic variants associated with schizophrenia might have act upon the same functional pathways in the development of schizophrenia.

Providing that recent study suggest a n-3 polyunsaturated fatty acid (PUFA) rich diet can help to reduce the schizophrenia-like behaviour in mouse exposed to early MIA events (Q. Li, Leung, et al., 2015), we also investigated how the n-3 PUFA rich diet affect the gene expression pattern in the adult cerebellum. *Sgk1*, a gene that regulates the glutamatergic system, were found to be significant in PolyI:C exposed mouse given different diet. Moreover, we found that pathway related to ECM were affected not only by MIA, but also in PolyI:C samples given different diets. It is therefore possible that the ECM pathway or genes within the ECM pathway might have mediated the effect of n-3 PUFA diet on MIA exposed mouse, making them an important target for further research.

4.1 Challenge in SNP-Heritability Estimation

Although it is now possible to estimates the SNP heritability based on the summary statistic from GWAS, a lot of questions remain unanswered in the estimation of SNP heritability. One major problem of SHREK and LDSC is that they both heavily relies on the Linkage Disequilibrium (LD) structures from the reference panel. However, GWAS samples can come from large variety of ethnic background thus the LD pattern estimated from the reference panel might not be representative of the sample LD. If there is a significant difference between the LD pattern from the reference panel and the LD pattern from the samples, both SHREK and LDSC would fail to

provide an accurate estimate. For example, if a GWAS is conducted with 50% European and 50% African, population stratification may confound the results. Even if one control for the population stratification using the principle component analysis (PCA), the question remains whether if one should use the African reference panel or the European reference panel in the estimation of SNP heritability. Moreover, information regarding the population stratification (e.g. the Principle Component (PC)) were usually unavailable, making the problem more complicated. Although LDSC claims that it can delineate the population confounding factors from the SNP heritability, it is expected that different reference panel will generate different LD score, thus leads to different estimates. In our example of 50% European and 50% African, it is unclear which reference panel should be used and one can expect that as the pattern of population mixing becomes more complicated, it will be increasingly difficult to obtain a representative reference panel and thus the estimation of SNP heritability can be difficult. Further researches are therefore required to tackle the problem of population stratification before one can confidently estimate the SNP heritability from summary statistics from GWAS that might contain samples from large variety of ethnic background.

An important observation in our simulation study was that there was a general bias observed in all the SNP-heritability estimation algorithm under the case control scenario. This is likely due to the ascertainment bias introduced through case control sampling. Although the liability adjustment was performed, bias was still observed. This suggested that we will need a better liability adjustment algorithm if we would like to accurately estimate the SNP-heritability from case control studies.

As technology advances, researchers can now use the next generation sequencing (NGS) technology to sequence the genome at per base resolution. This brings great prospect in the genetic studies for now we can directly identify the causal variants and can even detect rare causal variants providing sufficient sample

size. However, both SHREK and LDSC are designed to work on the summary statistics from GWAS where common SNPs are usually the focus. Because of the huge sampling error associated with rare variants, the LD calculated for rare variants usually has a larger standard error (SE). As SHREK and LDSC are both heavily rely on an accurate LD estimation, they might be unsuitable for the estimation of the contribution of rare variants to schizophrenia. In fact, it was found that when all causal variants are rare (minor allele frequency (maf) $< 1\%$), LDSC will often generate a negative slope, and the intercept will exceed the mean χ^2 statistic (B. K. Bulik-Sullivan et al., 2015). As a result of that, a different algorithm must be developed in order to estimates the heritability from rare variants using only summary statistics.

4.2 Schizophrenia: Future Perspectives

With the success of the PGC schizophrenia GWAS, research in schizophrenia genetics has finally entered an era of success. Through international collaboration, the PGC has finally identified 108 genetic loci that were associated with schizophrenia using GWAS approach (Stephan Ripke, B. M. Neale, et al., 2014). However, the actual causal variants have not been identified. Functional analysis of these associated variants, and their contribution to the etiology of schizophrenia will become an important topic for further research in schizophrenia genetics.

