The Genetics of Schizophrenia - The Contribution of Common Variants and Genes Implicated in a Maternal Immune Activation Model

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
May 9, 2016

Abstract

Schizophrenia is a disabling disorder affecting approximately 1% of the population worldwide. Twins studies and family studies estimated that genetic factors contributes to 64%-80% of the liability of schizophrenia. With the development of Genome-Wide Association Study (GWAS), a total of 108 common genetic loci associated with schizophrenia has been identified through the meta-analysis conducted by the Schizophrenia Working Group of Psychiatric Genomics Consortium (PGC). Using the summary statistic from the PGC schizophrenia GWAS, **Bulik-Sullivan2015** estimated that the common variant contributes to around 55.5% of the liability of schizophrenia using LD SCore regression (LDSC). However, studies suggest that rare mutations, structural variance, copy number variation (CNV) and also genetic-environment interaction are all contributing to the risk of schizophrenia. This suggest that the estimate from **Bulik-Sullivan2015** might be too high. An independent estimation of the Single Nucleotide Polymorphism (SNP)-heritability of schizophrenia might provide insight as to whether if estimates from **Bulik-Sullivan2015** are inflated.

In this thesis, we developed SNP HeRitability Estimation Kit (SHREK), an alternative algorithm for the estimation of SNP-heritability. Our simulation results suggest that, SHREK provided a more robust estimate for oligogenic traits and for binary traits when no confounding variables was present when compared to LDSC. Most importantly, using the summary statistics from the schizophrenia GWAS, the SNP-heritability of schizophrenia is estimated to be 0.185 (SD=0.00450)

by SHREK and 0.198 (SD=0.0057) by LDSC, suggesting that the estimates from **Bulik-Sullivan2015** are inflated. Thus, the common variants contributes to less than 20% of the liability of schizophrenia and it is likely for other genetic variants such as rare mutations and epigenetic factors to contribute to the heritability of schizophrenia.

In addition, previous studies have reported the interaction between genetic variation and prenatal infection in the etiology of schizophrenia. There are evidences that the effect of prenatal infection is mediated by maternal immune response, thus it is likely for the perturbation induced by maternal immune activation (MIA) to interact with genetic variations in the development of schizophrenia.

We therefore performed a RNA-sequencing study to investigate whether there are any genetic overlaps between differential genes induced by MIA and genetic variations detected by schizophrenia GWAS using the polyriboinosinic-polyribocytidilic acid (PolyI:C) mouse model. We found that the functional gene sets associated with schizophrenia are also enriched in MIA. In addition, when investigating the treatment effect of n-3 polyunsaturated fatty acid (PUFA) rich diet in MIA, we found that the gene expression of *Sgk1*, a gene that regulates the glutamatergic system, is affected by the n-3 PUFA rich diet in the PolyI:C exposed mice. *Sgk1* is therefore a potential mediator of treatment effect of n-3 PUFA rich diet in the MIA model. In conclusion, our results suggested that genes related to neural function and calcium ion signaling, as well as glutamate-related genes such as *Sgk1*, are the potential targets for future schizophrenia research.

(488 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

CI	confidence interval.
cM	centiMorgan.
CNV	copy number variation.
DSM	Diagnostic and Statistical Manual of Mental Disorders.
DZ	dizygotic.
GC	Genomic Control.
GCTA	Genome-wide Complex Trait Analysis.
GRM	Genetic Relationship Matrix.
GWAS	Genome-Wide Association Study.
IQ	Intelligence Quotient.
kb	kilobase.
LD	Linkage Disequilibrium.
LDSC	LD SCore regression.
MAF	Minor Allele Frequency.
MAPK	mitogen-activated protein kinase.
MHC	major histocompatibility complex.
MIA	maternal immune activation.
MLM	mixed linear model.
MZ	monozygotic.
PGC	Psychiatric Genomics Consortium.
PolyI:C	polyriboinosinic-polyribocytidilic acid.
PUFA	polyunsaturated fatty acid.
REML	restricted maximum likelihood.
SE	standard error.
SHREK	SNP HeRitability Estimation Kit.
SNP	Single Nucleotide Polymorphism.
WHO	World Health Organization.
YLD	years lost due to disability.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Schizophrenia

Schizophreniais a devastating psychiatric disorder affecting approximately 0.3–0.7% of the population worldwide (**dsm2013diagnostic**). According to the Diagnostic and Statistical Manual of Mental Disorders (DSM)-V, one of the standard diagnostic tools in psychiatry, a diagnosis of schizophrenia (F20.9) can only be reached if the patient has suffered from 2 or more of the following symptoms for a significant portion of time during a 1-month period: 1) delusion; 2) hallucination; 3) disorganized speech; 4) grossly disorganized or catatonic behaviour; and 5) negative symptoms such as diminished emotional expression, where one of the symptom must be either (1), (2) or (3), which are known as positive symptoms. Signs of disturbance need to persist for at least 6-month before the patient can be diagnosed with schizophrenia. Current medical treatment of schizophrenia, based on dopamine D2 receptor blockage, is effective only for the amelioration of positive symptoms in approximately 2/3 of patients.

Because of its disabling symptoms and the lack of entirely effective treatments, schizophrenia imposes a serious and long lasting health, social and financial burden to patients and their families (**knapp2004global**), where an increased tendency to commit suicide is often observe in schizophrenia patients (**saha2007systematic**). Based on the World Health Organization (WHO) report, schizophrenia was one of

the top 20 leading cause of years lost due to disability (YLD) in 2012, ranking 16 among all possible causes (table 1.1). In view of its severity, schizophrenia has

Table 1.1: Top 20 leading causes of YLD calculated by WHO in year 2012. Schizophreniawas considered as one of the top 20 leading causes of YLD (**Geneva2013**).

Rank	Cause	YLD (000s)	% YLD	YLD per 100k population
0	All Causes	740,545	100	10466
1	Unipolar depressive disorders	76,419	10.3	1080
2	Back and neck pain	53,855	7.3	761
3	Iron-deficiency anaemia	43,615	5.9	616
4	Chronic obstructive pulmonary disease	30,749	4.2	435
5	Alcohol use disorders	27,905	3.8	394
6	Anxiety disorders	27,549	3.7	389
7	Diabetes mellitus	22,492	3	318
8	Other hearing loss	22,076	3	312
9	Falls	20,409	2.8	288
10	Migraine	18,538	2.5	262
11	Osteoarthritis	18,096	2.4	256
12	Skin diseases	15,744	2.1	223
13	Asthma	14,134	1.9	200
14	Road injury	13,902	1.9	196
15	Refractive errors	13,498	1.8	191
16	Schizophrenia	13,408	1.8	189
17	Bipolar disorder	13,271	1.8	188
18	Drug use disorders	10,620	1.4	150
19	Endocrine, blood, immune disorders	10,495	1.4	148
20	Gynecological diseases	10,227	1.4	145

drawn much attention from the research community to delineate disease etiology and mechanisms, and to identify risk factors associated with schizophrenia.

1.2 Understanding Disease Mechanism

The information of relative contribution of genetic and environmental variation to schizophrenia is vital for efficient study designs. Thus it is essential to estimate the *heritability* of schizophrenia, which is a measurement of the relative contribution of genetic and environmental influence to individual differences in the liability.

Heritability is divided into the broad sense heritability and the narrow sense heritability. Broad sense heritability is defined as the *proportion* of total variance of a trait in a population explained by the *total* variation of genetic factors in the population, whereas the narrow sense heritability only takes into account of the variation of *additive* genetic factors in the population instead of the total variation of genetic factors.

1.2.1 Broad Sense Heritability

For any phenotype, one can partition it into a combination of genetic and environmental components (**falconer1996introduction**)

$$\text{Phenotype (P)} = \text{Genotype (G)} + \text{Environment (E)}$$

In the absence of gene-environmental correlation or interaction, the variance of the observed phenotype (σ_P^2) can be expressed as the sum of the variance of genotype (σ_G^2) and variance of environment (σ_E^2)

$$\sigma_P^2 = \sigma_G^2 + \sigma_E^2$$

The ratio between the variance of the observed phenotype and the variance of the genetic effects is then defined as the broad sense heritability (H^2):

$$H^2 = \frac{\sigma_G^2}{\sigma_P^2}$$

One key feature of heritability is that it is a *ratio* of *population* measurements at a specific time point. As a result, the heritability of a trait can differ in different strata of the same population (because of differences in the environment), and in different populations (because of differences in both genes and environment). A classic example is Intelligence Quotient (IQ), which increases in heritability with increasing age (**Bouchard2013**). It was hypothesized that the shared environment has a relatively larger effect on individuals when they were young, and gradually diminishes when they grow older and become more independent. The reduction in shared environmental influences results in an *increased portion* of variance in IQ explained by genetic differences (**Bouchard2013**).

However, as there are multiple forms of genetic effects, the definition of heritability can be more complicated. This leads to the concept of the narrow sense heritability.

1.2.2 Narrow Sense Heritability

There are multiple forms of genetic effects other than the additive effect. The effects of the genes can differ depending on the other gene at the same locus (dominance) or genes at different loci (epistasis). As a result, the total genetic variance can be partitioned into variance due to additive genetic effects (σ^2_A), variance due to dominant genetic effects (σ^2_D), and variance due to other epistatic genetic effects (σ^2_I):

$$\sigma^2_G = \sigma^2_A + \sigma^2_D + \sigma^2_I$$

In reality, individuals can only transmit one copy of each gene at a single genetic locus to their offspring. As a result, relatives other than full siblings and monozygotic twins can only share a maximum of one gene for each locus. Dominance and epistatic genetic effects are interactive effects which involve more than one genes and

thus are unlikely to contribute substantially to the resemblance between relatives other than monozygotic twins and full siblings (**Visscher2008**). On the contrary, the additive genetic effects are more transmittable from parents to offspring, and are therefore an important factor to consider in the prediction of parent-offspring resemblance. The narrow sense heritability (h^2) is therefore defined as the proportion of total variance of a trait in a population explained by the variance of the additive genetic effects in the population:

$$h^2 = \frac{\sigma_A^2}{\sigma_P^2} \tag{1.1}$$

To obtain the additive genetic effect, the genetic effect of a parent is first defined as $G_p = A + D$ where A is the additive genetic effect and D is the non-additive genetic effects. As only half of the additive effect are transmitted to their offspring, the child will have a genetic effect of $G_c = \frac{1}{2}A + \frac{1}{2}A' + D'$ where A' is the additive genetic effect obtained from another parent at random and D' is the non-additive genetic effect in the offspring. The parent offspring covariance $\mathrm{Cov}(O, P)$ can therefore be defined as

$$\begin{aligned} &\mathrm{Cov}(O, P) \qquad (1.2)\\ &= \mathrm{Cov}(\frac{1}{2}A + \frac{1}{2}A' + D', A + D)\\ &= \frac{1}{2}\mathrm{Cov}(A, A) + \frac{1}{2}\mathrm{Cov}(A, D) + \frac{1}{2}\mathrm{Cov}(A', A) + \frac{1}{2}\mathrm{Cov}(A', D) + \mathrm{Cov}(D', A) + \mathrm{Cov}(D', D)\\ &= \frac{1}{2}\sigma_A^2 + \frac{1}{2}\mathrm{Cov}(A, D) + \frac{1}{2}\mathrm{Cov}(A', A) + \frac{1}{2}\mathrm{Cov}(A', D) + \mathrm{Cov}(D', A) + \mathrm{Cov}(D', D) \qquad (1.3) \end{aligned}$$

Under the assumption of random mating, A' should be independent from A and D and D' should be independent from A and D, with the covariance between the additive genetics and non-additive genetics being zero (**falconer1996introduction**). Therefore

$$\mathrm{Cov}(O, P) = \frac{1}{2}\sigma_A^2 \tag{1.4}$$

Now in a simple linear regression of $y = X\beta + \epsilon$, the regression slope can be calculated as

$$\beta_{Xy} = \frac{\mathrm{Cov}(X, y)}{\mathrm{Var}(X)} \tag{1.5}$$

Therefore, by performing a regression of offspring on one of the parent, the regression slope is:

$$\begin{aligned} \beta_{OP} &= \frac{\mathrm{Cov}(O, P)}{\mathrm{Var}(P)} \\ &= \frac{1}{2}\frac{\sigma_A^2}{\sigma_P^2} \end{aligned} \tag{1.6}$$

which is half of the narrow sense heritability (h^2). eq. (1.6) can be further generalized to all possible relativeness (r):

$$h^2 = \frac{\beta_{XY}}{r} \tag{1.7}$$

where r is the proportion of genome shared by descent by relatives X and Y.

A key assumption in this calculation is that only additive genetic factors are shared among relatives. However, this is unlikely to be entirely true as relatives tend to be in the same cultural group and might have similar socio-economic status. These might all contribute to the variance of the trait, thus lead to bias in eq. (1.7). This will be discussed in more details in the later sections.

Nonetheless, eq. (1.7) provides a simple example for the calculation of the narrow sense heritability. However, for discontinuous traits (e.g. disease status), the calculation becomes more complicated because the variance of the phenotype is dependent on the population prevalence. As eq. (1.7) does not account for the trait prevalence, it cannot be directly applied to discontinuous traits. To perform heritability estimation on discontinuous traits, the concept of liability threshold model proposed by **Falconer1965** is necessary with the calculation.

1.2.3 Liability Threshold

According to the central limit theorem, if a phenotype is determined by a multitude of genetics and environmental factors with relatively small effect, then its distribution will likely be normal (**Visscher2008**). Indeed, many human traits (e.g. height and weight) do have a normal distribution in the population, or can be mathematically transformed to have a normal distribution. However, it is not possible for diseases like schizophrenia, where only a dichotomous disease status ("affected" and "normal") is observed.

To address this issue, **Falconer1965** proposed the liability-threshold model, which suggests that these discontinuous traits are determined by an underlying "liability" that follows a continuous distribution. Under the liability-threshold model, a discontinuous traits is assumed to be affected by a combination of numerous genetic and environmental factors, each with small effect. The dichotomous phenotype of an individual is determined by whether the combined effect of these factors ("liability") is above a particular threshold ("liability threshold") (fig. 1.1), i.e. only when an individual has a liability above the liability threshold will he/she be affected. One can then estimate the heritability of the discontinuous traits by comparing the mean liability of the general population to that of the relatives of the affected individuals. For example, considering a single threshold model of a dichotomous trait, where

T_G = Liability threshold of the general population

T_R = Liability threshold of relatives of the index case

q_G = Prevalence in the general population

q_R = Prevalence in relatives of the index case

L_a = Mean Liability of the index case

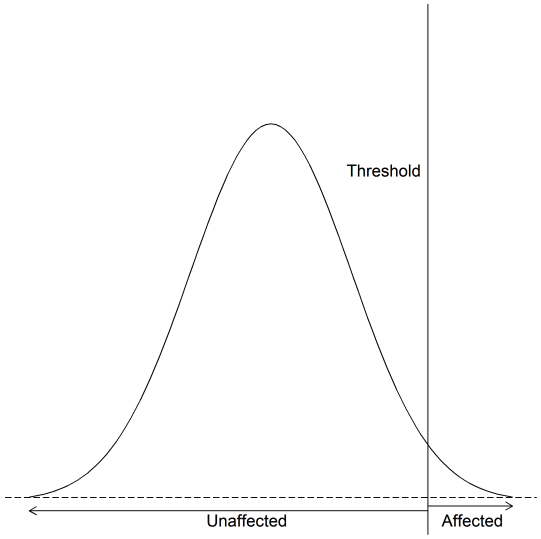


Figure 1.1: The liability threshold model. Only when an individual has a liability above the liability threshold will he/she be affected.

by assuming both the liability distribution of the general population and the relative of the index case follow the standard normal distribution, the two distributions can be aligned with respect to T_G and T_R. The mean liability of the index case L_a can then be calculated as $L_a = \frac{z_G}{q_G}$ where z_G is the value of the density function of the standard normal distribution at the liability threshold T_G. The regression of relatives' liability on the liability of the index case can be expressed as

$$\beta = \frac{T_G - T_R}{L_a} \tag{1.8}$$

Thus, by applying eq. (1.8) to eq. (1.7),

$$h^2 = \frac{T_G - T_R}{rL_a} \tag{1.9}$$

1.2.4 Adoption Study

One key limitation of eq. (1.7) is its inability to discriminate the genetic factors from the shared environmental factors. Relatives can share not only the additive genetic effect of alleles, but also some of the environmental factors such as diet and

socio-economic status.

Adoption studies are one of the classic tools for discriminating the effect of additive genetic factors from the effects of environmental factors. In adoption studies, the child is separated from their family soon after birth, thus minimized the effect of shared environmental factors. Any resemblance between the parent and offspring should be driven primarily by the shared genetic factors.

In the classical adoption study carried out by **HESTON1966** in 1966, 47 individuals who were born to schizophrenic mothers during the period from 1915 to 1947 were collected. The child were separated from their mother within three days of birth and sent to a foster family. 50 matched controls were also recruited in this study. An increased risk of schizophrenia was observed in individuals born to schizophrenic mothers when compared to the controls, even-though they were separated from their mother early on. This provide strong support for schizophrenia as a genetic disorder.

1.2.5 Twin Studies

Despite the usefulness of adoption studies, collection of adoption data are extremely difficult. Moreover, adoption studies cannot control the prenatal environment such as alcohol abuse and malnutrition during pregnancy, which remain confounded with genetic effects. For example, a high comorbidity of alcohol use disorders and schizophrenia was observed (**boyd1984**; **drake1990**). High consumption level of alcohol during pregnancy may bring about long-term alterations in nervous system functioning of the offspring (**garrett2014brain**). If the prenatal environmental effect increase the risk of schizophrenia, the effects of these environmental effect might be mis-interpreted as genetic effect.

Alternatively, twin studies, which utilize the genetic relationship between

monozygotic (MZ) and dizygotic (DZ) twins, can be performed to estimate the heritability of the target trait.

By definition, the common environmental factors (C) is the same for the twins whereas the non-shared environmental factors (E) is unique to each individual. For MZ twins, both the additive (A) and non-additive (D) genetic factors are shared between the siblings. On the other hand, only $\frac{1}{2}$ of the additive genetic factors and $\frac{1}{4}$ of the non-additive genetic factors are shared among the DZ twins (**rijsdijk2002analytic**).

In view of this, **falconer1996introduction** derived the heritability as

$$h^2 = 2(\rho_{MZ} - \rho_{DZ}) \tag{1.10}$$

where ρ_{MZ} and ρ_{DZ} are the phenotype correlation between the MZ twins and DZ twins respectively. However, if assortative mating occurs in the population, the additive genetic factors shared between the DZ twins might be higher than $\frac{1}{2}$, thus leading to an underestimation of the heritability. Nonetheless, twin studies can be served as the first step in the study of genetic architecture and provide vital understanding of the trait.

By combining Falconer's formula and the concept of liability threshold model, **Gottesman1967** estimated that the heritability of schizophrenia to be $>$ 60% based on previously collected twin data. This provides strong evidence that the genetic variation contributes more to the variance of schizophrenia.

The result was further supported by one of the landmark meta-analysis study conducted by **sullivan2003schizophrenia** Based on data obtained from 12 published schizophrenia twin studies, **sullivan2003schizophrenia** found much contribution from genetics on the liability of schizophrenia (81%, confidence interval (CI)=$73\% - 90\%$). Furthermore, in the large scale population based studies performed by **Lichtenstein2009** the genetic contribution to schizophrenia was found

to be 64%. Together, these results provide strong support for schizophrenia as a genetic disorder.