On the other hand, when estimating the SNP-heritability of schizophrenia, it was found that no more than 20% of the heritability has been accounted for by the current GWAS which is lower than the 81% estimated based on twin studies (Sullivan, Kendler, and M. C. Neale, 2003). This suggested that factors other than common SNP were contributing to the heritability of schizophrenia.

Clear evidences suggested that schizophrenia patients has a higher mor-

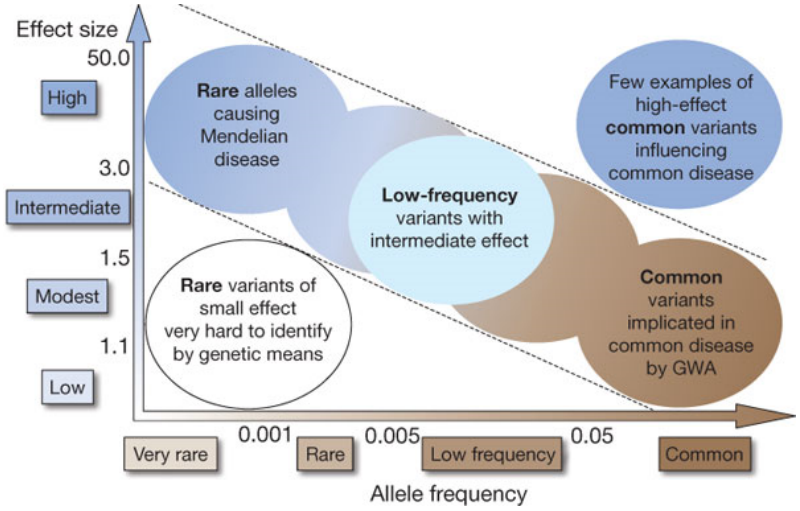


Figure 4.1: Relationship between effect size and allele frequency. It is expected that rare variants with large effect size were actively selected against in the population and therefore should be rare.

tality than the general population (Saha, Chant, and Mcgrath, 2007). Given this strong selective pressure, it is likely that the causal variants of schizophrenia with large effect size will be selected against in the population. As a result of that, causal variants with large effect size are likely to be rare (fig. 4.1). With the technological advancement in NGS, we are now able to investigate the human genome at per base resolution using Exome Sequencing and even Whole Genome Sequencing technology. Recent study by S M Purcell et al. (2014) was able to identify gene sets enriched by rare variants that were associated with schizophrenia using Exome Sequencing. This demonstrate the power of the sequencing technology in the identification of possible risk variants. Moreover, there was overlaps observed between genes harboring rare risk variants and those within the PGC schizophrenia GWAS (S M Purcell et al., 2014), suggesting that the rare variants and common variants studies are complementing each other. As more resources are devoted in to sequencing the genome of schizophrenia patients, more rare variants associated with schizophrenia are expected to be identified.

Currently, most of the focus in schizophrenia was directed to genetic variation yet it is possible that the heritability of schizophrenia is also transmitted in the

form of epigenetic changes such as methylation. It was observed that the risk for individual born from a schizophrenic mother is larger than that from a schizophrenic father. This suggests that maternal specific elements, such as maternal imprinting and mitochondria might account for part of the risk of schizophrenia. Epigenetic studies in schizophrenia (Wockner et al., 2014; Nishioka et al., 2012) has identified genes with differential DNA methylation patterns associated with schizophrenia, suggesting the importance of epigenetics in the etiology of schizophrenia.

As a highly heritable disorder, most of the research of schizophrenia has been focusing on the genetic factors. Although the genetic variation accounted for majority of the variations in schizophrenia, the environmental factors, especially prenatal infection is also an important factor to consider. It was estimated that prenatal infection accounts for roughly 33% of all schizophrenia cases (A S Brown and Derkits, 2010). The MIA rodent model has provide vital information on the possible interaction between the immune and neuronal system in the etiology of schizophrenia (U Meyer, Yee, and J Feldon, 2007). For example, Interleukin-6 (IL-6), a pro-inflammatory cytokine has been found to be an important mediator in generating the schizophrenia-like behaviour in rodent model (Smith et al., 2007). More importantly, there are evidence of the interaction between prenatal infection and genetic variation, supporting a mechanism of gene-environment interaction in the causation of schizophrenia (Clarke et al., 2009). As the SNP-heritability estimation does not take into account of the gene environmental interactions, it is possible that the "missing" heritability can be due to gene-environmental interactions. Efforts is now made by the European network of national schizophrenia networks studying Gene-Environmental Interaction (EUGEI) to identify possible genetic and environmental interaction that contributes to the disease etiology of schizophrenia.

With the sophistication of technologies, we can now perform whole genome sequencing with the HiSeq X Ten system costing less than $1,000. Therefore, the

largest challenge now resides in how to make sense of the data instead of data generation. For example, the alignment of sequence read to low complexity sequence or low-degeneracy repeats remains challenging and might be error prone, thus have a negative impact to the quality of the results(Sims et al., 2014). New sequencing technology such as Oxford Nanopore which can provide extra long-reads, might help to make alignment easier due to the extra information for each individual reads. However, the Oxford Nanopore is still under development and has a relatively high error rate (Mikheyev and Tin, 2014). Only until the error rate is dramatically decreased can the use of Oxford Nanopore system become feasible.

Even if the reads can be perfectly aligned to the genome, the functional annotation of variants remains challenging. When it comes to complex disease such as schizophrenia, there can be a lot of causal variants observed throughout the genome yet currently one can only provide estimates of the functional impact of variants on the exomic regions. The development of ENCODE project (ENCODE Project Consortium, 2012) and Genotype-Tissue Expression (GTEx) project (T. G. Consortium, 2015) have helped provide reference point for the annotation of genetic variations in the intergenic regions yet there are still many genetic variation in the genome where their function remains unknown. Only through the tireless effort of the molecular biologist can we gain sufficient information required to make sense of the sequencing data obtained.

In conclusion, we have only catch a glimpse of the etiology of schizophrenia and there are still a lot of questions left unanswered. It is expected that only by combining the study of epigenetic, genomic variation, gene expressions, and gene environmental interaction can provide a deeper understanding of the complex disease mechanism of schizophrenia be obtained.

Bibliography

Altshuler, David M et al. (2010). "Integrating common and rare genetic variation in diverse human populations." In: *Nature* 467.7311, pp. 52–58 (cit. on pp. 52, 54).

American Psychiatric Association (2013). *Diagnostic and Statistical Manual of Mental Disorders*. American Psychiatric Publishing, p. 991.

Anders, S and W Huber (2010). "Differential expression analysis for sequence count data". eng. In: *Genome Biol* 11.10, R106.

Andreasen, Nancy C and Ronald Pierson (2008). "The role of the cerebellum in schizophrenia." eng. In: *Biological psychiatry* 64.2, pp. 81–88.

Andrews, S. *FastQC A Quality Control tool for High Throughput Sequence Data*.

Bergeron, J D et al. (2013). "White matter injury and autistic-like behavior predominantly affecting male rat offspring exposed to group B streptococcal maternal inflammation". eng. In: *Dev Neurosci* 35.6, pp. 504–515.

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

Berretta, S (2012). "Extracellular matrix abnormalities in schizophrenia". eng. In: *Neuropharmacology* 62.3, pp. 1584–1597.

Berridge, Michael J (2014). "Calcium signalling and psychiatric disease: bipolar disorder and schizophrenia." eng. In: *Cell and tissue research* 357.2, pp. 477–492.

Bohmer, Christoph et al. (2004). "Stimulation of the EAAT4 glutamate transporter by SGK protein kinase isoforms and PKB." eng. In: *Biochemical and biophysical research communications* 324.4, pp. 1242–1248.

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

Brown, A S and E J Derkits (2010). "Prenatal infection and schizophrenia: a review of epidemiologic and translational studies". eng. In: *Am J Psychiatry* 167.3, pp. 261–280 (cit. on p. 154).