1.3 Schizophrenia Genetics

Although schizophrenia is highly heritable, little is known about the disease mechanism of schizophrenia and the genetic complexity of the disorder. It was observed that the lifetime morbid risk of MZ twins were only 48% (fig. 1.2), suggesting that it is unlikely for schizophrenia to follow the Mendelian framework (**Gottesman1967**; **Gottesman1982**; **gottesman1991schizophrenia**).

In view of this, **Gottesman1967** proposed that schizophrenia might follow a polygenic model, where the disease phenotype were determined by the additive effects from multiple genes.

By comparing the observed lifetime morbid risk and the expected risk from different models, **Risch1990** proposed that the causal variants of schizophrenia are more likely to have a risk less than 2 with no loci with risk larger than 3, suggesting a relatively small effect size. Large sample size are therefore required to detect these susceptibility loci through linkage studies (**Risch1990**).

As genetic data from large, multi-generational pedigrees with both affected and unaffected individuals are difficult to recruit, it is challenging for linkage studies to collect adequate samples. Thus early linkage studies of schizophrenia results in inconsistent findings (**Harrison2005**). Other methods are therefore required to identify the susceptibility loci of schizophrenia.

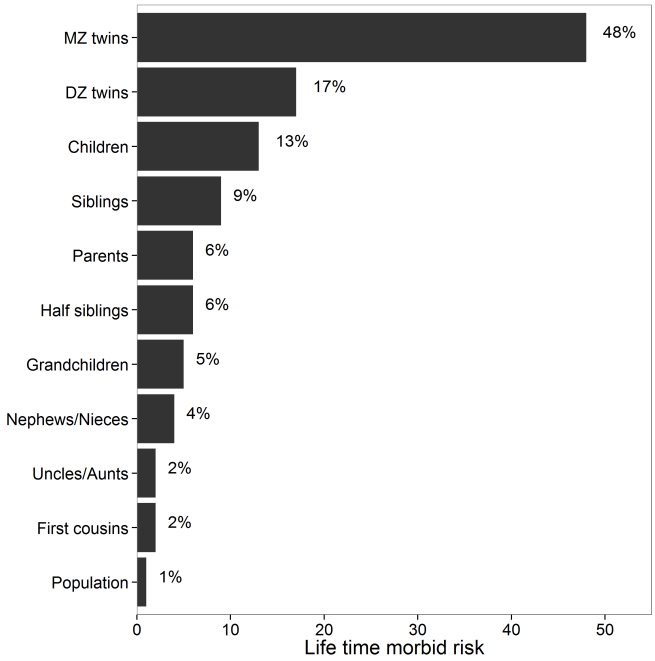


Figure 1.2: Lifetime morbid risks of schizophrenia in various classes of relatives of a proband. It was noted that the morbid risk of MZ twins was only 48%, much lower than one would expect if schizophrenia follows a Mendelian pattern. Reproduced with permission from journal (**Riley2006**).

1.3.1 The Human Genome Project and HapMap Project

In 1990, the Human genome project was initiated, aiming at constructing the first physical map of the human genome at per nucleotide resolution (**Lander2001**). The completion of the human genome project has opened up a new era of genetic research, allowing researchers to identify Single Nucleotide Polymorphisms (SNPs), which is one of the major source of genetic variation in the human genome.

Soon after the completion of the human genome project, the HapMap Project was initiated (**Consortium2005**), aiming to provide a genome-wide database of common human sequence variation such as SNPs with Minor Allele Frequency (MAF) ≥ 0.05.

More importantly, the HapMap Project provided a detailed Linkage Disequilibrium (LD) map of the human genome. LD is the non-random correlation

of genotypes between 2 genetic loci. SNPs in high LD are usually observed together in the human genome. When a large amount of SNPs are in high LD, a LD block is formed. By performing association testing on SNPs representing majority of the information within the LD block ("tagging"), genome-wide association can be performed. This is the fundamental concept of Genome-Wide Association Study (GWAS), which is now extensively used in genetic researches.

1.3.2 Genome Wide Association Study

In GWAS, genome-wide genotyping array are commonly used to systematically detect genetic variants such as SNP and copy number variation (CNV) in genome-wide scale. For quantitative traits, the association between the trait and frequency of the variants are calculated using methods such as linear regression. On the other hand, for dichotomous traits such as schizophrenia, the frequency of the variants are compared between the case and control samples using chi-square test or logistic regression.

However, when a large number of SNPs were tested, the frequency of type I error increases (**Peters2010**). The simplest method for the correction of GWAS is to use the genome wide threshold (p-value $\leq 5\times10^{-8}$), where only SNPs with p-value less then the genome wide threshold are considered to be significant in GWAS. Another possible method to decide the significant threshold is to consider the "effective number" of tests (**Li2011**), which reduced the genome-wide threshold according to the LD structure. Most importantly, the same genome-wide significance threshold should be used even for low density SNP arrays in which the SNPs are *randomly* selected. This is because the interpretation of the result of a random SNP should not depend on whether another SNP is tested or not (**gelman2016statistical**; **Sham2014**). Thus, multiple testing correction is vital in the analysis of SNP association.

Finally, when designing a GWAS, the magnitude of effect, sample size, and required level of statistical significance (the false-positive, or type I, error rate) are all important factors determining the detection power of the GWAS (**Purcell2003**). Similar to linkage studies, a larger sample size are required to identify susceptible loci with a smaller effect.

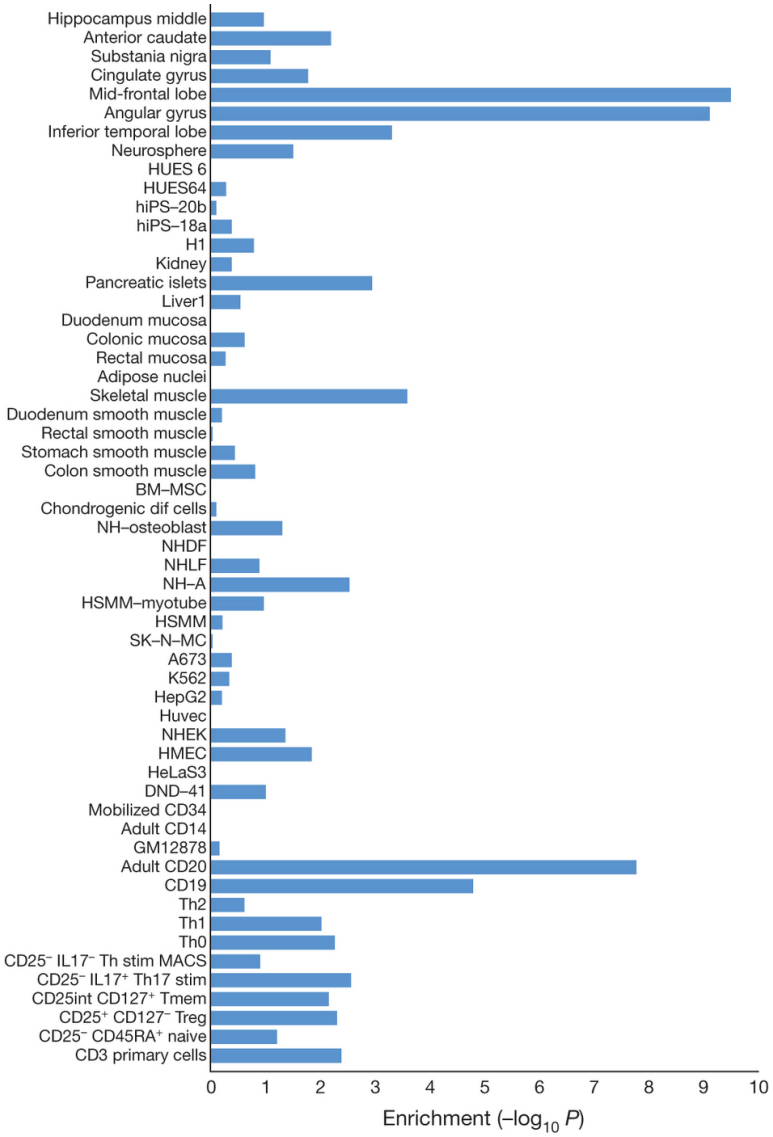
1.3.2.1 The Success of Psychiatric Genomic Consortium

Due to the relatively small sample size, early GWAS in schizophrenia were largely disappointing, where no robust genetic markers associated with schizophrenia were identified.

To overcome the problem of sample size, large consortium were formed such that genetic data from different research groups can be combined and analyzed. Finally, in 2014, the Schizophrenia Working group of the Psychiatric Genomics Consortium (PGC) has conducted a multi-stage schizophreniaGenome-Wide Meta Analysis of up to 36,989 schizophrenia samples and 113,075 controls (**Ripke2014**). A total of 128 linkage-disequilibrium-independent SNPs were found to exceeded the genome-wide significance (p-value $\leq 5 \times 10^{-8}$), correspond to 108 independent genetic loci. 75% of these loci contain protein coding genes and a further 8% of these loci were within 20 kilobase (kb) of a gene. It was found that genes involved in glutamatergic neurotransmission (e.g. *GRM3*, *GRIN2A* and *GRIA1*), synaptic plasticity and genes encoding the voltage-gated calcium channel subunits (e.g. *CACNA1C*, *CACNB2* and *CACNA1I*) were among the genes associated within these loci. Moreover, associations were significantly enriched at enhancers active in brain and in tissues with important immune functions (fig. 1.3) (**Ripke2014**).

The enrichment of immune related enhancers remains significant even after the removal of major histocompatibility complex (MHC) region from the analysis, suggesting that the significance association of the immune system with schizophrenia

Figure 1.3: Enrichment of enhancers of SNPs associated with schizophrenia. It was observed that the largest enrichment were in cell lines related to the brain and in tissues with important immune functions. Graphs reproduced with permission from the journal (**Ripke2014**).



is not driven only by the MHC region. Considering the role of immune system in neural development (**Zhao1998**; **Deverman2009**), perturbation in the immune system is likely to disrupt the brain development. Therefore, the immune system might have an important role in the etiology of schizophrenia.

Given the success of PGC schizophrenia GWAS, it is interesting to investigate the relative contribution of common variants, which are captured in the GWAS, to the genetic predisposition of schizophrenia. This will provide vital information as to whether other genetic variations such as rare mutations and methylation are also important.

1.3.3 Contribution of Common SNPs

By imposing a stringent genome-wide significant threshold to the results of GWAS, the Type I error can be reduced. However, real associations with a small effect size might be filtered out through multiple testing correction. Therefore, to estimate the true contribution of common SNPs to a trait (SNP-heritability), it is necessary to consider all the SNPs in the estimation.

1.3.3.1 Genome-wide Complex Trait Analysis

Currently, the most popular algorithm for the estimation of SNP-heritability is Genome-wide Complex Trait Analysis (GCTA), which utilize information from the Genetic Relationship Matrix (GRM) (**Yang2011**). The GRM represents the genetic relationship between all individuals within the GWAS. The genetic relationship between individual j and k can be estimated as

$$A_{jk} = \frac{1}{N} \sum_{i=1}^{N} \frac{(x_{ij} - 2p_i)(x_{ik} - 2p_i)}{2p_i(1 - p_i)} \tag{1.11}$$

where x_{ij} is the number of copies of the reference allele for the i^{th} SNP of the j^{th} individual and p_i is the frequency of the reference allele. Because genotypes are usually code as 0, 1 or 2 (homozygous reference, heterozygous and homozygous alternative respectively), one can model the distribution of genotype using the binomial distribution. Therefore, the expected mean and variance of genotype i is $2p_i$ and $2p_i(1-p_i)$ respectively, and the GRM can be represented as $A_{jk} = \frac{1}{N}\sum_{i=1}^{N} z_{ij}z_{ik}$, where z_{ij} is the standardized genotype for the i^{th} SNP of the j^{th} individual.

Using the information from the GRM, **Yang2011** fitted the effects of all the SNPs as random effects in a mixed linear model (MLM)

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{g} + \epsilon \tag{1.12}$$

$$\text{Var}(\boldsymbol{y}) = \boldsymbol{A}\sigma_g^2 + \boldsymbol{I}\sigma_\epsilon^2 \tag{1.13}$$

where $\boldsymbol{y}$ is an $n \times 1$ vector of phenotypes with n samples, $\boldsymbol{\beta}$ is a vector of fixed effects such as sex and age, $\boldsymbol{g}$ is an $n \times 1$ vector of the total genetic effects of the individuals, σ_g^2 is the variance explained by all the SNPs and finally, σ_ϵ^2 is the variance explained by residual effects.

By fitting the effects of *all* SNPs as random effects in a MLM, a single parameter can be estimated, i.e. the variance explained by all SNPs or SNP-heritability. Based on eq. (1.13), **Yang2011** implemented the restricted maximum likelihood (REML) using the average information algorithm to provide an unbiased estimates of σ_g^2 and σ_ϵ^2. The SNP-heritability of the trait can then be defined as $\frac{\sigma_g^2}{\sigma_g^2+\sigma_e^2}$.

Yang2010a were able to estimate the variance in height explained by GWAS SNPs to be around 45%, much larger than previously reported 5%. However, the estimate was still less than 80%, which is the expected heritability of height. **Yang2010a** hypothesized that one possible source of "missing heritability" might be due to incomplete LD of the GWAS chip. Taken into consideration of incomplete

LD, **Yang2010a** estimated that the proportion of variance explained by causal variants to be as high as 0.84 with standard error (SE) of 0.16, within the range of the expected heritability of height, providing support to their hypothesis.

By far, the biggest limitation of GCTA is the requirement of individual genotype data to calculate the GRM. For studies where only summary statistics are available, e.g. the meta-analysis of schizophrenia, GCTA analysis cannot be performed. Therefore, other methods are required.

1.3.3.2 LD SCore regression

In large scale GWAS studies, a general inflation of summary statistics can sometimes be observed. The inflation was usually considered to be contributed by the presence of confounding factors such as population stratification, under the assumption that most of the SNPs were not associated with the disease. Therefore, Genomic Control (GC) inflation factor were usually used to control for the inflation in GWAS results (**Zheng2006**).

However, for complex diseases such as schizophrenia, there might be a large number of causal SNPs, therefore violating the underlying assumption of GC. Through careful simulation, **Yang2011b** demonstrated that in the absence of population stratification and other form of technical artifacts, the presence of polygenic inheritance can inflate the summary statistic. It was observed that the magnitude of inflation was determined by the *heritability*, the LD structure, sample size and the number of causal SNPs of the trait.

Bulik-Sullivan2015 noted that the variants in LD with a causal variant show an elevation in summary statistic proportional to their LD with the causal variant. Therefore, the more SNPs that are in LD with the index SNP, the higher the likelihood for the index SNP to tag the causal variant. However, inflation in

summary statistic resulting form cryptic relatedness and population stratification will not correlate with LD. Therefore, **Bulik-Sullivan2015** developed the LD score, where the LD score of a SNP j is defined as the sum of r^2 of k neighboring SNPs within a 1 centiMorgan (cM) window:

$$l_j = \sum_k r_{jk}^2 \tag{1.14}$$

The expected χ^2s association of SNP_j with the trait can then be defined as a function of the LD score (l_j), the number of samples (N), the number of SNPs in the analysis (M) and most importantly, the SNP heritability (h^2):

$$\text{E}[\chi_j^2|l_j] = \frac{Nh^2}{M} l_j + 1 \tag{1.15}$$

Alternatively, if confounding factors are present in the study (e.g. population stratification), eq. (1.15) can be defined as

$$\text{E}[\chi_j^2|l_j] = \frac{Nh^2}{M} l_j + Na + 1 \tag{1.16}$$

where a is the contribution of the confounding bias.

By considering eq. (1.16) as a regression model, **Bulik-Sullivan2015** observed that the contribution of common variants (the SNP heritability h^2) is the slope of the regression, whereas the intercept minus one will represent the mean contribution of the confounding bias, such as population stratification. Using eq. (1.16), **Bulik-Sullivan2015** implemented the LD SCore regression (LDSC) to delineate the contribution from confounding factors and common genetic variants.

To assess the performance of LDSC, **Bulik-Sullivan2015** performed a number of controlled simulation. First, they simulated polygenic traits without any confounding factors. The result shows that the average intercept estimated by LDSC was closed to one, correctly suggest the absence of confounding factors. Moreover, the heritability estimates provided by LDSC were unbiased for all simulated

conditions. Only when the number of causal variants was small did the standard error of the LDSC estimates become very large. Similarly, when non-genetic trait was simulated with only confounding factors such as population stratification, the intercepts estimated by LDSC were approximately equal to the GC inflation factor, with only a small positive bias in the regression slope. Most importantly, when polygenic traits with confounding factors were simulated, the intercepts estimated by LDSC were approximately equal to the mean χ^2 statistics among the null SNPs. This provide strong evidence that the LD score regression can partition the inflation in test statistics even in the presence of both bias and polygenicity.

Following the success of their simulation, **Bulik-Sullivan2015** utilized LDSC to estimate the SNP heritability of schizophrenia using the summary statistics from the PGC schizophrenia GWAS (**Ripke2014**). The estimated SNP heritability of schizophrenia is around 0.555 with standard error of 0.008 after adjusting for ascertainment bias. Comparing the SNP heritability estimated with the heritability estimated from the population study (64% (**Lichtenstein2009**)) and the twin studies (81% (**sullivan2003schizophrenia**)), it seems as if the common variants have accounted for most if not all of the heritability of schizophrenia.

1.3.3.3 Partitioning of Heritability

Other than the estimation of SNP heritability, LDSC also allows for the partitioning of heritability into different pathways, which can be used for functional enrichment analysis.

Traditionally, functional enrichment analysis in GWAS only consider significant SNPs. However, SNPs with small effect size that do not reach genome-wide significance threshold might be filtered out. For example, in 2013, only 13 risk loci were detected using 13,833 schizophrenia samples and 18,310 controls (**Ripke2013**). Yet when the sample size increased to 34,241 schizophrenia samples and 45,604 con-

trols in 2014, 108 risk loci were identified (**Ripke2014**). Thus, only when sample size are larger can those risk loci with a smaller effect size to be able to reach the genome-wide significance threshold. By filtering the insignificant loci, the functional enrichment analysis might loses power.

On the other hand, LDSC utilize the summary statistic of all the SNPs included in the GWAS when estimating the association of a functional category with the trait. Specifically, eq. (1.16) modified into

$$\mathrm{E}[\chi_j^2] = N \sum_C \tau_C l(j, C) + Na + 1 \tag{1.17}$$

with $\frac{h^2}{M} l_j$ substituted by $\sum_C \tau_C l(j, C)$ where $l(j, C)$ is the LD Score of SNP j with respect to category C and τC is the per-SNP heritability in category C.

By applying eq. (1.17) to the summary statistics from the PGC schizophrenia GWAS, **Finucane2015** found that brain cell types and immune related cell types were most enriched in schizophrenia. Of all the functional categories, H3K4me3 mark in the fetal brain (table 1.2) was the most enriched in schizophrenia. As H3K4me3 is mostly linked to active promoters, genes activated in fetal brain (e.g. genes related to brain development) are therefore likely to be associated with schizophrenia, supporting the idea of schizophreniaas a neuro-developmental disorder.