Brown, Alan S (2012). "Epidemiologic studies of exposure to prenatal infection and risk of schizophrenia and autism." eng. In: *Developmental neurobiology* 72.10, pp. 1272–1276.

Bulik-Sullivan, Brendan (2015). *Replicating MDD heritability Estimation* (cit. on pp. 100, 101).

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 (cit. on pp. 39, 40, 57, 66, 72, 100, 101, 152).

Busby, Michele A et al. (2013). "Scotty: a web tool for designing RNA-Seq experiments to measure differential gene expression". In: *Bioinformatics* 29.5, pp. 656–657.

Cadenhead, K S et al. (2000). "Modulation of the startle response and startle laterality in relatives of schizophrenic patients and in subjects with schizotypal personality disorder: evidence of inhibitory deficits." eng. In: *The American journal of psychiatry* 157.10, pp. 1660–1668.

Cingolani, Pablo et al. (2012). "A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w(1118); iso-2; iso-3". In: *Fly* 6.2, pp. 80–92.

Clandinin, M T (1999). “Brain development and assessing the supply of polyunsaturated fatty acid.” eng. In: *Lipids* 34.2, pp. 131–137.

Clarke, Mary C et al. (2009). “Evidence for an interaction between familial liability and prenatal exposure to infection in the causation of schizophrenia.” eng. In: *The American journal of psychiatry* 166.9, pp. 1025–1030 (cit. on p. 154).

Cline, H (2005). “Synaptogenesis: a balancing act between excitation and inhibition”. eng. In: *Curr Biol* 15.6, R203–5.

Consortium, The GTEx (2015). “The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans”. In: *Science* 348.6235, pp. 648–660 (cit. on p. 155).

Consortium, The International HapMap (2005). “A haplotype map of the human genome”. In: *Nature* 437, pp. 1299–1320.

Costa, E et al. (2001). “Dendritic spine hypoplasticity and downregulation of reelin and GABAergic tone in schizophrenia vulnerability.” eng. In: *Neurobiology of disease* 8.5, pp. 723–742.

Derosa, G et al. (2009). “Effects of long chain omega-3 fatty acids on metalloproteinases and their inhibitors in combined dyslipidemia patients.” eng. In: *Expert opinion on pharmacotherapy* 10.8, pp. 1239–1247.

Deverman, B E and P H Patterson (2009). “Cytokines and CNS development”. eng. In: *Neuron* 64.1, pp. 61–78.

Dobin, A et al. (2013). “STAR: ultrafast universal RNA-seq aligner”. eng. In: *Bioinformatics* 29.1, pp. 15–21.

Eastwood, S L et al. (2003). “The axonal chemorepellant semaphorin 3A is increased in the cerebellum in schizophrenia and may contribute to its synaptic pathology.” eng. In: *Molecular psychiatry* 8.2, pp. 148–155.

ENCODE Project Consortium (2012). “An integrated encyclopedia of DNA elements in the human genome”. In: *Nature* 489.7414, pp. 57–74 (cit. on p. 155).

Engstrom, Par G et al. (2013). "Systematic evaluation of spliced alignment programs for RNA-seq data". In: *Nat Meth* 10.12, pp. 1185–1191.

Falconer, Douglas S (1965). "The inheritance of liability to certain diseases, estimated from the incidence among relatives". In: *Annals of Human Genetics* 29.1, pp. 51–76.

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

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

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

Fromer, M et al. (2014). "De novo mutations in schizophrenia implicate synaptic networks". eng. In: *Nature* 506.7487, pp. 179–184.

Garbett, K a et al. (2012). "Effects of maternal immune activation on gene expression patterns in the fetal brain". In: *Translational Psychiatry* 2.4, e98.

Gilad, Yoav and Orna Mizrahi-Man (2015). "A reanalysis of mouse ENCODE comparative gene expression data." eng. In: *F1000Research* 4, p. 121.

Giles, Peter J and David Kipling (2003). "Normality of oligonucleotide microarray data and implications for parametric statistical analyses." eng. In: *Bioinformatics (Oxford, England)* 19.17, pp. 2254–2262.

Giovanoli, S. et al. (2013). "Stress in puberty unmasks latent neuropathological consequences of prenatal immune activation in mice". eng. In: *Science* 339.6123, pp. 1095–1099.

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 (cit. on pp. 39, 78, 81, 95).

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

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

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

— (1967b). "A polygenic theory of schizophrenia". In: *Proceedings of the National Academy of Sciences* 58.1, pp. 199–205.

Guennebaud, Gaël, Benoît Jacob, et al. (2010). *Eigen v3*. http://eigen.tuxfamily.org.

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 (cit. on pp. 49, 63).

Gui, Hongsheng et al. (2013). "RET and NRG1 interplay in Hirschsprung disease." eng. In: *Human genetics* 132.5, pp. 591–600 (cit. on p. 60).

Hansen, Per Christian (1987). "The truncated SVD as a method for regularization". In: *Bit* 27.4, pp. 534–553 (cit. on p. 52).

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

Hennessy, Bryan T et al. (2005). "Exploiting the PI3K/AKT Pathway for Cancer Drug Discovery". In: *Nat Rev Drug Discov* 4.12, pp. 988–1004.

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

Hinrichs, A S et al. (2006). "The UCSC Genome Browser Database: update 2006." eng. In: *Nucleic acids research* 34.Database issue, pp. D590–8 (cit. on p. 65).

Ho, Nghia (2011). *OPENCV VS. ARMADILLO VS. EIGEN ON LINUX* (cit. on p. 104).

Hoyle, David C et al. (2002). “Making sense of microarray data distributions.” eng. In: *Bioinformatics (Oxford, England)* 18.4, pp. 576–584.

Jiang, Lichun et al. (2011). “Synthetic spike-in standards for RNA-seq experiments”. In: *Genome Research* 21.9, pp. 1543–1551.

Kavazos, Kristyn et al. (2015). “Dietary supplementation with omega-3 polyunsaturated fatty acids modulate matrix metalloproteinase immunoreactivity in a mouse model of pre-abdominal aortic aneurysm.” eng. In: *Heart, lung & circulation* 24.4, pp. 377–385.

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

Kim, Daehwan et al. (2013). “TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions”. In: *Genome Biology* 14.4, R36.

Kitajka, Klára et al. (2002). “The role of n-3 polyunsaturated fatty acids in brain: Modulation of rat brain gene expression by dietary n-3 fatty acids”. In: *Proceedings of the National Academy of Sciences of the United States of America* 99.5, pp. 2619–2624.

Knable, M B and D R Weinberger (1997). “Dopamine, the prefrontal cortex and schizophrenia.” eng. In: *Journal of psychopharmacology (Oxford, England)* 11.2, pp. 123–131.

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

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

Lang, Florian, Christoph Böhmer, et al. (2006). “(Patho)physiological Significance of the Serum- and Glucocorticoid-Inducible Kinase Isoforms”. In: *Physiological Reviews* 86.4, pp. 1151–1178.

Lang, Florian, Nathalie Strutz-Seebohm, et al. (2010). "Significance of SGK1 in the regulation of neuronal function". In: *The Journal of Physiology* 588.18, pp. 3349–3354.

Lee, Eminy H Y et al. (2003). "Enrichment enhances the expression of sgk, a glucocorticoid-induced gene, and facilitates spatial learning through glutamate AMPA receptor mediation." eng. In: *The European journal of neuroscience* 18.10, pp. 2842–2852.

Lein, E S et al. (2007). "Genome-wide atlas of gene expression in the adult mouse brain". eng. In: *Nature* 445.7124, pp. 168–176.

Li, Bo and Colin N Dewey (2011). "RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome." eng. In: *BMC bioinformatics* 12, p. 323.

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 (cit. on p. 47).

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. 55).