Cell type	cell-type group	Mark	P-value
Fetal brain**	CNS	H3K4me3	3.09×10^{-19}
Mid frontal lobe**	CNS	H3K4me3	3.63×10^{-15}
Germinal matrix**	CNS	H3K4me3	2.09×10^{-13}
Mid frontal lobe**	CNS	H3K9ac	5.37×10^{-12}
Angular gyrus**	CNS	H3K4me3	1.29×10^{-11}
Inferior temporal lobe**	CNS	H3K4me3	1.70×10^{-11}
Cingulate gyrus**	CNS	H3K9ac	5.37×10^{-11}
Fetal brain**	CNS	H3K9ac	5.75×10^{-11}
Anterior caudate**	CNS	H3K4me3	2.19×10^{-10}
Cingulate gyrus**	CNS	H3K4me3	4.57×10^{-10}
Pancreatic islets**	Adrenal/Pancreas	H3K4me3	2.24×10^{-09}
Anterior caudate**	CNS	H3K9ac	3.16×10^{-9}
Angular gyrus**	CNS	H3K9ac	4.68×10^{-9}

Mid frontal lobe**	CNS	H3K27ac	7.94×10^{-9}
Anterior caudate**	CNS	H3K4me1	1.20×10^{-8}
Inferior temporal lobe**	CNS	H3K4me1	3.72×10^{-8}
Psoas muscle**	Skeletal Muscle	H3K4me3	4.17×10^{-8}
Fetal brain**	CNS	H3K4me1	6.17×10^{-8}
Inferior temporal lobe**	CNS	H3K9ac	9.33×10^{-8}
Hippocampus middle**	CNS	H3K9ac	9.33×10^{-7}
Pancreatic islets**	Adrenal/Pancreas	H3K9ac	1.62×10^{-6}
Penis foreskin melanocyte primary**	Other	H3K4me3	2.09×10^{-6}
Angular gyrus**	CNS	H3K27ac	2.34×10^{-6}
Cingulate gyrus**	CNS	H3K4me1	2.82×10^{-6}
Hippocampus middle**	CNS	H3K4me3	2.82×10^{-6}
CD34 primary**	Immune	H3K4me3	4.68×10^{-6}
Sigmoid colon**	GI	H3K4me3	5.01×10^{-6}
Fetal adrenal**	Adrenal/Pancreas	H3K4me3	6.31×10^{-6}
Inferior temporal lobe**	CNS	H3K27ac	8.32×10^{-6}
Peripheralblood mononuclear primary**	Immune	H3K4me3	9.33×10^{-6}
Gastric**	GI	H3K4me3	1.17×10^{-5}
Substantia nigra*	CNS	H3K4me3	1.95×10^{-5}
Fetal brain*	CNS	H3K4me3	2.63×10^{-5}
Hippocampus middle*	CNS	H3K4me1	3.31×10^{-5}
Ovary*	Other	H3K4me3	6.46×10^{-5}
CD19 primary (UW)*	Immune	H3K4me3	7.08×10^{-5}
Small intestine*	GI	H3K4me3	8.51×10^{-5}
Lung*	Cardiovascular	H3K4me3	1.17×10^{-4}
Fetal stomach*	GI	H3K4me3	1.29×10^{-4}
Fetal leg muscle*	Skeletal Muscle	H3K4me3	1.51×10^{-4}
Spleen*	Immune	H3K4me3	1.70×10^{-4}
Breast fibroblast primary*	Connective/Bone	H3K4me3	2.04×10^{-4}
Right ventricle*	Cardiovascular	H3K4me3	2.14×10^{-4}
CD4+ CD25- Th primary*	Immune	H3K4me3	2.19×10^{-4}
CD4+ CD25- IL17- PMA Ionomycin stim MACS Th sprimary*	Immune	H3K4me1	2.19×10^{-4}
CD8 naive primary (UCSF-UBC)*	Immune	H3K4me3	2.24×10^{-4}
Pancreas*	Adrenal/Pancreas	H3K4me3	2.34×10^{-4}
CD4+ CD25- Th primary*	Immune	H3K4me1	2.75×10^{-4}
CD4+ CD25- CD45RA+ naive primary*	Immune	H3K4me1	2.75×10^{-4}
Colonic mucosa*	GI	H3K4me3	3.24×10^{-4}
Right atrium*	Cardiovascular	H3K4me3	3.31×10^{-4}
Fetal trunk muscle*	Skeletal Muscle	H3K4me3	3.39×10^{-4}
CD4+ CD25int CD127+ Tmem primary*	Immune	H3K4me3	3.47×10^{-4}

Substantia nigra*	CNS	H3K9ac	3.63×10^{-4}
Placenta amnion*	Other	H3K4me3	4.17×10^{-4}
Breast myoepithelial*	Other	H3K9ac	5.50×10^{-4}
CD8 naive primary (BI)*	Immune	H3K4me1	5.75×10^{-4}
Substantia nigra*	CNS	H3K4me1	6.61×10^{-4}
Cingulate gyrus*	CNS	H3K27ac	7.94×10^{-4}
CD4+ CD25- CD45RA+ naive primary*	Immune	H3K4me3	8.71×10^{-4}

Table 1.2: Enrichment of Top Cell type of Schizophrenia. * = significant at False Discovery Rate < 0.05. ** = significant at p < 0.05 after correcting for multiple hypothesis. Reproduce with permission from Journal.(**Finucane2015**)

1.3.4 Contribution of Other Genetic Variants

Although the estimated SNP heritability of schizophrenia by **Bulik-Sullivan2015** suggest common SNPs to be the major contributor to the heritability of schizophrenia, other variants, such as copy number variation (CNV) have been found to be associated with schizophrenia.

1.3.4.1 Copy Number Variation

CNV are classified as segment of DNA that is 1 kb or larger, and is present at a different copy number when compared to the reference genome, usually in the form of insertion, deletion or duplication (**Feuk2006**). Due to the length of CNV, these variants might contain the entire genes and their regulatory regions, which might contribute to significant phenotypic differences (**Feuk2006**).

Recently, **Szatkiewicz2014** conducted a GWAS for CNV association with schizophrenia using the Swedish national sample (4,719 schizophrenia samples and 5,917 controls). A number of risk CNVs such as 16p11.2 duplications, 22q11.2 deletions, 3q29 deletions and 17q12 duplications were identified. In general, CNVs associated with schizophrenia are rare (≤ 12 in 4,719 samples (**Szatkiewicz2014**)) and

have a relative large effect (e.g. odd ratio > 2 (**Szatkiewicz2014**; **Walsh2008**)).

Szatkiewicz2014 also performed the gene set enrichment analysis and they found that calcium ion channel signaling and binding partners of the fragile X mental retardation protein are enriched by CNV observed in the schizophrenic samples (**Szatkiewicz2014**).

Similar pathways were also found to be enriched with structure variants in schizophrenic samples in a separate study conducted by **Walsh2008** Pathways important for brain development, including neuregulin signaling, extracellular signal-regulated kinase/mitogen-activated protein kinase (MAPK) signaling, synaptic long-term potentiation, axonal guidance signaling, integrin signaling, and glutamate receptor signaling were all found to be significantly overrepresented with structural variants found in schizophrenic samples.

1.3.4.2 Rare Single Nucleotide Mutation

In addition to CNV, there are also evidence of an increased burden of rare variants in schizophrenia (**purcell2014polygenic**). By sequencing the exome of 2,536 schizophrenia cases and 2,543 normal controls, **purcell2014polygenic** identified a common missense allele on *CCHCR1* in the MHC region to be associated with schizophrenia. Although none of the genes showed a significant burden of rare mutation in schizophrenia cases, a significant increased burden of rare nonsense and disruptive variants was observed in gene sets such as voltage-gated calcium ion channel, genes affected by *de novo* mutations in schizophrenia (**Fromer2014**) and the postsynaptic density, all of which have been reported to be associated with schizophrenia in previous genetic studies (**Ripke2014**).

As might be expected, evidence of gene sets enriched by common variants, structural variations, CNV and rare variants all converge to the same set

of functional pathways, suggesting a common functional pathway is disrupted in schizophrenia patients. Together, the evidences support for the involvements of genetic variations other than common SNPs in the risk of schizophrenia.

1.4 Environmental Risk Factors

In the estimation of heritability of schizophrenia, only the additive genetic factors are considered. However, **zuk2012mystery** suggested that failure in accounting for gene-environment interaction might result in "phantom heritability" (**zuk2012mystery**). Specifically, presence of the gene-environment interaction can increase the total heritability of a trait.

In 2004, **Tienari2004** conducted an adoption study where they found that individuals with higher genetic risk were significantly more sensitive to "adverse" vs "healthy" rearing patterns in adoptive families than are adoptees at low genetic risk (**Tienari2004**). Moreover, using the national registers in Finland, **Clarke2009** found that the effect of prenatal infection was five times greater in those who had a family history of psychosis when compared to those who did not. These evidences provide support of the presence of gene-environment interaction in schizophrenia, which might suggest the total heritability of schizophrenia was overestimated (**zuk2012mystery**).

Additionally, the gene(G)-environment(E) interaction loads on G in twin studies, but do not contribute to heritability estimate in GWAS, therefore gene-shared environment interaction is presented, it is expected to observe a discrepancy between the heritability estimated from twin-studies and estimated from GWAS.

With the possible overestimation of heritability for schizophrenia, the environmental contribution to the disease etiology might be higher than expected. It

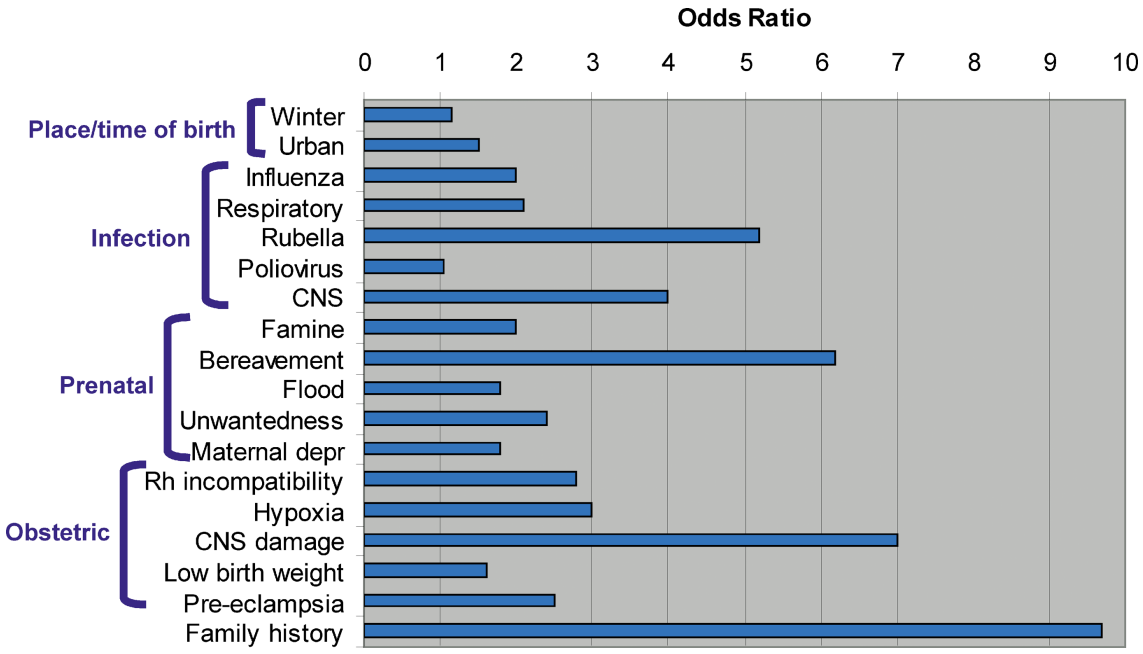


Figure 1.4: Risk factors of schizophrenia. Of all risk factors, family history of schizophrenia was the largest, where risk of schizophrenia can be more than 9 times higher than the general population for individual with a family history of schizophrenia (**Sullivan2005**)

might then be important to also study the environmental effect in schizophrenia.

Environmental factors such as prenatal infection (**Brown2010**), winter birth (**o1991season**), tobacco consumption (**Kelly1999**) and socio economic status (**mcgrath2008schizophrenia**) have been found to be associated with schizophrenia.

1.4.1 Prenatal Infection

Because evidence suggest there to be an interaction between prenatal infection and genetic variations (**Clarke2009**), it might be of particular interest to investigate the effect of prenatal infection to schizophrenia.

Initial clues of involvements of prenatal infection in the etiology of schizophrenia originates from the increased risk of schizophrenia in individuals who were fetuses during the 1957 influenza epidemic (**Mednick1958**). Moreover, it was observed

that other than influenza, infection of HSV-2 and *T.gondii* during gestation also increase the risk of schizophrenia(**Brown2010**). These evidences suggested prenatal infection might be associated with schizophrenia.

Early studies of prenatal infection in schizophrenia relied on ecological data to define the exposure status without any confirmation of maternal infection during pregnancy (**Brown2010**). Therefore, the exposure status were inaccurate and unreliable, leading to inconsistent findings (**Brown2010**). Subsequently, birth cohorts, where infection was documented using different biomarkers during pregnancies, were conducted in order to obtain a better labeling of the exposure status (**Brown2010**). From these cohorts, it was found that as long as an individual's mother was infected by any form of infectious agents (e.g. influenza, HSV-2 and *T.gondii*) during gestation, an individual's risk of schizophrenia increases (**Brown2010**). This leads to the hypothesis that maternal immune activation (MIA) (**Brown2010**) rather than a particular infectious agent, is the main risk factor. **Garbett2012a** suggested that one possible mechanism is that the maternal immune response to infection might have disrupted the brain development in the fetus, therefore predispose the fetus to schizophrenia.

1.4.2 Maternal Immune Activation Model

It is challenging to study the mechanism of MIA, because it is not possible to carry out controlled experiment on human samples due to ethical concerns. Therefore rodent models is used as an alternative. However, unlike physiological traits, psychiatric disorder such as schizophrenia are characterized by symptoms related to higher level functioning such as hallucinations, delusion, disorganized speech etc. (**dsm2013diagnostic**), which are not readily detectable in rodents. Therefore, as a compromise, rodent expressing "schizophrenia-like" behaviours such as impaired prepulse inhibition, impaired working memory and reduced social interaction are

considered as "cases" (**Meyer2007a**).

However, it is crucial to note that the behavioral abnormality is not unique to schizophrenia, but can also be observed in autistic samples. Moreover, MIA is also one of the risk factor for autism (**Brown2012**). As a results, results from these rodent studies are non-specific to schizophrenia or autism. Discussion of the similarity and difference between autism and schizophrenia is beyond the scope of the current thesis, therefore, the discussion is limited to schizophrenia.

A common rodent model in the study of MIA is to use the viral analogue polyriboinosinic-polyribocytidilic acid (PolyI:C) to induce the maternal immune response during pregnancy. In this model, offspring exposed to PolyI:C displays phenotypes mirrors those observed in schizophrenia (**Li2009c**; **Meyer2009b**; **Li2010a**), such as deficiency in prepulse inhibition (**Cadenhead2000**). Because PolyI:C only induce the MIA without infecting the fetuses, this model provide strong evidence that MIA, instead of the specific infection, contributes to the increased risk of schizophrenia.

Smith2007 were able to demonstrate that a single injection of Interleukin-6 (IL-6) to the pregnant mouse can induce schizophrenia-like behavior in the adult offspring. By eliminating the IL-6 from the maternal immune response, the behavior deficits associated with MIA were observed in the adult offspring (**Smith2007**). This indicates that IL-6 is central to the process by which MIA causes long-term behavioral changes.

Recent studies of global gene expression patterns in MIA-exposed rodent fetal brains (**Oskvig2012**; **Garbett2012a**) suggest that the post-pubertal onset of schizophrenic and other psychosis-related phenotypes might stem from attempts of the brain to counteract the environmental stress induced by MIA during its early development (**Garbett2012a**). For example, genes with neuroprotective function such as crystallins also have additional roles in neuronal differentiation and ax-

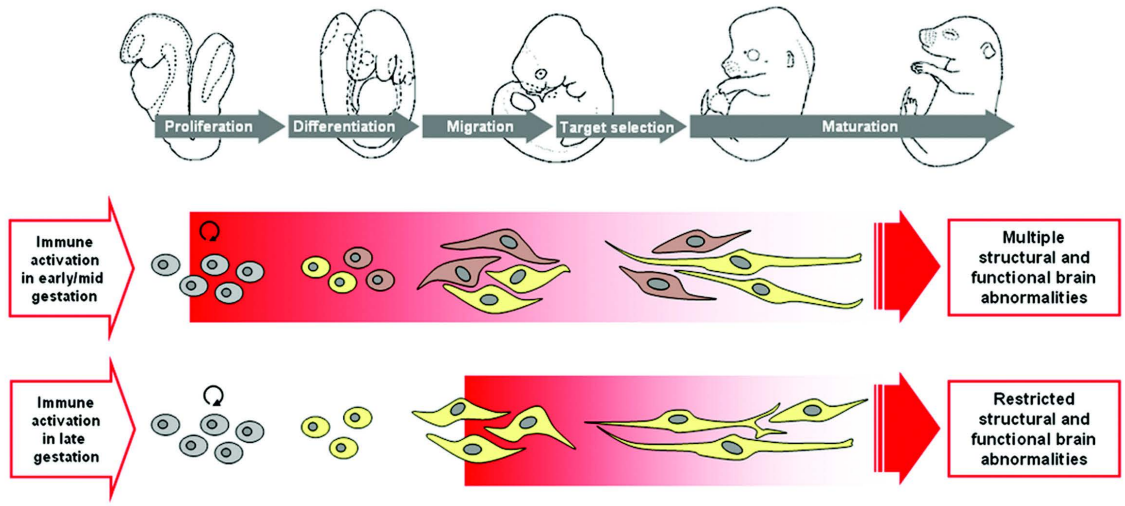


Figure 1.5: Hypothesized model of the impact of prenatal immune challenge on fetal brain development. Maternal infection in early/mid pregnancy may affect early neurodevelopmental events in the fetal brain, thereby influencing the differentiation of neural precursor cells (grey) into particular neuronal phenotype (yellow or brown). This may predispose the developing fetal nervous system to additional failures leading to multiple structural and functional brain abnormalities in later life. Figure used with permission from Journal (**Meyer2007a**)

onal growth (**Garbett2012a**). By over-expressing these genes to counteract the environmental stress, the balance between neurogenesis and differentiation in the embryonic brain maybe disrupted. Based on these observations, **Garbett2012a** propose that once the immune activation disappears, the normal brain development programme resumes with a time lag, result in permanent changes in connectivity and neurochemistry that might ultimately leads to schizophrenia-like behaviours.

On the other hand, an age dependent structural abnormalities in the mesoaccumbal and nigrostriatal dopamine systems were also found to be induced by MIA (**Vuillermot2010**). Specifically, MIA induces an early abnormality in specific dopaminergic systems such as those in the striatum and midbrain region (**Vuillermot2010**). Based on these observations, **Meyer2007a** hypothesized that inflammation in the fetal brain during early gestation not only disrupts neurodevelopmental processes such as cell proliferation and differentiation, it also predispose the developing nervous system to additional failures in subsequent cell mi-

gration, target selection, and synapse maturation (fig. 1.5) (**Meyer2007a**).