Li, Q, C Cheung, R Wei, V Cheung, et al. (2010). "Voxel-based analysis of postnatal white matter microstructure in mice exposed to immune challenge in early or late pregnancy". eng. In: *Neuroimage* 52.1, pp. 1–8.

Li, Q, C Cheung, R Wei, E S Hui, et al. (2009). "Prenatal immune challenge is an environmental risk factor for brain and behavior change relevant to schizophrenia: evidence from MRI in a mouse model". eng. In: *PLoS One* 4.7, e6354.

Li, Q, Y O Leung, et al. (2015). "Dietary supplementation with n-3 fatty acids from weaning limits brain biochemistry and behavioural changes elicited by prenatal

exposure to maternal inflammation in the mouse model." eng. In: *Translational psychiatry* 5, e641 (cit. on p. 150).

Liao, Yang, Gordon K Smyth, and Wei Shi (2014). "featureCounts: an efficient general purpose program for assigning sequence reads to genomic features." eng. In: *Bioinformatics (Oxford, England)* 30.7, pp. 923–930.

Lichtenstein, Paul et al. (2009). "Common genetic determinants of schizophrenia and bipolar disorder in Swedish families: a population-based study". In: *The Lancet* 373.9659, pp. 234–239 (cit. on p. 149).

Lidow, Michael S (2003). "Calcium signaling dysfunction in schizophrenia: a unifying approach." eng. In: *Brain research. Brain research reviews* 43.1, pp. 70–84.

Love, Michael I, Wolfgang Huber, and Simon Anders (2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." eng. In: *Genome biology* 15.12, p. 550.

Maloku, Ekrem et al. (2010). "Lower number of cerebellar Purkinje neurons in psychosis is associated with reduced reelin expression". In: *Proceedings of the National Academy of Sciences* 107.9, pp. 4407–4411.

Marioni, J C et al. (2008). "RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays". eng. In: *Genome Res* 18.9, pp. 1509–1517.

Martin, Marcel (2011). "Cutadapt removes adapter sequences from high-throughput sequencing reads". In: *EMBnet.journal; Vol 17, No 1: Next Generation Sequencing Data Analysis*.

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

Mednick (1988). "Schizophrenia Following Prenatal Exposure to an Influenza Epidemic". In: *Arch Gen Psychiatry* 45.1.

Meyer, U, B K Yee, and J Feldon (2007). "The neurodevelopmental impact of prenatal infections at different times of pregnancy: the earlier the worse?" eng. In: *Neuroscientist* 13.3, pp. 241–256 (cit. on p. 154).

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

Mikheyev, Alexander S and Mandy M Y Tin (2014). "A first look at the Oxford Nanopore MinION sequencer." eng. In: *Molecular ecology resources* 14.6, pp. 1097–1102 (cit. on p. 155).

Neumaier, Arnold (1998). "Solving Ill-Conditioned and Singular Linear Systems: A Tutorial on Regularization". In: *SIAM Review* 40.3, pp. 636–666 (cit. on p. 50).

Nishioka, Masaki et al. (2012). "DNA methylation in schizophrenia: progress and challenges of epigenetic studies." eng. In: *Genome medicine* 4.12, p. 96 (cit. on p. 154).

Nugent, Tom F. et al. (2007). "Dynamic mapping of hippocampal development in childhood onset schizophrenia". In: *Schizophrenia Research* 90.1-3, pp. 62–70.

O'Callaghan, E et al. (1991). "Season of birth in schizophrenia. Evidence for confinement of an excess of winter births to patients without a family history of mental disorder." eng. In: *The British journal of psychiatry : the journal of mental science* 158, pp. 764–769.

Olivo, Susan E and Leena Hilakivi-Clarke (2005). "Opposing effects of prepubertal low- and high-fat n-3 polyunsaturated fatty acid diets on rat mammary tumorigenesis." eng. In: *Carcinogenesis* 26.9, pp. 1563–1572.

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. 55).

Oskviga, Devon B. et al. (2012). "Maternal immune activation by LPS selectively alters specific gene expression profiles of interneuron migration and oxidative stress in the fetus without triggering a fetal immune response". In: *Brain, Behavior, and Immunity* 26.4, pp. 623–634.

Peloso, Gina M et al. (2015). "Phenotypic extremes in rare variant study designs." ENG. In: *European journal of human genetics : EJHG* (cit. on p. 82).

Perlstein, W M et al. (2001). "Relation of prefrontal cortex dysfunction to working memory and symptoms in schizophrenia." eng. In: *The American journal of psychiatry* 158.7, pp. 1105–1113.

Project, Genomes et al. (2012). "An integrated map of genetic variation from 1,092 human genomes". In: *Nature* 491.7422, pp. 56–65 (cit. on pp. 54, 65).

Psychiatric GWAS Consortium Bipolar Disorder Working Group (2011). "Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4." eng. In: *Nature genetics* 43.10, pp. 977–983 (cit. on p. 65).

Purcell, S M et al. (2014). "A polygenic burden of rare disruptive mutations in schizophrenia". eng. In: *Nature* 506.7487, pp. 185–190 (cit. on p. 153).

Purcell, S, S S Cherny, and P C Sham (2003). "Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits". en. In: *Bioinformatics* 19, pp. 149–150 (cit. on p. 63).

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 (cit. on pp. 56, 59).

Reeves, P G, F H Nielsen, and G C Jr Fahey (1993). *AIN-93 purified diets for laboratory rodents: final report of the American Institute of Nutrition ad hoc writing committee on the reformulation of the AIN-76A rodent diet.* eng.

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

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

Ripke, Stephan, Benjamin M. Neale, et al. (2014). "Biological insights from 108 schizophrenia-associated genetic loci". In: *Nature* 511, pp. 421–427 (cit. on pp. 65, 88, 152).

Ripke, Stephan, Naomi R Wray, et al. (2013). "A mega-analysis of genome-wide association studies for major depressive disorder." eng. In: *Molecular psychiatry* 18.4, pp. 497–511 (cit. on p. 65).

Ripke, S et al. (2013). "Genome-wide association analysis identifies 13 new risk loci for schizophrenia". eng. In: *Nat Genet* 45.10, pp. 1150–1159.

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

Robinson, M D, D J McCarthy, and G K Smyth (2010). "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data". eng. In: *Bioinformatics* 26.1, pp. 139–140.

Saha, Sukanta, David Chant, and John Mcgrath (2007). "A Systematic Review of Mortality in Schizophrenia". In: *Archives of general psychiatry* 64.10, pp. 1123–1131 (cit. on p. 153).

Sanderson, Conrad (2010). *Armadillo: An Open Source C++ Linear Algebra Library for Fast Prototyping and Computationally Intensive Experiments*. Tech. rep. (cit. on p. 104).

Seyednasrollah, Fatemeh, Asta Laiho, and Laura L Elo (2015). "Comparison of software packages for detecting differential expression in RNA-seq studies". In: *Briefings in Bioinformatics* 16.1, pp. 59–70.

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 (cit. on pp. 49, 50, 63, 64, 97).

Sims, David et al. (2014). "Sequencing depth and coverage: key considerations in genomic analyses". In: *Nat Rev Genet* 15.2, pp. 121–132 (cit. on p. 155).

Smith, S E et al. (2007). "Maternal immune activation alters fetal brain development through interleukin-6". eng. In: *J Neurosci* 27.40, pp. 10695–10702 (cit. on p. 154).

Stamenkovic, Ivan (2003). "Extracellular matrix remodelling: the role of matrix metalloproteinases." eng. In: *The Journal of pathology* 200.4, pp. 448–464.

Su, Zhan, Jonathan Marchini, and Peter Donnelly (2011). "HAPGEN2: Simulation of multiple disease SNPs". In: *Bioinformatics* 27.16, pp. 2304–2305 (cit. on pp. 53, 55, 58).

Subramanian, Aravind et al. (2005). "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles". In: *Proceedings of the National Academy of Sciences* 102.43, pp. 15545–15550.