In a separate study by **Giovanoli2013** a lower dosage of PolyI:C were injected to the pregnant mice during early gestation. The authors hypothesized that a low dose of PolyI:C will leads to restricted behavioral abnormalities in adulthood, thereby avoiding possible ceiling effects of the prenatal immunological manipulation on long-term brain and behavioral functions (**Giovanoli2013**). The offspring were then left undisturbed or exposed to unpredictable stress during peripubertal development.

An increased level of dopamine in the nucleus accumbens was observed in Offspring exposed to PolyI:C disregarding whether if they were exposed to postnatal stress. Whereas serotonin (5-HT) were found to be decreased in the medial prefrontal cortex when exposed to postnatal stress regardless of prenatal exposure. Only when the offspring were exposed to both PolyI:C and postnatal stress will they have an increased dopamine levels in the hippocampus or will sensorimotor gating and psychotomimetic drug sensitivity be affected (**Giovanoli2013**). **Giovanoli2013** therefore suggest that the prenatal insult serves as a "disease primer" that increase offspring's vulnerability to subsequent insults.

One of the critical consideration in the study of MIA is the specific gestation period of vulnerability to infection-mediated disturbance (**Meyer2007a**). Early epidemiological studies have suggested that the second trimester of human pregnancy might be the vulnerability period. However, in birth cohorts such as the Prenatal Determinants of Schizophrenia, it was found that the time window with maximum risk for infection-mediated disturbance in brain development is earlier than the second trimester of human pregnancy, which can be as early as the first trimester (**Meyer2007a**). By reviewing existing MIA studies, **Meyer2007a** suggested the effect of MIA during late pregnancy is restricted to the late developmental programmes, thus have a more restricted pathological phenotype in the

grown offspring compared to MIA during early pregnancy (**Meyer2007a**). Subsequent MIA studies using the PolyI:C mouse model also support the hypothesis proposed by **Meyer2007a** where it was observed that MIA early in gestation event exert a more extensive impact on the phenotype of offspring (**Li2009c**; **Li2010a**).

Despite the more severe impact of MIA during early gestation, most MIA studies have been focusing on the mid-gestation period. Therefore, there is a lack of understanding of the full molecular implication of early MIA events in adult brain.

1.4.3 Dietary Effects

The main goal of schizophrenia research is to identify a effective treatment for schizophrenia. One potential target is the n-3 polyunsaturated fatty acid (PUFA) (**Li2015**; **Trebble2003**). In mouse, it was found that n-3 PUFA can inhibits the production of IL-6 (**Trebble2003**) - a major mediator in MIA model (**Smith2007**). Apart from its anti-inflammatory property, n-3 PUFA such as docosahexaenoic acid (DHA) also plays a critical role in the development of central nervous system (**Clandinin1999**; **Kitajka2002**). Given its strong implication in neuronal functioning, it is possible that n-3 PUFA rich diet may reduce the symptoms of schizophrenia, as reported by a recent study (**Li2015**).

Together, it is interesting for us to not only investigate the effect of MIA during early gestation, but also study the effect of different diet to the treatment of schizophrenia. As technology advances, global messenger RNA (mRNA) expression changes can now be examined using RNA Sequencing, therefore allowing us to investigate effect of early MIA exposure and n-3 PUFA rich diet to the expression pattern in the brain.

1.4.4 RNA Sequencing

Before the development of next generation sequencing (NGS) technology, the global gene expression changes can only be inspected by performing microarray analysis, which is based on probe hybridization. With the development of NGS technology, sequencing can be performed on the mRNA fragments.

Comparing to microarray, RNA Sequencing has a number of advantages, most notably, because RNA Sequencing does not rely on specific probe hybridization, it does not suffer from bias introduced by probe performances such as signal saturation, cross-hybridization, background noises and non-specific hybridization (**Zhao2014**). Furthermore, alternative splicing analysis and de-novo transcript assembly can be readily performed on the same set of RNA Sequencing data. On the other hand, de-novo assembly cannot be detected using microarray and specialized chips are required in order to perform alternative splicing analysis.

However, the superior performance of RNA Sequencing does not come without a cost. The first hurdle in the analysis of RNA Sequencing data is the sequence alignment. As RNA sequencing generates sequence reads from the mRNA transcripts where the introns are usually spliced out, special consideration are required during alignment. Specifically, one can either align the sequence reads directly to the transcriptome or to the reference genome.

The alignment to the transcriptome is relatively simple, the challenge however is that However, multiple isoform can share the same exon. Therefore there are a high level of mapping uncertainties, e.g. a single read can be aligned to multiple transcripts (**Li2011e**). This results in difficulties when trying to quantify the expression level of individual transcripts.

On the other hand, one can align the sequence reads directly to the reference genome. However, due to alternative splicing, splice aware algorithms such as

TopHat2 (**Kim2013**), STAR (**Dobin2013**) and MapSplice (**Wang2010**) must be used instead.

Another difficulty is the differential expression analysis. In RNA Sequencing, the expression of a gene is represented by the number of reads aligned to the gene. However, unlike microarray, where the signal usually follows a normal distribution (**Hoyle2002**; **Giles2003**), the distribution of the RNA Sequencing count data are more complicated.

Early RNA Sequencing experiment assumes the gene expression counts to follows the Poisson distribution (**Marioni2008**), where the variance of the expression is expected to be equal to the mean of the expression. However, it was found that the assumption of Poisson distribution is too restrictive, as an over-dispersion was typically observed in RNA Sequencing data (**Anders2010**). Therefore, to overcome the problem of over-dispersion, differential expression analysis of RNA Sequencing data are required to model the expression using the negative binomial distribution (**Anders2010**; **Robinson2010**) or the beta negative binomial distribution (**trapnell2012differential**), instead of the Poisson distribution.

By using the appropriate aligner for the alignment of the RNA sequencing data, and using the appropriate statistical modeling, RNA Sequencing can provide unprecedented power for the analysis of expression changes. Therefore, RNA Sequencing might be an appropriate tool for the analysis of gene expression changes induced by MIA event and study the effect of the n-3 PUFA diet.

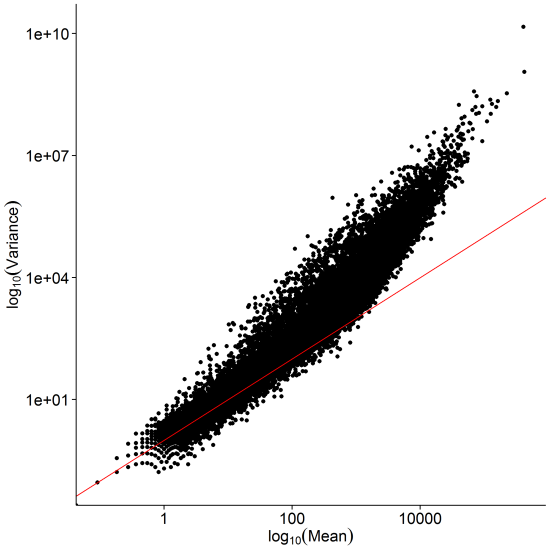


Figure 1.6: Over-dispersion observed in RNA Sequencing Count Data. If the RNA Sequencing count data follows the Poisson distribution, then the mean and variance of the data should be equal (follow the diagonal). However, it was observed that as the mean increases, the variance increases even more, suggesting that there is an over-dispersion in the data.

1.5 Summary

The estimated SNP-heritability for schizophrenia by **Bulik-Sullivan2015** suggest that the common variants might have accounted for most if not all of the heritability of schizophrenia. However, there are multiple concerns, firstly, **Bulik-Sullivan2015** only considered a limited number of cases in the binary trait simulation. It is therefore unclear how the LDSC perform in different binary traits. Moreover, studies suggest that CNV (**Szatkiewicz2014**), structural variations (**Walsh2008**) and also rare mutations (**purcell2014polygenic**) are contributing to the risk of schizophrenia. Also, considering the presence of gene-environment interactions can inflate the total heritability estimated for a disease **zuk2012mystery** it is likely for the SNP-heritability estimated by **Bulik-Sullivan2015** to be too high. It might therefore be interesting to see if an alternative method can result in the same estimates as provided by LDSC.

In this thesis, we would first like to perform an extensive simulation to

investigate the performance of LDSC given different genetic architectures, sampling strategies and types of traits. We are especially interested in the performance of LDSC for binary traits. In addition, we would like to develop an independent algorithm to estimate the SNP-heritability of schizophrenia. This will allow us to assess whether if the SNP-heritability estimated by LDSC is correct.

On the other hand, with the presence of the gene-environment interaction (**Tienari2004**; **Clarke2009**), the total heritability of schizophrenia might be over-estimated **zuk2012mystery** Therefore, the proportion of environmental influence to schizophrenia might be higher than expected. One of the environmental factor interacting with genetic factors is the prenatal infection **Clarke2009** Given its importance in the etiology of schizophrenia, it is of our interest to study the impact of prenatal infection to the transcriptome pattern of the brain. Hence, we perform a RNA Sequencing study to capture gene expression changes induced by early (Gestation Day (GD)9) MIA events in the mouse cerebellum using the PolyI:C mouse model. Additionally, because our lab has recently found that n-3 PUFA rich diet might help to reduce schizophrenia-like behaviour in mice exposed to early MIA insults (**Li2015**),we are also interested in investigating the effect of n-3 PUFA rich diet in the gene expression pattern in the brain of the MIA samples.

In summary, this thesis is divided into three parts. We first investigate the performance of LDSC in Chapter 2 and introduce SNP HeRitability Estimation Kit (SHREK), an alternative algorithm to LDSC for the estimation of SNP-heritability using GWAS summary statistics. This should allow us to assess whether if the SNP-heritability of schizophrenia estimated by LDSC is correct.

In the next chapter, we introduce the RNA Sequencing study on the effect of early MIA insult and n-3 PUFA rich diet on gene expression of the mouse cerebellum. This should provide understanding to the impact of MIA in early gestation and also provide insight into possible treatment target for schizophrenia induced by

MIA.

Lastly, we will summarize and conclude all findings in Chapter 4 where future perspectives on genetic studies of schizophrenia will be provided.

2 Heritability Estimation

2.1 Introduction

Bulik-Sullivan2015 estimated the SNP-heritability of schizophrenia to be around 55.5%, which suggest the common variants have accounted for majority of the heritability of schizophrenia. However, studies have found that CNV (**Szatkiewicz2014**), structural variations (**Walsh2008**) and also rare mutations (**purcell2014polygenic**) can also contribute to the risk of schizophrenia. Additionally, with the presence of gene-environment interaction, the total heritability estimated for schizophrenia might be inflated **zuk2012mystery** It is therefore likely that **Bulik-Sullivan2015** has overestimated the SNP-heritability of schizophrenia.

It is noted that **Bulik-Sullivan2015** only conducted a limited number of simulation on binary trait data, therefore it is unclear how LDSC performs in other condition. Additionally, because **Bulik-Sullivan2015** found that LDSC has a larger standard error for oligogenic traits, it might be beneficial for us to develop an alternative algorithms that has robust performance for all traits.

Herein, we introduce SNP HeRitability Estimation Kit (SHREK), an alternative algorithm to LDSC for the estimation of SNP-heritability using GWAS summary statistics. We also investigate the performance of SHREK and LDSC on different traits by performing a comprehensive simulation, using GCTA as a

golden standard. Finally, we estimate the SNP-heritability of schizophrenia using SHREK and LDSC to investigate whether the SNP-heritability estimated by **Bulik-Sullivan2015** is correct.

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 Basic Concept of SHREK

The fundamental idea of SHREK is that the summary statistics of a SNP from a GWAS should represent the among of phenotypic variance it explain. By definition, the coefficient of determination (r^2) is the variance in the dependent variable that is predictable by the independent variable. If we define the phenotype as the dependent variable and the genotype as the independent variable, the coefficient of determination between the phenotype and genotype equals to the heritability explained by the specific SNP.

Therefore, if we assume the summary statistic of SNP i be t_i, which follow the student-t distribution under the null, the coefficient of determination between

SNP i and the phenotype can be represented as

$$r_i^2 = \frac{t_i^2}{n - 2 + t_i^2} \tag{2.1}$$

where n is the number of samples. When n is large, t_i^2 will converge to χ^2 distribution under the null, with mean equal to 1. Therefore, we can express eq. (2.1) as

$$\hat{r_i}^2 = \frac{t_i^2 - 1}{n - 2 + t_i^2} \tag{2.2}$$

If all SNPs are independent of each other, the SNP-heritability will simply be

$$\text{SNP-heritability} = \sum_i \hat{r_i}^2 \tag{2.3}$$

However, in reality SNPs are correlated with each other and will inflate the summary statistic. Therefore, we must take into account of LD when we estimate the SNP-heritability.

2.2.2 Derivation of SHREK

Here, we provide a step by step derivation of SHREK showing how can the LD be incorporate into the calculation of SNP-heritability using only the summary statistic. First, we assume the phenotype $\boldsymbol{y}$ and genotype matrix $\boldsymbol{X}$ 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 $\boldsymbol{R}$ as the LD matrix.

The linear regression between $\boldsymbol{X}$ and $\boldsymbol{y}$ can then be defined as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2.4}$$

where $\boldsymbol{\epsilon}$ is the error term, account for the non-genetic elements contributing to the phenotype.

As heritability is defined as $\frac{V_G}{V_P}$ where V_G is the variance of the genetic factor and V_P is the variance of the phenotype, we can obtain

$$\text{Heritability} = \frac{\text{Var}(\boldsymbol{X}\boldsymbol{\beta})}{\text{Var}(\boldsymbol{y})} \tag{2.5}$$

$$= \text{Var}(\boldsymbol{X}\boldsymbol{\beta}) \tag{2.6}$$

We then assume $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_m)^t$ follows the distribution:

$$\boldsymbol{\beta} \sim f(0, H)$$

$$\boldsymbol{H} = diag(\boldsymbol{h^2})$$

$$\boldsymbol{h^2} = (h_1^2, h_2^2, \ldots, h_m^2)^t$$

where $\boldsymbol{h^2}$ is the variance of the "true" effect. $\text{Var}(\boldsymbol{X}\boldsymbol{\beta})$ can then be expressed as

$$\begin{aligned}
\text{Var}(\boldsymbol{\beta}\boldsymbol{X}) &= \text{E}_X\text{Var}_{\beta|X}(\boldsymbol{X}\boldsymbol{\beta}) + \text{Var}_X\text{E}_{(\beta|X)}(\boldsymbol{X}\boldsymbol{\beta}) \\
&= \text{E}_X(\boldsymbol{X}^t\text{Var}(\boldsymbol{\beta})\boldsymbol{X}) \\
&= \text{E}_X(\boldsymbol{X}^t\boldsymbol{H}\boldsymbol{X}) \\
&= \text{Tr}(\text{Var}(\boldsymbol{X}\boldsymbol{H})) \\
&= \sum_i h_i^2
\end{aligned} \tag{2.7}$$

Therefore

$$\text{SNP-heritability} = \sum_i h_i^2 \tag{2.8}$$

Now the covariance between SNP_i ($\boldsymbol{x_i}$) and $\boldsymbol{y}$ can be expressed as

$$\begin{aligned}\text{Cov}(\boldsymbol{x_i}, \boldsymbol{y}) &= \text{Cov}(\boldsymbol{x_i}, \boldsymbol{X\beta} + \boldsymbol{\epsilon}) \\ &= \text{Cov}(\boldsymbol{x_i}, \boldsymbol{X\beta}) \\ &= \text{Cov}(\boldsymbol{x_i}, \boldsymbol{X})\boldsymbol{\beta} \\ &= \boldsymbol{R_i\beta} \end{aligned} \tag{2.9}$$

As both $\boldsymbol{X}$ and $\boldsymbol{y}$ are standardized, their covariance is equal to their correlation

$$\text{Cov}(\boldsymbol{x_i}, \boldsymbol{y}) = \text{Cor}(\boldsymbol{x_i}, \boldsymbol{y}) = \rho_i$$

In reality, the *observed* correlation is usually error-prone. Therefore the observed correlation between SNP_i and the phenotype is defined as

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

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

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

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

$$\text{E}(\epsilon_i|\rho_i) \sim 0 \tag{2.10}$$

$$\text{Var}(\epsilon_i|\rho_i) \sim 1 \tag{2.11}$$

If we define the z-statistic and χ^2-statistic as

$$z_i = \hat{\rho}_i\sqrt{n} \tag{2.12}$$

$$\begin{aligned}\chi_i^2 &= z_i^2 \\ &= \hat{\rho_i^2}n \end{aligned} \tag{2.13}$$

We can express χ^2 as

$$\begin{aligned}\chi_i^2 &= \hat{\rho_i^2} n \\ &= n\left(\boldsymbol{R_i}\boldsymbol{\beta} + \frac{\epsilon_i}{\sqrt{n}}\right)^2 \end{aligned} \tag{2.14}$$

and

$$\begin{aligned}\mathrm{E}(\chi_i^2) &\approx n\boldsymbol{R_i^t}\boldsymbol{H}\boldsymbol{R_i} + 1 \\ &= n\sum_j R_{ij}^2 h_j^2 + 1 \end{aligned} \tag{2.15}$$

$$\tag{2.16}$$

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

$$\begin{aligned}&\sum_i \left(\chi_i^2 - \mathrm{E}\left(\chi_i^2\right)\right)^2 \\ =&\sum_i \left(\chi_i^2 - \left(n\sum_j R_{ij}^2 \hat{h_j^2} + 1\right)\right)^2 \end{aligned} \tag{2.17}$$

By defining

$$f_i = \frac{\chi_i^2 - 1}{n} \tag{2.18}$$

eq. (2.17) becomes

$$\sum_i \left(\chi_i^2 - \mathrm{E}\left(\chi_i^2\right)^2\right) = \sum_i \left(f_i - \sum_j R_{ij}^2 h_j^2\right)^2 \tag{2.19}$$

$$= \boldsymbol{f^t f} - \boldsymbol{2 f^t R_{sq}\hat{h^2}} + \boldsymbol{\hat{h^2} R_{sq}^t R_{sq}\hat{h^2}} \tag{2.20}$$

where $\boldsymbol{R_{sq} = R \circ R}$ and $\circ$ denotes the element-wise product (Hadamard product). By differentiating with respect to $\hat{h^2}$ and set to 0,we get

$$\begin{aligned}2\boldsymbol{R_{sq}^t R_{sq}\hat{h^2}} - 2\boldsymbol{R_{sq} f} &= 0 \\ \boldsymbol{R_{sq}\hat{h^2}} &= \boldsymbol{f} \end{aligned} \tag{2.21}$$

Remember from eq. (2.8) that SNP-heritability can be represent as $\sum_i h_i^2$,

therefore we can express the estimated SNP heritability as

$$\begin{aligned} \boldsymbol{R_{sq}}\hat{\boldsymbol{h^2}} &= \boldsymbol{f} \\ \hat{\boldsymbol{h^2}} &= \boldsymbol{R_{sq}^{-1}}\boldsymbol{f} \\ \text{SNP-heritabilility} &= \mathbf{1}^t\boldsymbol{R_{sq}^{-1}}\boldsymbol{f} \end{aligned} \tag{2.22}$$

2.2.3 Calculation of Standard Error

It is also essential to estimate the standard error of the estimate. Based on eq. (2.22), the variance of the SNP-heritability can be defined as

$$\text{Var}(\hat{\text{SNP-heritability}}) = \mathbf{1}^t\boldsymbol{R_{sq}^{-1}}\text{Var}(\boldsymbol{f})\boldsymbol{R_{sq}^{-1}}\mathbf{1} \tag{2.23}$$

Based on eq. (2.23), it is observed that in order to calculate the variance of the estimate, $\text{Var}(\boldsymbol{f})$ is required.