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

Sullivan, Patrick F, Kenneth S Kendler, and Michael C Neale (2003). "Schizophrenia as a Complex Trait". In: *Archives of general psychiatry* 60, pp. 1187–1192 (cit. on pp. 149, 152).

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

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

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

Trapnell, Cole et al. (2012). “Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks”. In: *Nat. Protocols* 7.3, pp. 562–578.

Trebble, Timothy et al. (2003). “Inhibition of tumour necrosis factor-alpha and interleukin 6 production by mononuclear cells following dietary fish-oil supplementation in healthy men and response to antioxidant co-supplementation.” eng. In: *The British journal of nutrition* 90.2, pp. 405–412.

Tsai, Kuen J et al. (2002). “sgk, a primary glucocorticoid-induced gene, facilitates memory consolidation of spatial learning in rats.” eng. In: *Proceedings of the National Academy of Sciences of the United States of America* 99.6, pp. 3990–3995.

Velakoulis, Dennis et al. (2006). “Hippocampal and amygdala volumes according to psychosis stage and diagnosis”. In: *Archives of general psychiatry* 63, pp. 139–149.

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

Vogel, Christine and Edward M Marcotte (2012). “Insights into the regulation of protein abundance from proteomic and transcriptomic analyses.” eng. In: *Nature reviews. Genetics* 13.4, pp. 227–232.

Vuillermot, Stéphanie et al. (2010). “A longitudinal examination of the neurodevelopmental impact of prenatal immune activation in mice reveals primary defects in dopaminergic development relevant to schizophrenia”. eng. In: *J Neurosci* 30.4, pp. 1270–1287.

Walsh, Tom et al. (2008). “Rare Structural Variants Disrupt Multiple Genes in Neurodevelopmental Pathways in Schizophrenia”. In: *Science* 320.5875, pp. 539–543.

Wang, K et al. (2010). "MapSplice: accurate mapping of RNA-seq reads for splice junction discovery". eng. In: *Nucleic Acids Res* 38.18, e178.

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 (cit. on p. 54).

Wassef, A, J Baker, and L D Kochan (2003). "GABA and schizophrenia: a review of basic science and clinical studies". eng. In: *J Clin Psychopharmacol* 23.6, pp. 601–640.

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. 54).

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 (cit. on p. 57).

Wockner, L F et al. (2014). "Genome-wide DNA methylation analysis of human brain tissue from schizophrenia patients". In: *Transl Psychiatry* 4, e339 (cit. on p. 154).

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

Yang, Jian, Beben Benyamin, et al. (2010). "Common SNPs explain a large proportion of the heritability for human height." eng. In: *Nature genetics* 42.7, pp. 565–569.

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

Yang, Jian, Naomi R. Wray, and Peter M. Visscher (2010). "Comparing apples and oranges: Equating the power of case-control and quantitative trait association studies". In: *Genetic Epidemiology* 34.3, pp. 254–257 (cit. on p. 48).

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 (cit. on p. 57).

Yeganeh-Doost, Peyman et al. (2011). "The role of the cerebellum in schizophrenia: from cognition to molecular pathways". In: *Clinics* 66.Suppl 1, pp. 71–77.

Yue, Feng et al. (2014). "A comparative encyclopedia of DNA elements in the mouse genome." eng. In: *Nature* 515.7527, pp. 355–364.

Zhao, B and J P Schwartz (1998). "Involvement of cytokines in normal CNS development and neurological diseases: recent progress and perspectives". eng. In: *J Neurosci Res* 52.1, pp. 7–16.

Zhao, Shanrong et al. (2014). "Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells". In: *PLoS ONE* 9.1. Ed. by Shu-Dong Zhang, e78644.

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

Zimmer, Geraldine et al. (2010). "Chondroitin sulfate acts in concert with semaphorin 3A to guide tangential migration of cortical interneurons in the ventral telencephalon." eng. In: *Cerebral cortex (New York, N.Y. : 1991)* 20.10, pp. 2411–2422.