We first consider genotype x_i has a standard normal mean z_i and non-centrality parameter μ_i such that

$$\begin{aligned} \text{E}[x_i] &= \text{E}[z_i + \mu_i] \\ &= \mu_i \end{aligned} \tag{2.24}$$

$$\begin{aligned} \text{Var}(x_i) &= \text{E}[x_i^2] - \text{E}[x_i]^2 \\ &= \text{E}[z_i^2 + \mu_i^2 + 2z_i\mu_i] - \mu_i^2 \\ &= \text{E}[z_i^2] + \text{E}[\mu_i^2] - \mu_i^2 \\ &= 1 \end{aligned} \tag{2.25}$$

$$\begin{aligned} \text{Cov}(x_i, x_j) &= \text{E}[x_i x_j] - \text{E}[x_i]\text{E}[x_j] \\ &= \text{E}[(z_i + \mu_i)(z_j + \mu_j)] - \mu_i\mu_j \\ &= \text{E}[z_i z_j] \end{aligned} \tag{2.26}$$

Because the genotypes are standardized, $\text{Cov}(x_i, x_j) = \text{Cor}(x_i, x_j) = R_{ij}$, with R_{ij}

being the LD between SNP_i and SNP_j. $\text{Cov}(\chi_i^2, \chi_j^2)$ can then be calculated as

$$\begin{aligned}
\text{Cov}(\chi_i^2, \chi_j^2) &= \text{Cov}\left((z_i + \mu_i)^2, (z_j + \mu_j)^2\right) \\
&= \text{Cov}\left(z_i^2 + \mu_i^2 + 2z_i\mu_i, z_j^2 + \mu_j^2 + 2z_j\mu_j\right) \\
&= \text{Cov}(z_i^2, z_j^2) + 4\mu_i\mu_j\text{Cov}(z_i, z_j) \\
&= \text{E}(z_i^2 z_j^2) - \text{E}(z_i^2)\text{E}(z_j^2) + 4\mu_i\mu_j\text{E}(z_i z_j) - \text{E}(z_i)\text{E}(z_j) \\
&= \text{E}(z_i^2 z_j^2) + 4\mu_i\mu_j\text{E}(z_i z_j) - 1
\end{aligned} \tag{2.27}$$

As

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

and $\text{E}[z_i z_j] = R_{ij}$, $\text{E}[z_i^2 z_j^2]$ can be expressed as

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

As a result, $\text{Cov}(\chi_i^2, \chi_j^2)$ becomes

$$\text{Cov}(\chi_i^2, \chi_j^2) = 2R_{ij}^2 + 4R_{ij}\mu_i\mu_j \tag{2.28}$$

But then,

$$\begin{aligned}
\text{E}(z_i z_j) &= R_{ij} + \mu_i\mu_j \\
\text{E}(z_i z_j - R_{ij}) &= \mu_i\mu_j
\end{aligned} \tag{2.29}$$

Therefore

$$\mathrm{Cov}(\chi_i^2, \chi_j^2) = 4R_{ij}z_i z_j - 2R_{ij}^2 \tag{2.30}$$

Additionally, it was observed that

$$\begin{aligned} \mathrm{E}[\hat{z}_i\hat{z}_j] &= \mathrm{E}[(z_i + \mu_i)(z_j + \mu_j)] \\ &= \mu_i\mu_j + \mathrm{E}[z_i z_j] \\ &= \mu_i\mu_j + R_{ij} \end{aligned} \tag{2.31}$$

where $\hat{z}_i$ is the *observed* z-statistic ($\hat{\boldsymbol{z}} = \sqrt{\boldsymbol{\chi^2}}$), with the direction of effect as its sign). Substituting eq. (2.31) and eq. (2.30) into eq. (2.23) in matrix form, the variance of the estimate can be expressed as

$$\mathrm{Var}(\hat{\text{SNP-heritability}}) = \mathbf{1}^t \boldsymbol{R}_{\boldsymbol{sq}}^{-1} \frac{2\boldsymbol{R}_{\boldsymbol{sq}} + 4\boldsymbol{R} \circ \hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^t}{n^2} \boldsymbol{R}_{\boldsymbol{sq}}^{-1} \mathbf{1} \tag{2.32}$$

Given the direction of effect, eq. (2.32) should in theory provide the correct estimate of the standard error. However, the direction of effect are not always available and in such scenario, eq. (2.32) cannot provide an accurate estimation of the standard error. As $n \times \boldsymbol{f} + 1$ is approximately χ^2 distributed, we might view eq. (2.21) 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, the "effective number"(e) of association (**Li2011**) can be calculated. Substituting e into the variance equation of non-central χ^2 distribution yields

$$\mathrm{Var}(H) = \frac{2(e + 2H)}{n^2} \tag{2.33}$$

Here, eq. (2.33) serves as a heuristic estimation of the standard error, which does not require the direction of effect.

2.2.4 Liability Threshold Model

In case control studies, the proportion of cases is usually (much) larger than the prevalence in the population, leading to ascertainment bias (**lee2011estimating**). Therefore, to adjust for the ascertainment bias, the estimated heritability needs to be transported to a liability scale.

Under the liability threshold model, a sample is consider as affected if his/her liability is above the liability threshold. The mean values of disease liability in affected and unaffected individuals are then defined as $\frac{z}{K}$ and $-\frac{z}{1-K}$, respectively, where z and K are the height of the standard normal curve at the threshold liability and the population risk. Let v be the proportion of affected individuals in a sample, then the expected mean of liability ($\mathrm{E}(L)$) in the sample is

$$\begin{aligned}\mathrm{E}(L) &= v\frac{z}{K} + (1-v)(-\frac{z}{1-K}) \\ &= \frac{z(v-K)}{K(1-K)}\end{aligned} \tag{2.34}$$

and the square of mean liability is

$$\mathrm{E}(L^2) = \frac{z^2(K^2 - 2vK - v)}{K^2(1-K)^2} \tag{2.35}$$

Based on the above equations, the variance of liability can be calculated as

$$\begin{aligned}\mathrm{Var}(L) &= \mathrm{E}(L^2) - (\mathrm{E}(L))^2 \\ &= \frac{z^2(v(1-v))}{K^2(1-K)^2}\end{aligned} \tag{2.36}$$

Therefore, in case-control sample, the summary statistic is attenuated from the standard scenario by a factor of $\mathrm{Var}(L)$ and the SNP-heritability estimated by SHREK can be adjusted as follow

$$\hat{SNP - Heritability} = \frac{K^2(1-K)^2}{z^2 v(1-v)}\mathbf{1}^t \boldsymbol{R}_{\boldsymbol{sq}}^{-1}\boldsymbol{f} \tag{2.37}$$

2.2.5 Extreme Phenotype Selection

Similarly, in scenarios where extreme phenotype select was performed, both the summary statistics from the association and the heritability estimation can be increased 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 (**Sham2014**). Thus, to adjust for the inflation, $\frac{V_P}{V_{P'}}$ can be multiplied to the estimate from SHREK

$$\text{SNP-H}\hat{\text{e}}\text{ritability} = \frac{V_P}{V_{P'}}\mathbf{1}^t \boldsymbol{R}_{\boldsymbol{sq}}^{-1}\boldsymbol{f} \tag{2.38}$$

2.2.6 Inverse of the Linkage Disequilibrium Matrix

Using eq. (2.22), the SNP-heritability can be estimated from the GWAS summary statistics. When $\boldsymbol{R}_{\boldsymbol{sq}}$ is full rank and positive definite, eq. (2.21) can be solved using the QR decomposition or LU decomposition without explicitly calculating the inverse of $\boldsymbol{R}_{\boldsymbol{sq}}$. We can then take the sum of $\hat{\boldsymbol{h}}^{\mathbf{2}}$ to get the answer for eq. (2.22).

However, LD matrices are usually ill-conditioned. Therefore the solution of eq. (2.21) is prone to large numerical errors (**Neumaier1998**). In order to solve eq. (2.21), regularization techniques such as Tikhonov Regularization (also known as Ridge Regression) and Truncated Singular Value Decomposition (tSVD) can be performed (**Neumaier1998**). 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

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^t \tag{2.39}$$

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. (2.39), the inverse of $\boldsymbol{A}$ can be expressed as

$$\boldsymbol{A}^{-1} = \boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^t \tag{2.40}$$

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

When the vector $\boldsymbol{b}$ is collected with some error $\boldsymbol{\epsilon}$, the solution to $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ becomes:

$$\begin{aligned} \boldsymbol{x} &= \boldsymbol{A}^{-1}(\boldsymbol{b} + \boldsymbol{\epsilon}) \\ &= \boldsymbol{A}^{-1}\boldsymbol{b} + \boldsymbol{A}^{-1}\boldsymbol{\epsilon} \\ &= \boldsymbol{x}^* + \boldsymbol{A}^{-1}\boldsymbol{\epsilon} \end{aligned} \tag{2.41}$$

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

$$\begin{aligned} \delta\boldsymbol{x} &= \boldsymbol{x} - \boldsymbol{x}^* \\ &= \boldsymbol{A}^{-1}\boldsymbol{\epsilon} \end{aligned} \tag{2.42}$$

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

$$\begin{aligned} \frac{||\delta\boldsymbol{x}||}{||\boldsymbol{x}||} \div \frac{||\boldsymbol{\epsilon}||}{||\boldsymbol{b}||} &= \frac{||\delta\boldsymbol{x}||}{||\boldsymbol{\epsilon}||}\frac{||\boldsymbol{b}||}{||\boldsymbol{x}||} \\ &= \frac{||\boldsymbol{A}^{-1}\boldsymbol{\epsilon}||}{||\boldsymbol{\epsilon}||}\frac{||\boldsymbol{A}\boldsymbol{x}||}{||\boldsymbol{x}||} \\ &= ||\boldsymbol{A}^{-1}||\ ||\boldsymbol{A}|| \end{aligned} \tag{2.43}$$

where $||.||$ is the matrix norm. This is defined as the condition number of matrix $\boldsymbol{A}$ ($\kappa(\boldsymbol{A})$). A matrix with a high condition number is said to be ill-conditioned, which might result in unstable or inaccurate approximation to the matrix equation.

To obtain a meaningful solution for ill-conditioned/singular matrix $\boldsymbol{A}$,

tSVD method can be performed to obtain a pseudo inverse of $\boldsymbol{A}$. The tSVD of $\boldsymbol{A}$ is defined as

$$\boldsymbol{A}^{+} = \boldsymbol{U}\boldsymbol{\Sigma}_k\boldsymbol{V}^t \qquad \text{and} \qquad \boldsymbol{\Sigma}_k = \text{diag}(\sigma_1, \ldots, \sigma_k, 0, \ldots, 0) \tag{2.44}$$

where $\boldsymbol{\Sigma}_k$ equals to replacing the smallest $n-k$ singular value by 0 (**Hansen1987**). Alternatively, we can define

$$\sigma_i = \begin{cases} \sigma_i & \text{for} \quad \sigma_i \geq t \\ 0 & \text{for} \quad \sigma_i < t \end{cases} \tag{2.45}$$

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, providing a reasonable approximation to $\boldsymbol{x}$. A problem with tSVD is that it only works when matrix $\boldsymbol{A}$ has a well determined numeric rank (**Hansen1987**). That is, the performance of tSVD is optimal 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. Small numerical error might change the threshold, thus leads to unstable results. One way to calculate if a matrix has well-defined rank is to calculate the "gap" in its singular values

$$gap = \frac{\sigma_k}{\sigma_{k+1}} \tag{2.46}$$

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

To examine whether if the tSVD is suitable for SHREK, we examined the LD matrix formed using the 1,000 genome European samples. In brief, only SNPs on chromosome 22 was used. SNPs within a 1 megabase (mb) region is considered as a "block" where three "blocks" form a single "window". The maximum "gap" of this LD window is calculated. We then transverse the genome one block at a time and then calculate the mean maximum "gap". It was found that the mean

maximum "gap" for all windows on chromosome 22 is 5,262,198,714,018,811, which suggest that the LD matrix have a well-defined rank. Thus, the choice of tSVD for the regularization is appropriate.

In view of this, tSVD was selected as the method for regularization for solving eq. (2.21). 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 SHREK.

2.2.7 Implementation

SHREK is implemented using the C++ programming languages (version C++11) and the matrix algebra was performed using the EIGEN C++ header library (**eigenweb**). Although the Armadillo library (**Sanderson2010**) is faster than EIGEN (**Ho2011**), the speed can only be achieved when addition libraries such as OpenBLAS were installed. The use of EIGEN therefore simplify the programme installation, making it more user friendly.

Although tSVD can provide an approximation to the ill-posed eq. (2.21), it is an $O(n^3)$ algorithm, making the computation run time prohibitive when the number of SNPs is large. Unfortunately, the number of SNPs in a GWAS is generally high, making it impossible to calculate the tSVD of the whole genome.

From eq. (2.39), it is noted that the matrix $\boldsymbol{U}$ and $\boldsymbol{V}$ are the eigenvectors of $\boldsymbol{A}\boldsymbol{A}^t$ and $\boldsymbol{A}^t\boldsymbol{A}$ respectively, for any symmetric matrix such as the LD matrix, $\boldsymbol{U}$ and $\boldsymbol{V}$ are identical. Therefore, tSVD can be performed using eigenvalue decomposition, which is more efficient than SVD. However, eigenvalue decomposition is still an $O(n^3)$ algorithm. Therefore, it is still difficult to perform the full decomposition across the whole genome.

Given that it is unlikely for SNPs more than 1 mb apart to be in LD with each other, the genome is separated into 1 mb blocks where every 3 blocks form one single window. eq. (2.21) is solved for every window where only the middle block is updated. The only exception is the first and last window where the first and last block are also updated respectively. Sliding window is then performed where the algorithm will transverse through the genome with a step size of 1 until the whole genome is decomposed. This essentially break down the problem into small pieces where the computation of eq. (2.21) is feasible.

2.2.8 Comparing with LDSC

When compared with LDSC, SHREK is taking a bottom up approach. To illustrate this difference, one can imagine in an extreme unlikely scenario where the GWAS contains only one single SNP, the SNP-heritability estimated from SHREK rely solely on the summary statistic of the single SNP and does not rely on any additional information. Whereas for LDSC, as it relies on the LD score for the estimation of the SNP-heritability, one can see that for the same summary statistic, the estimation can change as the LD score of the target SNP changed. Although this scenario is almost impossible for today's GWAS, it serves as an example to illustrate the fundamental different between SHREK and LDSC.

Additionally, when estimating the SNP-heritability, SHREK only requires SNPs found on both the GWAS chip and the reference panel. On the other hand, LDSC can use any number of SNPs in the calculation of the LD score. Although it allows for flexibility, it introduce ambiguity in the estimation. It is possible that when a different number of SNPs are included in the calculation of the LD score, a different estimation is obtained.

Finally, for the estimation of the standard errors, LDSC uses delete-one

jackknife whereas SHREK uses both eq. (2.33) and eq. (2.32). Under ideal situation, SHREK should in theory provide a more accurate estimate for its standard error when compared to LDSC.

2.2.9 Comparing Different LD correction Algorithms

Similar to LDSC, SHREK relies heavily on the LD matrix for the accurate estimation of SNP-heritability. Therefore, it is crucial that the LD estimates are accurate.

When using SHREK, the LD matrix is usually estimated from the reference panel such as the 1000 genome project (**Project2012**) or the HapMap project (**Altshuler2010**), which are samples from the population. Therefore, LD estimated from these reference panel may contain sampling errors, which might result in bias in the estimate of SHREK.

For an accurate estimation of the SNP-heritability, it might be important to correct for the sampling error in the LD. Varies method for the correction of sample R^2 has been proposed by **Weir1980**; **Wang2007**

$$\text{Ezekiel} : \tilde{R^2} = 1 - \frac{n-1}{n-2}(1-\hat{R^2}) \tag{2.47}$$

$$\text{Olkin-Pratt} : \tilde{R^2} = 1 - \frac{(n-3)(1-\hat{R^2})}{n-2}(1+\frac{2(1-\hat{R^2})}{n}) \tag{2.48}$$

$$\text{Pratt} : \tilde{R^2} = 1 - \frac{(n-3)(1-\hat{R^2})}{n-2}(1+\frac{2(1-\hat{R^2})}{n-3.3}) \tag{2.49}$$

$$\text{Smith} : \tilde{R^2} = 1 - \frac{n}{n-1}(1-\hat{R^2}) \tag{2.50}$$

$$\text{Weir} : \tilde{R^2} = \hat{R^2} - \frac{1}{2n} \tag{2.51}$$

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

To assess the performance of each individual correction methods, a simulation was performed. 5,000 SNPs with MAF ≥ 0.05 were first randomly selected

from chromosome 22 from the 1,000 genome Northern Europeans from Utah (CEU) haplotypes and were used as an input to HAPGEN2 (**Su2011**) to simulate 1,000 individuals. HAPGEN2 is a simulation tool 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 (**Li2003**). This allows for the simulation of genotypes with LD structures comparable to those observed in CEU population. We than randomly select 100 SNPs as the causal variants. As **Orr1998** suggested that the exponential distribution could be used to approximate the genetic architecture of adaptation, effect size of the causal variants were simulated with an exponential distribution with $\lambda = 1$

$$\begin{aligned} \theta &= \exp(\lambda = 1) \\ \beta &= \pm\sqrt{\frac{\theta \times h^2}{\sum \theta}} \end{aligned} \tag{2.52}$$

with a random direction of effect.

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 where $h^2 \in \{0.0, 0.1, \ldots, 0.9\}$ using

$$\begin{aligned} \epsilon_i &\sim N(0, \mathrm{Var}(\boldsymbol{X\beta})\frac{1-h^2}{h^2}) \\ \boldsymbol{\epsilon} &= (\epsilon_1, \epsilon_2, ..., \epsilon_n)^t \\ \boldsymbol{y} &= \boldsymbol{X\beta} + \boldsymbol{\epsilon} \end{aligned} \tag{2.53}$$

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. Finally, PLINK (**Purcell2007**) was used to perform the genotype association and the resulting summary statistics were used as an input to SHREK with different LD correction algorithms used.

The whole process was repeated 100 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$>$ 0.05 from chromosome 22
2. Simulate 500 samples using HAPGEN2 and used as the reference panel
3. Randomly generate 100 effect sizes with eq. (2.52)
4. Randomly assign the effect sizes to 100 SNPs with $h^2 \in \{0.0, 0.1, \ldots, 0.9\}$
5. Simulate 1,000 samples using HAPGEN2 and calculate their phenotype according to eq. (2.53)
6. Perform heritability estimation using SHREK with different LD correction algorithm
7. Repeat 100 times

2.2.10 Simulation Study

One of the main purpose of this chapter is to assess the performance of LDSC when dealing with various phenotypes. Therefore, simulations were performed for quantitative and binary traits with different genetic architectures. The effect of sampling strategies, such as random sampling and extreme phenotype selection, on the heritability estimation were also investigated.

As GCTA (**Yang2011**) is the most commonly used programme for the estimation of SNP-heritability from GWAS data, it was also included in the analysis as a reference point. It is noted that as no confounding factors were simulated, the intercept estimation function in LDSC will be penalized with a larger standard error, leading to an unfair comparison. Therefore, performance of LDSC with fixed

intercept (--no-intercept) were also inspected to avoid bias against LDSC.

2.2.10.1 Sample Size

Of all the parameters, sample size is of the most importance in determining the standard error of the estimates. As sample size increases, the samples will be more representative of the true population, thus provide more accurate estimation of the parameters, resulting in a smaller standard error. Using simple text mining, the sample size distribution of GWAS was obtained from the GWAS catalog (**Welter2014**). 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 reduce the computation time required for the simulation, only 1,000 samples were simulated in each simulations unless otherwise stated.

2.2.10.2 Number of SNPs in Simulation

In each simulation, with the exception of the binary trait simulation, 50,000 SNPs from chromosome 1 were simulated , which correspond to 200 SNPs within a 1 mb region. 50,000 SNPs from chromosome 1 was selected because this is the largest amount of SNPs with adequate density that can be simulated given the computation limitation.

2.2.10.3 Genetic Architecture

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 (**Su2011**), 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.

Finally, 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. (2.52) and the phenotype were simulated using eq. (2.53).

2.2.10.4 Quantitative Trait Simulation

As a prove of concept study, quantitative traits were simulated to investigate the performance of the algorithms, especially for SHREK.

First, 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 SHREK, 500 independent samples were simulated as the reference panel for the calculation of LD scores and LD matrix respectively. The association between the genotype and phenotype were calculated using PLINK (**Purcell2007**). The summary statistics and the reference panel were then provided for LDSC and SHREK 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 was repeated 50 times such that a

distribution of the estimate can be obtained. In total, 10 independent populations were simulated.

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. (2.52), 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. (2.53)
6. Perform heritability estimation using SHREK, GCTA, LDSC with fixed intercept and LDSC with intercept estimation.
7. Repeat step 5-6 50 times
8. Repeat step 1-7 10 times

2.2.10.5 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. **Gui2013** 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

$$\begin{aligned} \beta_{eL} &= \pm\sqrt{\frac{0.5h^2}{m}} \\ \beta_{eS} &= \pm\sqrt{\frac{0.5h^2}{100-m}} \\ \beta &= \{\beta_{eL}, \beta_{eS}\} \end{aligned} \tag{2.54}$$

The effect sizes were then randomly assigned to 100 causal SNPs and phenotypes were calculated using eq. (2.53). 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. (2.54), 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. (2.53)
6. Perform heritability estimation using SHREK, LDSC with fixed intercept, LDSC with intercept estimation and GCTA
7. Repeat step 5-6 50 times
8. Repeat step 1-7 10 times

2.2.10.6 Binary Traits

As **Bulik-Sullivan2015** only performed a limited amount of binary trait simulation, a comprehensive simulation on binary trait is therefore required to thoroughly assess the performance of LDSC when dealing with binary traits.

To simulate binary traits, two additional parameters, the population prevalence (p) and observer prevalence (q), have to be taken into consideration. In order to simulate a trait with population prevalence of p and observed prevalence of q with n cases, $\min(\frac{n}{p}, \frac{n}{q})$ samples were required to be simulated. For example, if the observed prevalence is 50% with the population prevalence of 1%, a minimum of 100,000 needs to be simulated in order to obtain 1,000 cases. Therefore when the population prevalence is small, a tremendous amount of computational resources are required in order to perform the simulation. To reduce the burden of computation, the observed prevalence was limited to 50% and only 5,000 SNPs were simulated from chromosome 22. By changing from chromosome 1 to chromosome 22, the number of SNPs simulated can be reduced without significantly reducing the SNP density.

To investigate the effect of population prevalence and the heritability of the traits to the performance of the algorithms, different population prevalence (p) were simulated with $p \in \{0.5, 0.1, 0.05, 0.01\}$. The heritability of the trait were also varied from 0 to 0.9 with increment of 0.1.

In brief, 5,000 SNPs with MAF > 0.05 were randomly selected from chromosome 22 as an input to HAPGEN2. k causal SNPs with $k \in \{10, 50, 100, 500\}$ were randomly selected, each with effect sizes simulated based on eq. (2.52). $\frac{1{,}000}{p}$ samples were then simulated and their phenotype were calculated using eq. (2.53). The phenotypes were then standardized and samples with phenotype passing the liability threshold with respect to p were defined as case. An equal amount of controls were then randomly selected from samples with phenotype below the liability threshold.

In summary, the binary trait simulation follows

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. (2.52) 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. (2.53)
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 SHREK, LDSC with fixed intercept, LDSC with intercept estimation and GCTA
8. Repeat step 5-7 50 times
9. Repeat step 1-8 10 times

2.2.10.7 Extreme Phenotype Sampling

GWAS provides unprecedented power to perform hypothesis-free genetic association throughout the whole genome. However, due to budget constraint, selecting individuals with extreme phenotypes for GWAS is often one way to maximize the genetic signal by enriching the protective/risk common allele at both ends of the distribution, thus increasing the statistical power (**Guey2011**). 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 (**Sham2014**). To investigate the effect of extreme phenotype sampling on the performance of SNP-heritability estimation, we also performed the extreme phenotype simulation.

Again, 50,000 SNPs with MAF > 0.05 were selected from chromosome 1

and were used as an input for HAPGEN2 and 500 samples were simulated to serve as the reference panel for LDSC and SHREK.

From the 50,000 SNPs, 100 SNPs were randomly selected as the causal SNPs and their effect sizes were simulated using eq. (2.52). Two settings are considered: sampling 10% high and 10% low extreme phenotypes ($K = 0.1$); sampling 20% high and 20% low extreme phenotypes ($K = 0.2$). A total of $\frac{500}{K}$ samples were simulated where the sample phenotypes were calculated using eq. (2.53). Phenotypes were then standardized and 500 samples from each extreme ends of the phenotype distribution such that a total of 1,000 samples were obtained. PLINK was used to perform the quantitative trait association analysis and provide the required summary statistics for down-stream analysis. 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 all samples.

As the extreme phenotype sampling were not natively supported by the LDSC and GCTA, to allow for a fair comparison, extreme phenotype adjustment from **Sham2014** 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. (2.52), with heritability ranging from 0 to 0.9 (increment of 0.1)
4. Randomly assign the effect sizes to 100 SNPs

5. Simulate $\frac{500}{K}$ 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. (2.53) 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 SHREK, 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 **Sham2014**

11. Repeat step 5-10 50 times

12. Repeat step 1-11 10 times

2.2.11 Application to Real Data

The main goal of the current study is to investigate whether if the SNP-heritability of schizophrenia estimated by LDSC is correct. Therefore, we estimate the SNP-heritability of schizophrenia with SHREK alongside LDSC, using the summary statistics from the PGC schizophrenia GWAS (**Ripke2014**) as an input. As a control, we also estimated the SNP-heritability for bipolar and major depression disorder (**sklar2011large**; **Ripke2013b**) as a control.

The reference genome were downloaded from 1000 genome (hg19) (**Project2012**) and were converted to PLINK binaries using the PLINK --vcf function. The Euro-

pean 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. LDSC analysis were then performed with and without the intercept estimation (--no-intercept) to serve as a baseline comparison.

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 (**Hinrichs2006**) 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 of the disorders were only estimated using the autosomal SNPs. Furthermore, the MHC region (chr6:25,000,000-35,000,000) was removed from the analysis due to its unusual LD and genetic architecture (**Bulik-Sullivan2015**).

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 **Bulik-Sullivan2015** 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 SHREK. This will results in an inflation in the final estimates. Therefore estimates from SHREK can only serve as an upper bound of the true SNP heritability.

2.3 Results

SHREK is now available on `https://github.com/choishingwan/shrek`.

2.3.1 LD Correction

As SHREK relies on the LD structure to estimate the SNP heritability, it is important to correct for bias in the LD estimates. The performance of the correct algorithms were tested through the HAPGEN2 simulation (fig. 2.1). It is observed that when no bias correction was applied, the mean estimates biased downward, as expected.

The main purpose of this simulation is to test how the LD correction methods affect the performance of SHREK. With the exception of the formula proposed by **Weir1980** (eq. (2.51)), all correction methods results in estimates that are upwardly biased. We hypothesize that it might be because of "over-adjustment" by these method, therefore lead to an inflated estimates. Given these results, it is concluded that **Weir1980**'s formula has the best performance and are therefore selected as the default LD correction algorithm for SHREK.

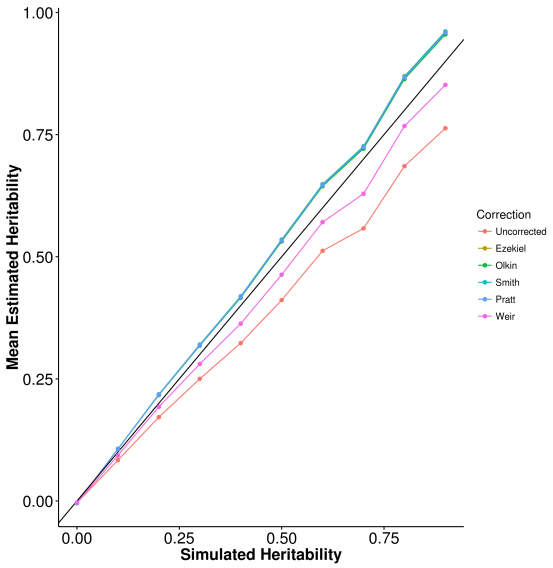


Figure 2.1: Effect of LD correction to Heritability Estimation. We compared the performance of SHREK when different R^2 bias correction algorithm was used. When no bias correction was carried out, a downward bias was observed. After the application of the bias correction algorithms, the mean estimations of all except in the case of Weir eq. (2.51) algorithms leads to an overestimation of heritability.

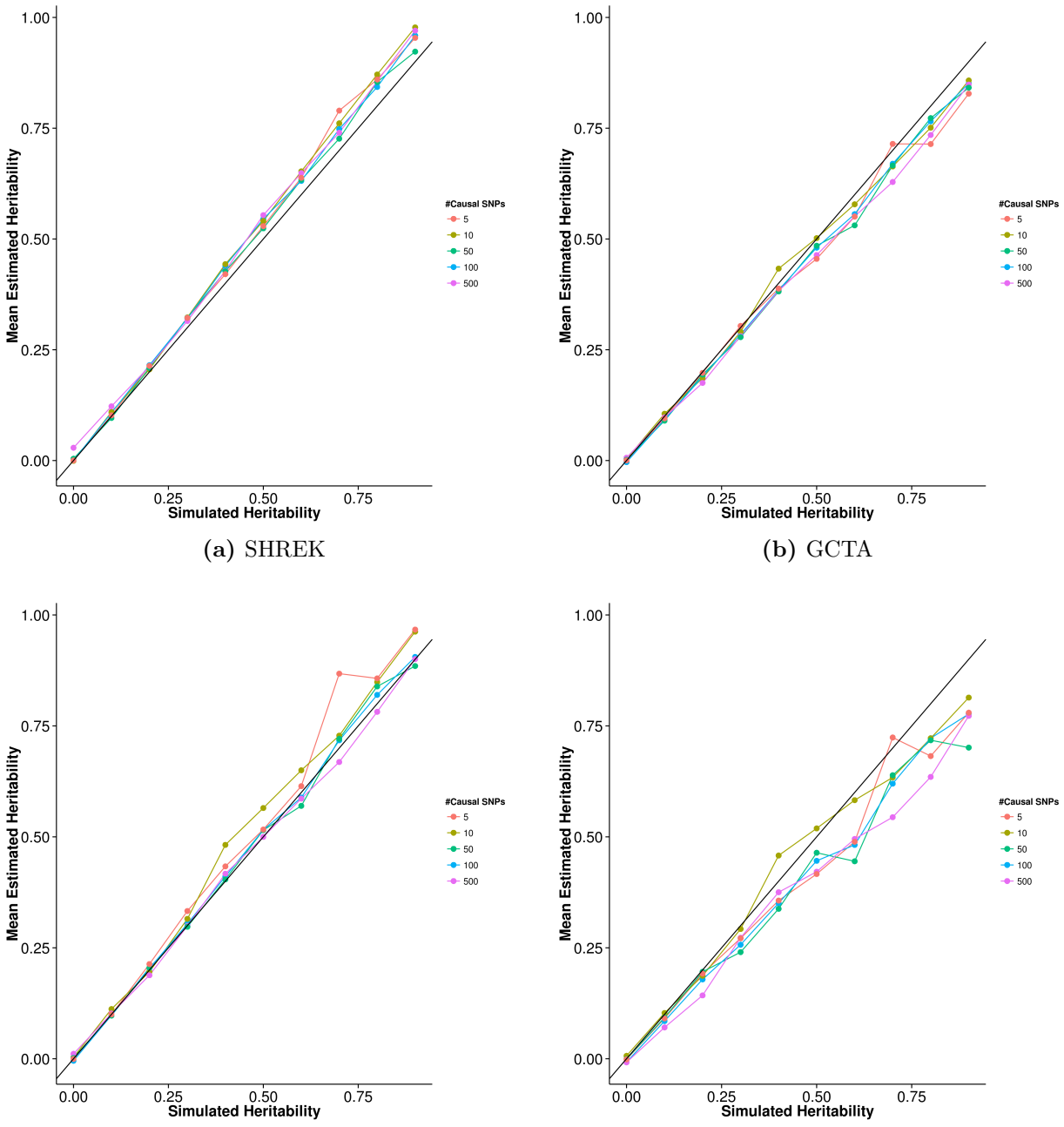
2.3.2 Simulation Study

To formally assess the performance of LDSC and SHREK under different scenarios, a number of simulations were performed.

2.3.2.1 Quantitative Trait Simulation

By varying the number of causal SNPs and the heritability of the trait, performance of LDSC, SHREK and GCTA can be investigated.

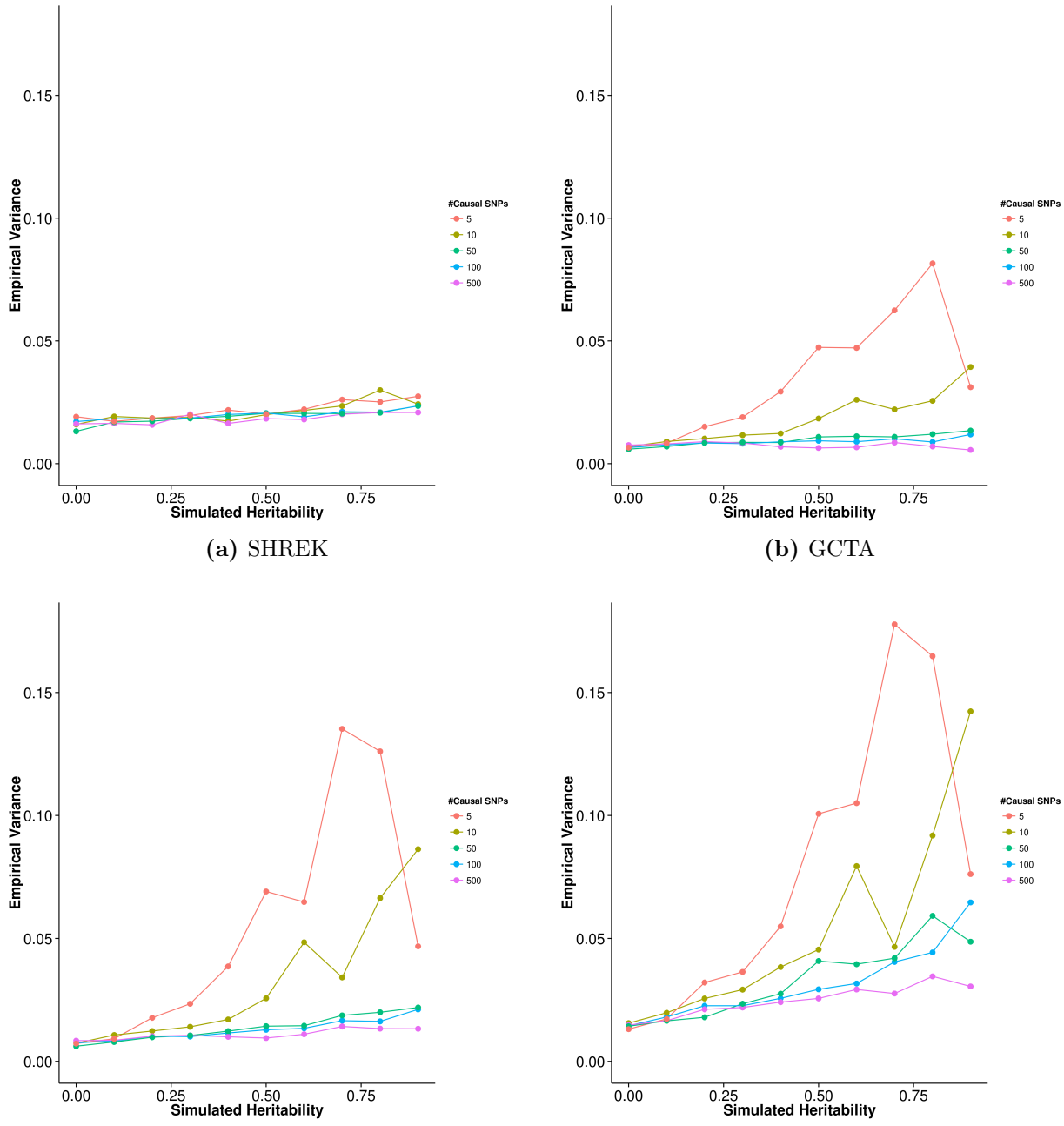
First, when comparing the mean estimates to the simulated heritability, a small upward bias is observed in the estimates from SHREK (fig. 2.2a). On the other hand, estimates from GCTA are moderately biased downward (fig. 2.2b), similar to the estimates from LDSC with intercept estimation (fig. 2.2d), but with a smaller variability. When the intercept was fixed, LDSC can accurately estimate the SNP heritability where an upward bias can only be observed when the number of SNPs



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

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

Figure 2.2: 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.



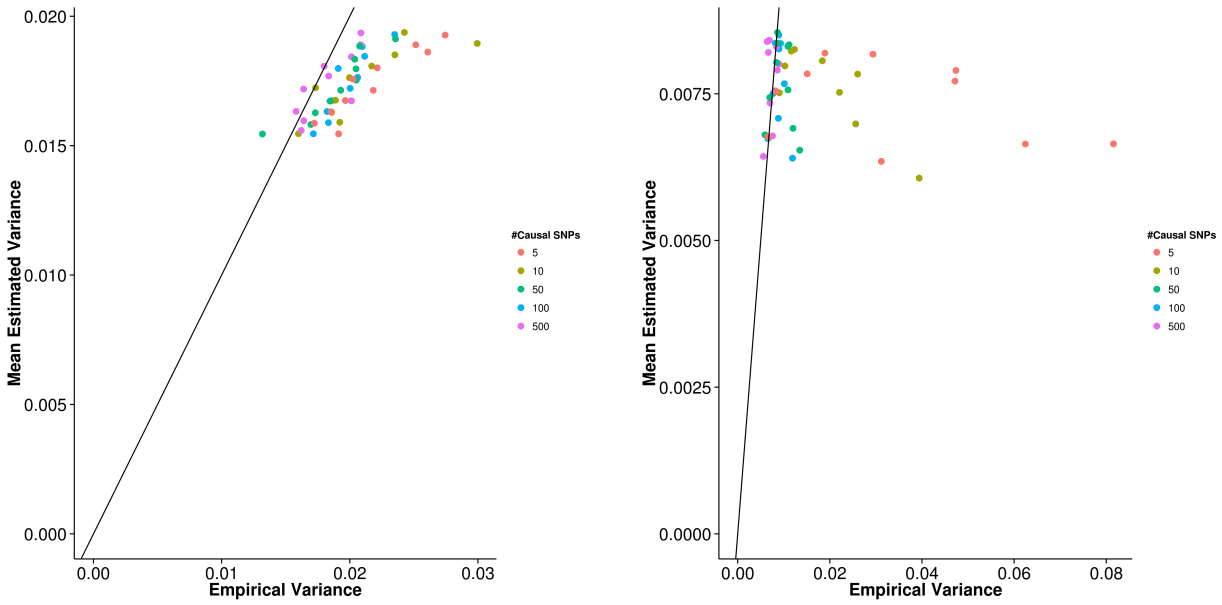
(a) SHREK

(b) GCTA

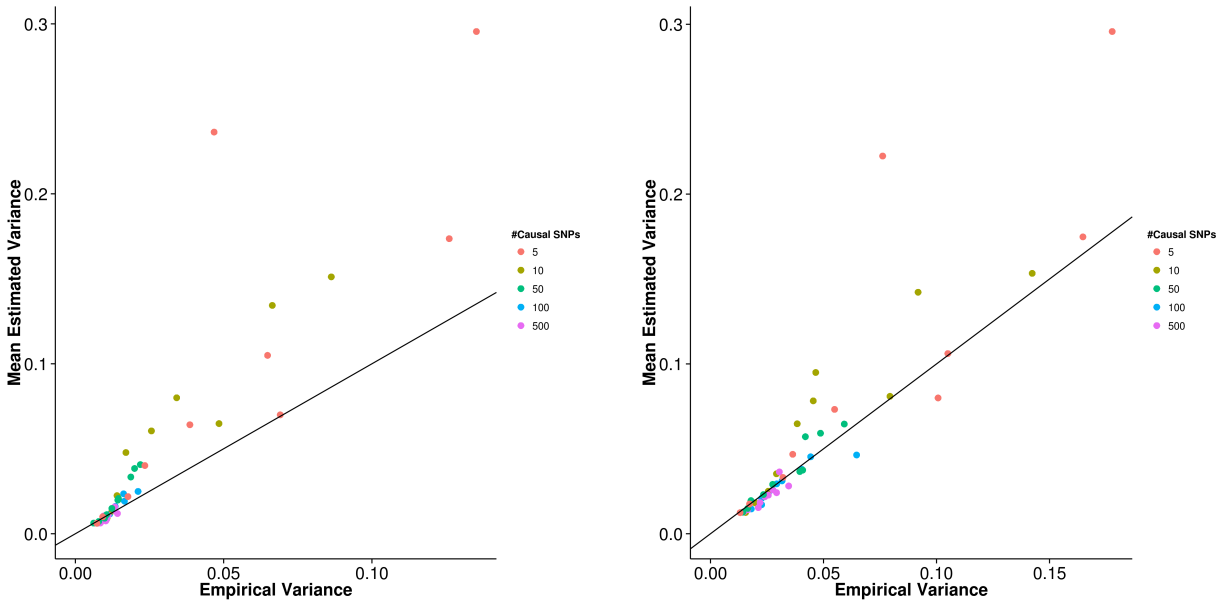
(c) LDSC with fix intercept

(d) LDSC with intercept estimation

Figure 2.3: 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 2.4: 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.

is small.

Second, the empirical variance of the estimates is another important indicator of the performance of the algorithms. It is clear that empirical variance of the estimates from LDSC are sensitive to the number of causal SNPs (figs. 2.3c and 2.3d). When the number of causal SNPs decreases, the variance of the estimates increases, as reported by **Bulik-Sullivan2015** Moreover, consistent with the results from **Bulik-Sullivan2015** the intercept estimation increases the variance of the estimates from LDSC. Similarly, the variance of the estimates form GCTA also increases when the number of causal SNPs decreases (fig. 2.3b), despite it has the lowest variance in its estimates. On the other hand, estimates from SHREK are relatively insensitive to the number of causal SNPs.

Finally, it is important for the algorithms to be able to estimate the standard error of its estimates. Therefore, the estimated variance is compared with the empirical variance of the estimates. It is observed that with a large number of causal SNPs, GCTA can accurately estimates its variance (fig. 2.4b). However, when the number of causal SNPs is small, GCTA underestimates the variance of its

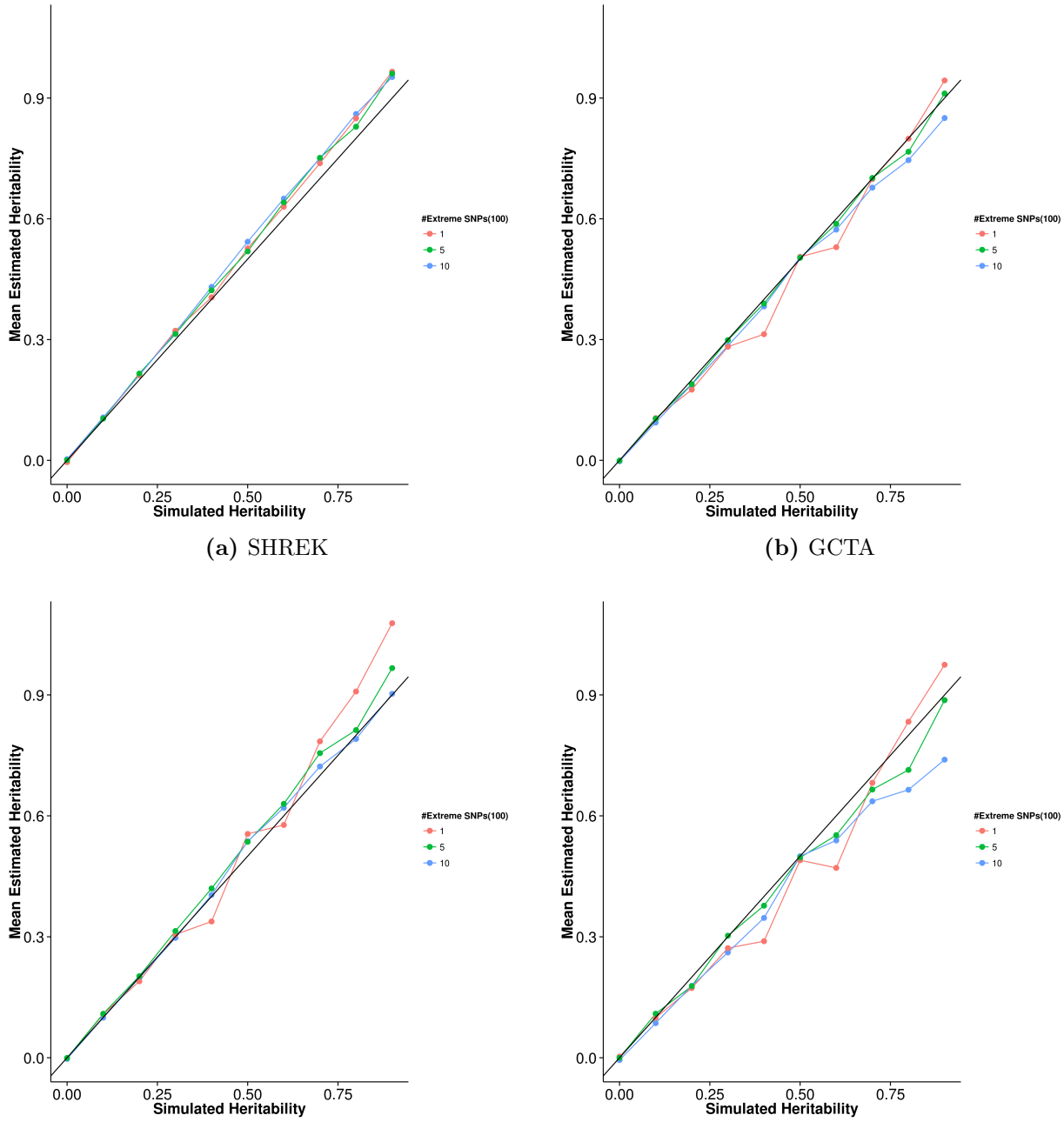
estimates. On the other hand, SHREK consistently underestimate the variance of its estimates (fig. 2.4a). But when compared to LDSC, the magnitude of bias of the variance estimated from SHREK is much smaller. LDSC tends to overestimate its variance (figs. 2.4c and 2.4d), and only when the intercept estimation was performed can LDSC has a better estimates of the variance when the number of causal SNPs is large. Although the bias of the variance estimated from SHREK is much smaller than LDSC, its variance estimate tends to be under-conservative which might misleading. Further development might be required for SHREK to provide a better estimate of standard error.

Taking into account of the bias and variance of the estimates, GCTA has the best overall performance. On the other hand, when the number of causal SNPs is small, SHREK has a better performance when compared to LDSC, whereas LDSC performs better under polygenic condition. It is also observed that estimates from SHREK is the least sensitive to changes in the genetic architecture among the algorithms tested (table 2.1).

2.3.2.2 Quantitative Trait Simulation with Extreme Effect Size

Sometimes, it is possible for a trait to have a small portion of causal variants with much larger effect size when compared to other causal variants. To investigate how LDSC and SHREK perform in such scenarios, simulations were performed with trait that has 100 causal SNPs where 1,5 or 10 of those SNP(s) has a large effect.

The overall performance of the algorithms are similar to the results observed in the quantitative trait simulation (fig. 2.5). However, when 1 of the causal SNPs was simulated with large effect, the mean estimates from LDSC and GCTA fluctuate (figs. 2.5b to 2.5d). The same fluctuation is not observed in SHREK (fig. 2.5a). Similarly, the empirical variance of the estimates (fig. 2.6) from GCTA and LDSC increase and fluctuate when only 1 of the causal SNPs was simulated



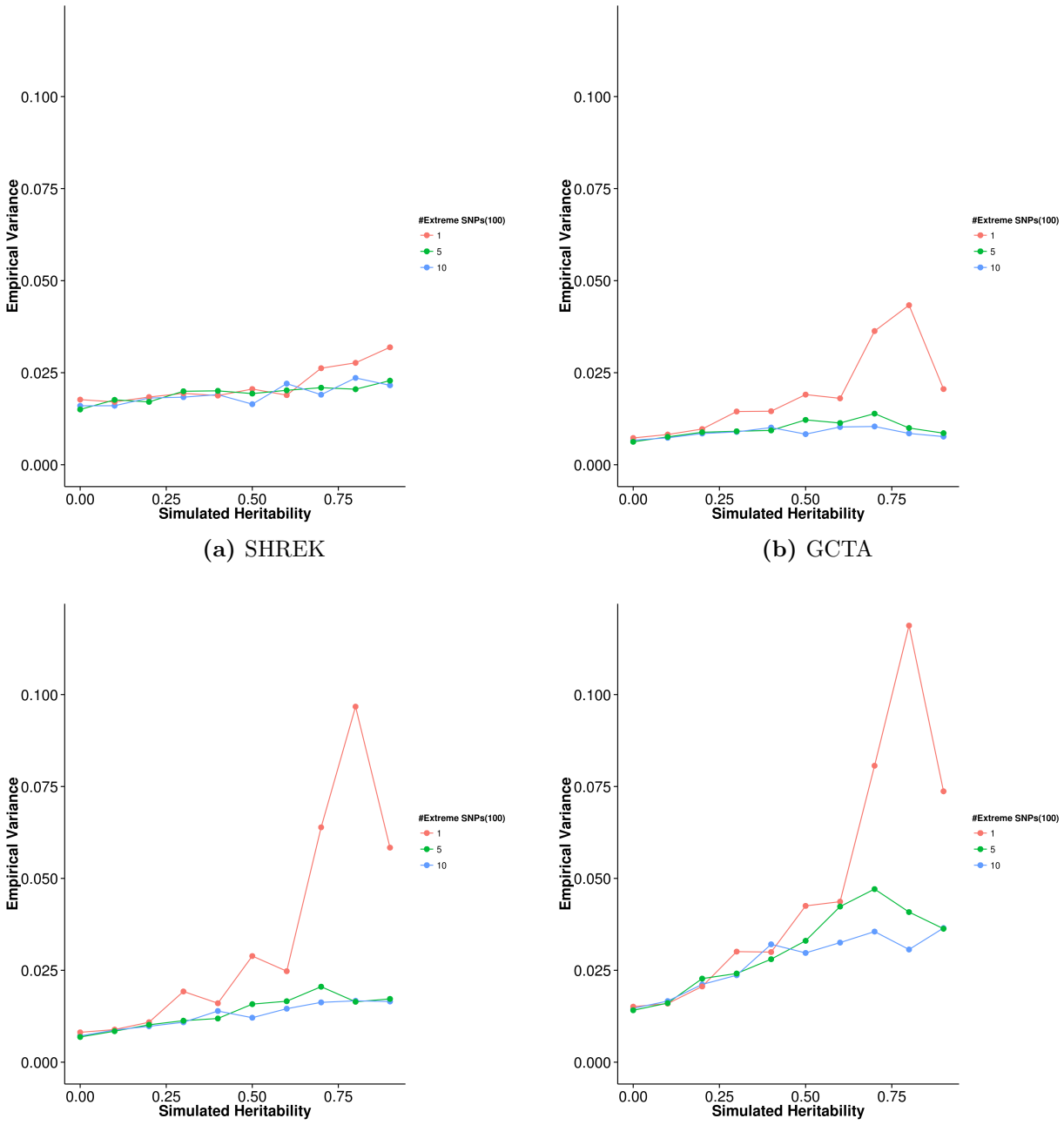
(a) SHREK

(b) GCTA

(c) LDSC with fix intercept

(d) LDSC with intercept estimation

Figure 2.5: 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.



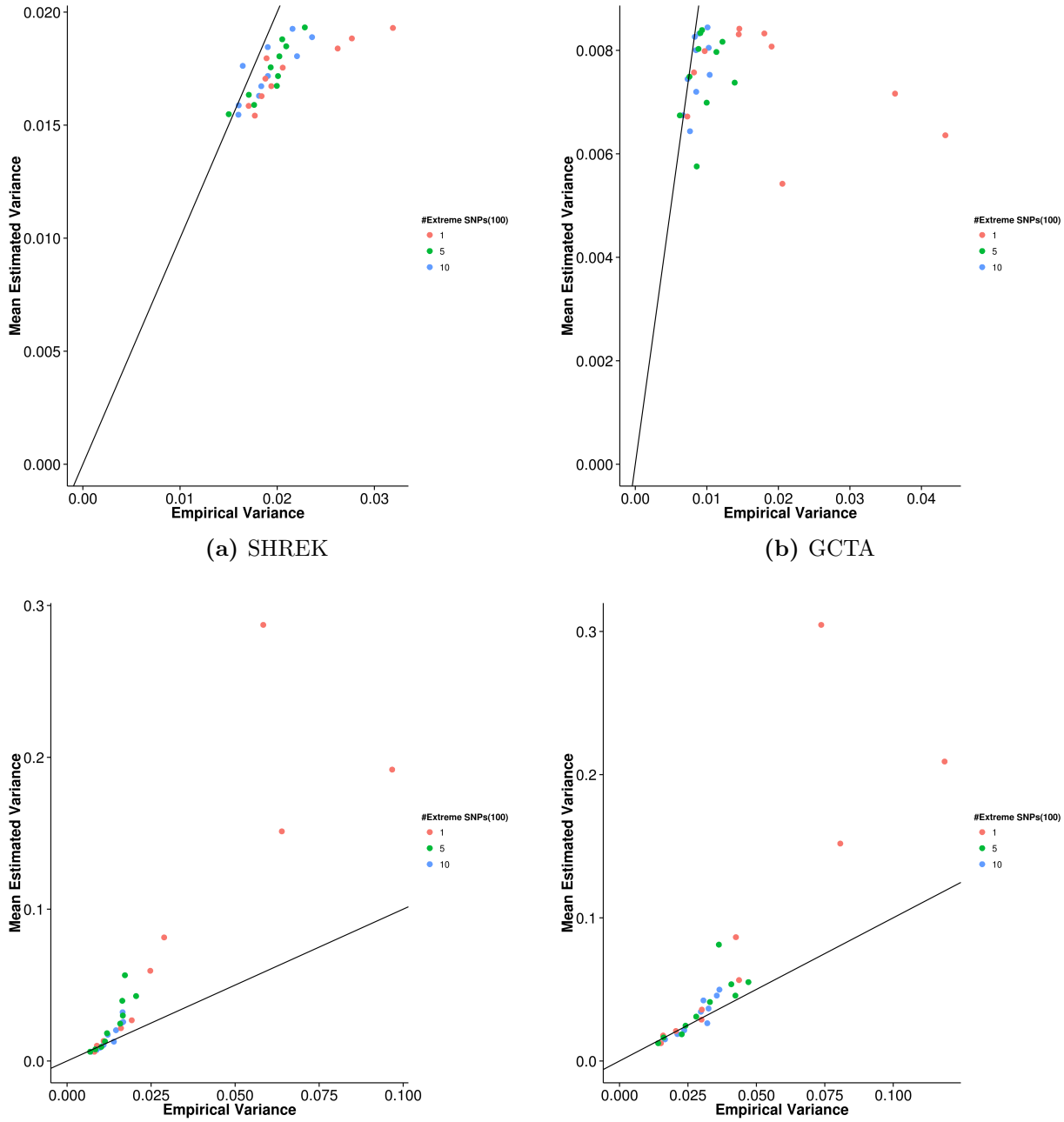
(a) SHREK

(b) GCTA

(c) LDSC with fix intercept

(d) LDSC with intercept estimation

Figure 2.6: 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 2.7: 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. When comparing the performance of SHREK and LDSC, it is observed that LDSC performs better unless only 1 of the causal SNP has a large effect size. However, it is also observed that the performance of SHREK is robust to the change in number of SNPs with extreme effect size.

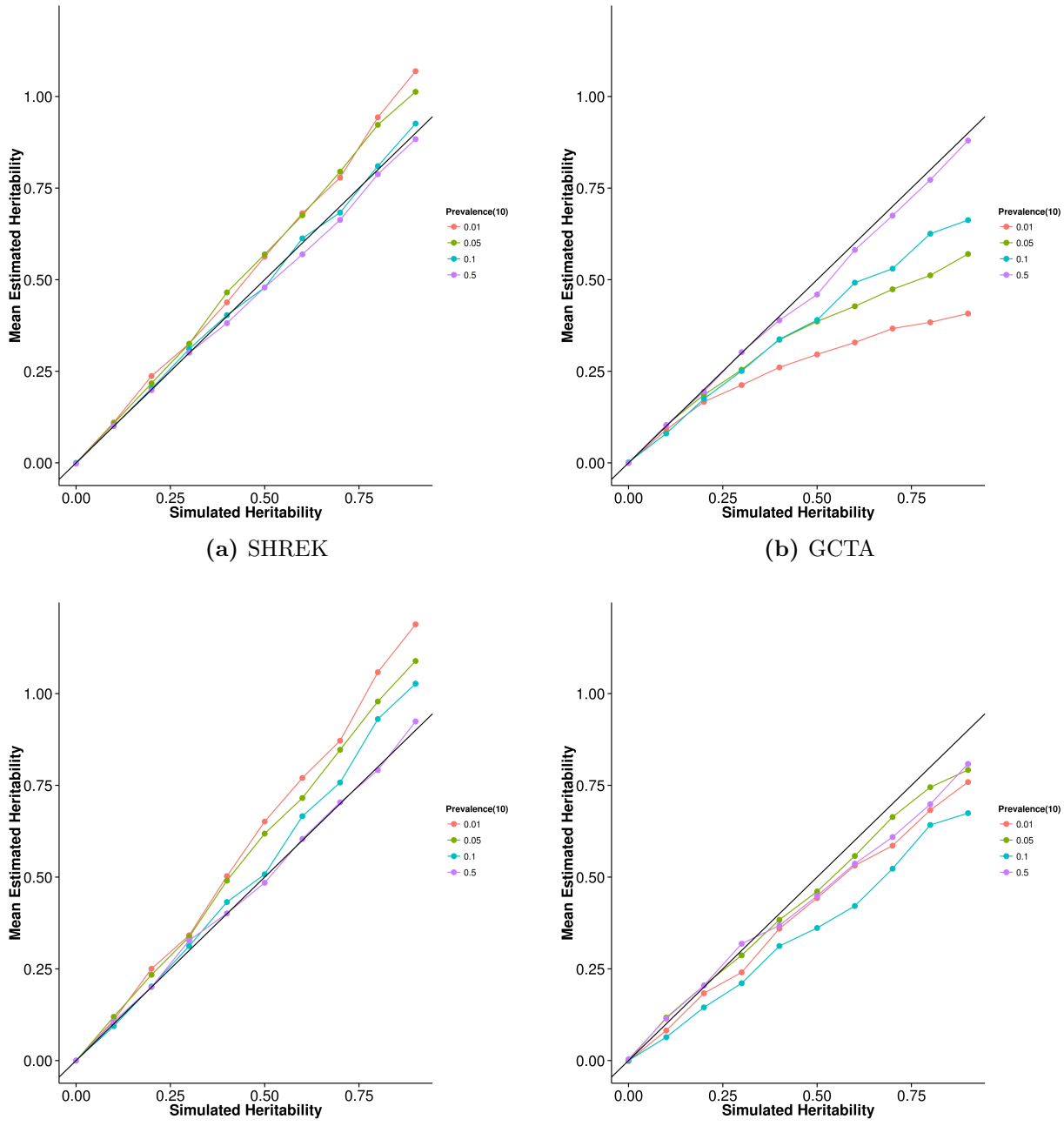
with large effect. Again, the estimates from SHREK are robust to change in number of SNP with large effect size.

When inspecting the variance estimation, it is observed that both SHREK and GCTA underestimate their empirical variance. As the number of SNP(s) with large effect size decreases, the magnitude of bias increases. On the other hand, it is observed that LDSC tends to overestimate its empirical variance. When the intercept is fixed, the estimated variance from LDSC can be as much as 3 fold larger than the empirical variance when only 1 of the causal SNP with large effect size was simulated.

To conclude, GCTA has the best performance among the algorithms tested (table 2.2). However, in the case where the individual genotypes are not available, SHREK has a better performance when compared to LDSC when only 1 of the causal SNPs carries a large effect size. Moreover, it is also observed that SHREK is more robust to change in number of SNPs with large effect size.

2.3.2.3 Binary Trait Simulation

To estimate the SNP-heritability for binary traits, it is important to model the diseases status under a liability threshold model, which require knowledge of the population prevalence of the disease. Due to the important influence of the popula-



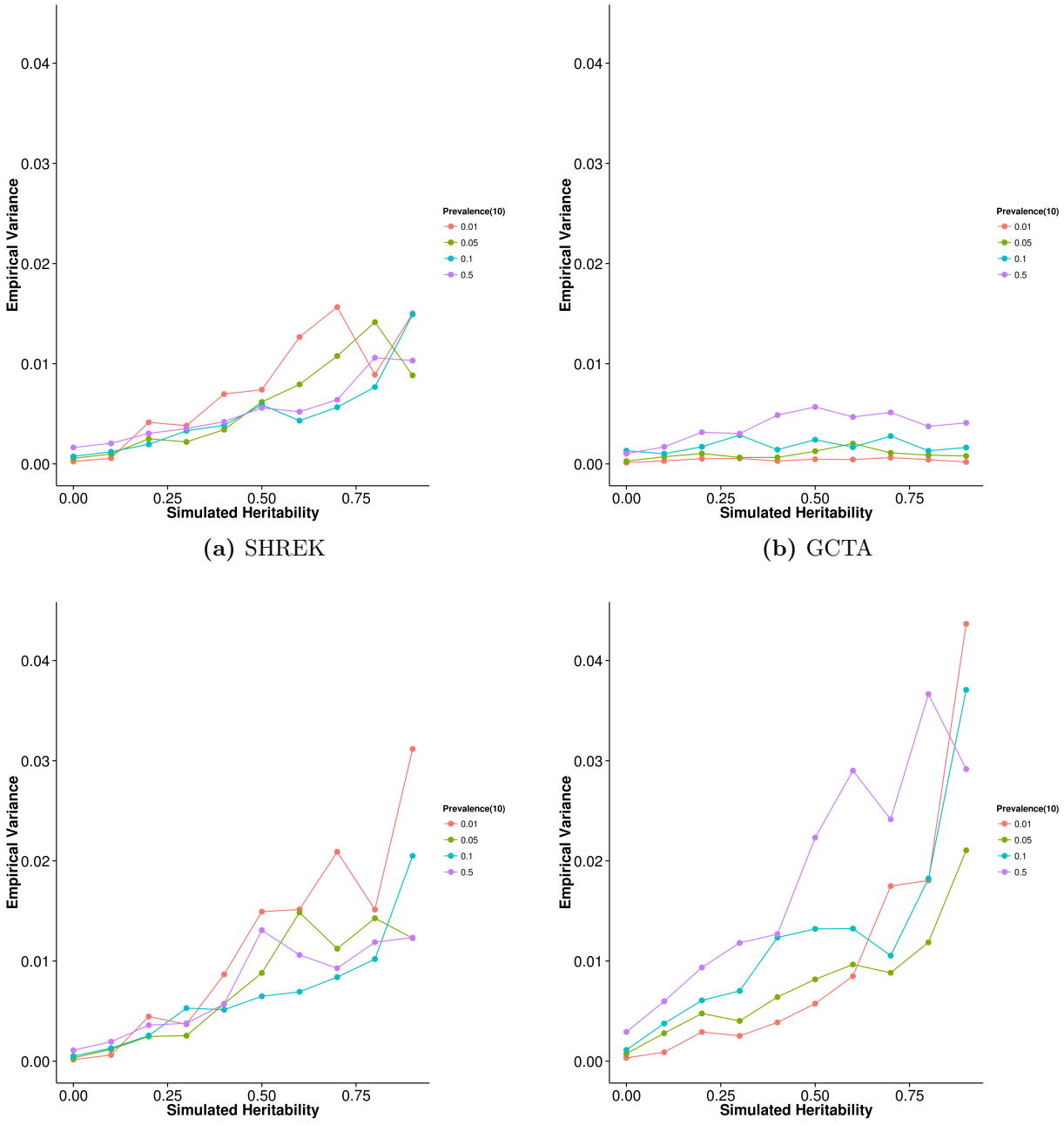
(a) SHREK

(b) GCTA

(c) LDSC with fix intercept

(d) LDSC with intercept estimation

Figure 2.8: Mean of results from binary trait simulation with random effect size simulation with 10 causal SNPs. The performance of GCTA was as suggested by **Golan2014** 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.



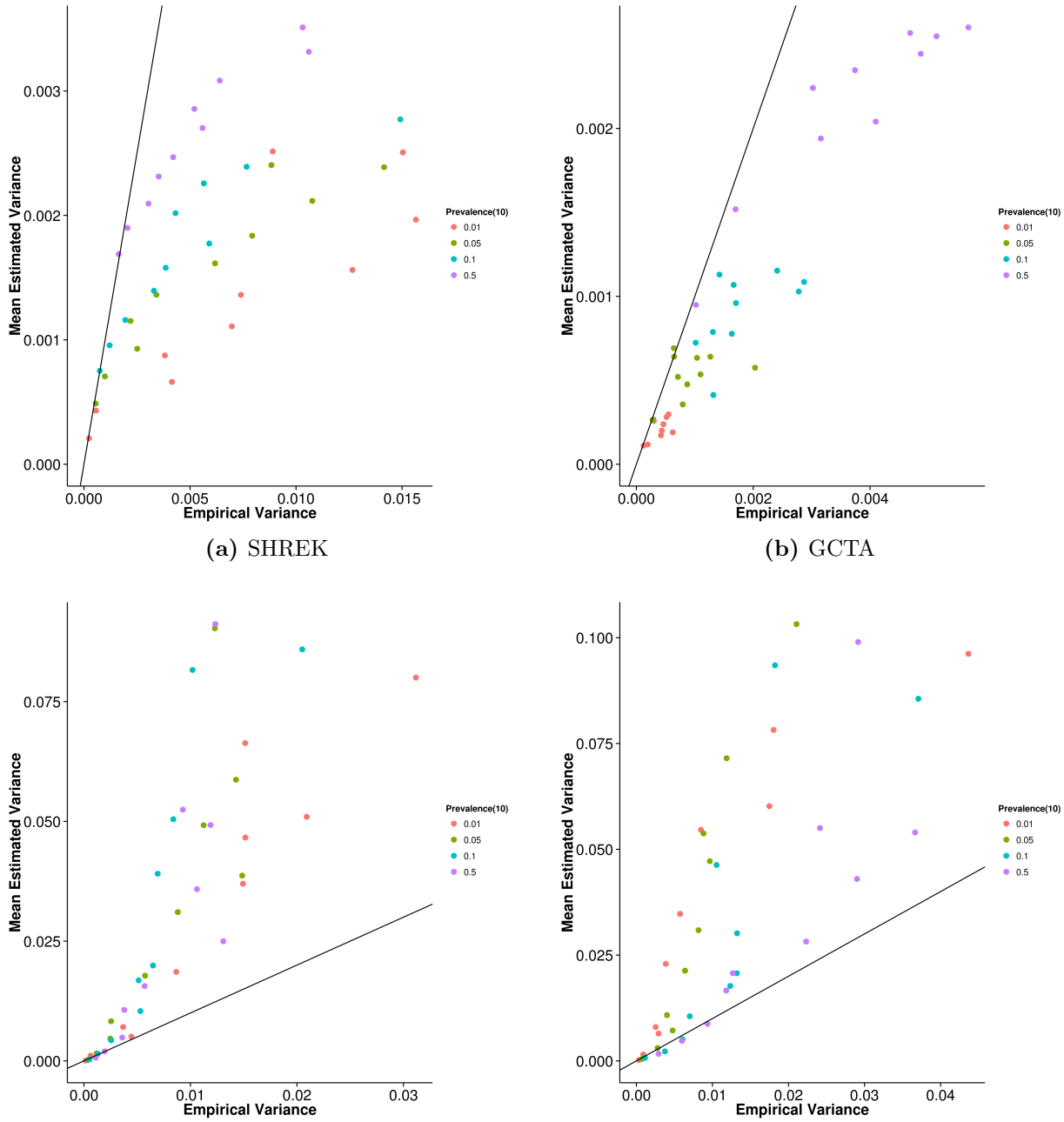
(a) SHREK

(b) GCTA

(c) LDSC with fix intercept

(d) LDSC with intercept estimation

Figure 2.9: Variance of results from binary trait 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 binary trait 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 2.10: Estimated variance of results from binary trait 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.

tion prevalence to the heritability, traits with different population prevalence were simulated such that the performance of LDSC and SHREK can be assessed.

When only 10 causal SNPs were simulated, it is clear that the population prevalence has a significant impact to the performance of the algorithms (fig. 2.8). Of all the algorithms tested, GCTA is the most affected by the population prevalence (fig. 2.8b), where the estimates are generally underestimated. As the population prevalence decreases, the magnitude of bias increases, as reported by **Golan2014** On the other hand, the estimates generated by LDSC with fixed intercept and SHREK are upwardly biased with the magnitude of bias increases as the population prevalence decreases. Surprisingly, when intercept estimation was performed, a downward bias is observed in the estimates from LDSC. The magnitude of bias is also relatively smaller when compared to LDSC with fixed intercept. The same pattern are observed when different number of causal SNPs were simulated (figs. 2.16, 2.19 and 2.22).

Of all the algorithms, GCTA has the smallest average empirical variance (fig. 2.9b) and LDSC with intercept estimation has the largest empirical variance. On the other hand, it is observed that the estimates from SHREK (fig. 2.9a) and LDSC (fig. 2.9c) with fixed intercept have similar empirical variance. As the number of causal SNPs increases, the empirical variance of all algorithms decreases (figs. 2.17, 2.20 and 2.23) similar to the results from the quantitative trait simulation.

It is observed that SHREK consistently underestimate its empirical variance where the magnitude of bias increases as population prevalence decreases (fig. 2.10a). On the other hand, GCTA can provide a more accurate estimation for its empirical variance, only moderately underestimated the variance (fig. 2.10b). Again, it is observed that LDSC consistently overestimate its empirical variance (fig. 2.10). However, as the number of causal SNPs increases, the magnitude of

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 Binary Trait simulation. Algorithm with the best performance under each condition were **bold-ed**. Of all the algorithms, SHREK has the best average performance. It is observed that as the number of causal SNPs increases, the MSE tends to decrease for all algorithms, similar to the results from quantitative trait simulation.

bias observed in the estimation of variance decreases for LDSC (figs. 2.18, 2.21 and 2.24). When 500 causal SNPs were simulated, LDSC can provide a relatively accurate estimates of its empirical variance (fig. 2.24c).

Overall, SHREK has the best average performance of all the algorithm tested (table 2.3). Interestingly, although no confounding factors were simulated, it is observed that LDSC with intercept estimation has a better performance than LDSC with fixed intercept when the prevalence is small. Therefore, it is possible for the intercept estimation to help correcting for some of bias introduced by case control sampling.

It is noted that when compared to the quantitative trait simulation, a smaller number of SNPs and larger sample size (2,000 samples with 1,000 cases and 1,000 controls) were simulated. Thus, the results from binary trait simulations are

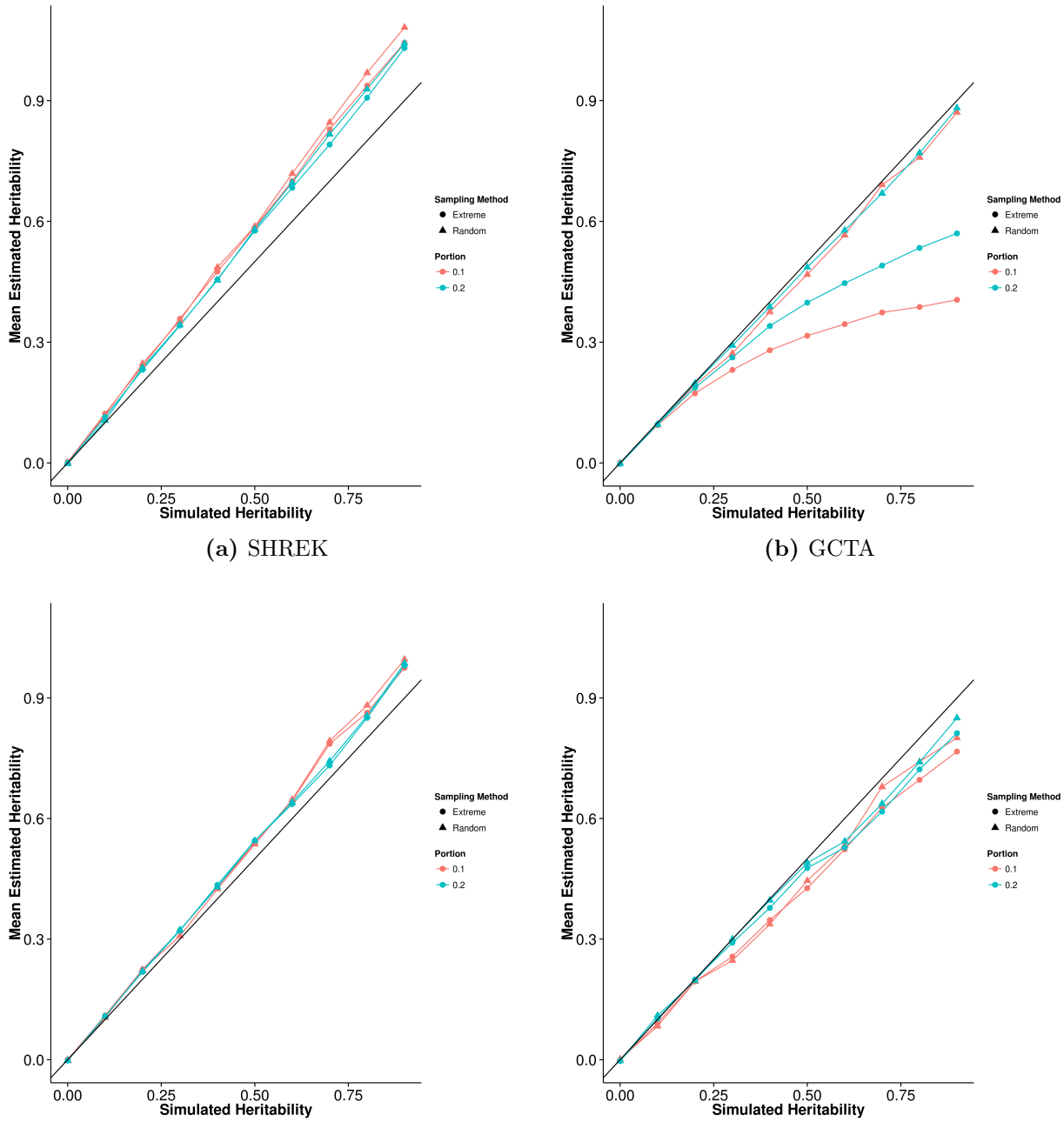
not directly comparable to the results from the quantitative trait simulations.

2.3.2.4 Extreme Phenotype Simulation

By using appropriate sampling strategy, such as extreme phenotype sampling (**Peloso2015**), one can increase the power of the association study. However, it is unclear how the extreme phenotype sampling will affect the performance of the SNP heritability estimation. Therefore, simulations were performed to investigate the effect of extreme phenotype sampling on SNP heritability estimation compared to random sampling approach.

It is observed that when the extreme phenotype sampling was performed, the estimates from GCTA biased downward in pattern similar to what was observed in the binary trait simulation (fig. 2.11b). On the other hand, estimates from SHREK and LDSC with fixed intercepts are slightly inflated whereas LDSC with intercept estimation slightly underestimated the SNP heritability (fig. 2.11).

When comparing the empirical variance, the random sampling consistently results in a larger empirical variance in the estimates of the algorithms (table 2.4). It is observed that when random sampling were performed, the resulting empirical variance from the algorithms are similar to the results in the quantitative trait simulation. However, there is a large discrepancy in the estimated variance of LDSC and GCTA, where there can be as much as a tenfold difference (fig. 2.13). On the other hand, the estimated variance of SHREK is unaffected. It is unclear what induces the inflation and further investigations are therefore required.



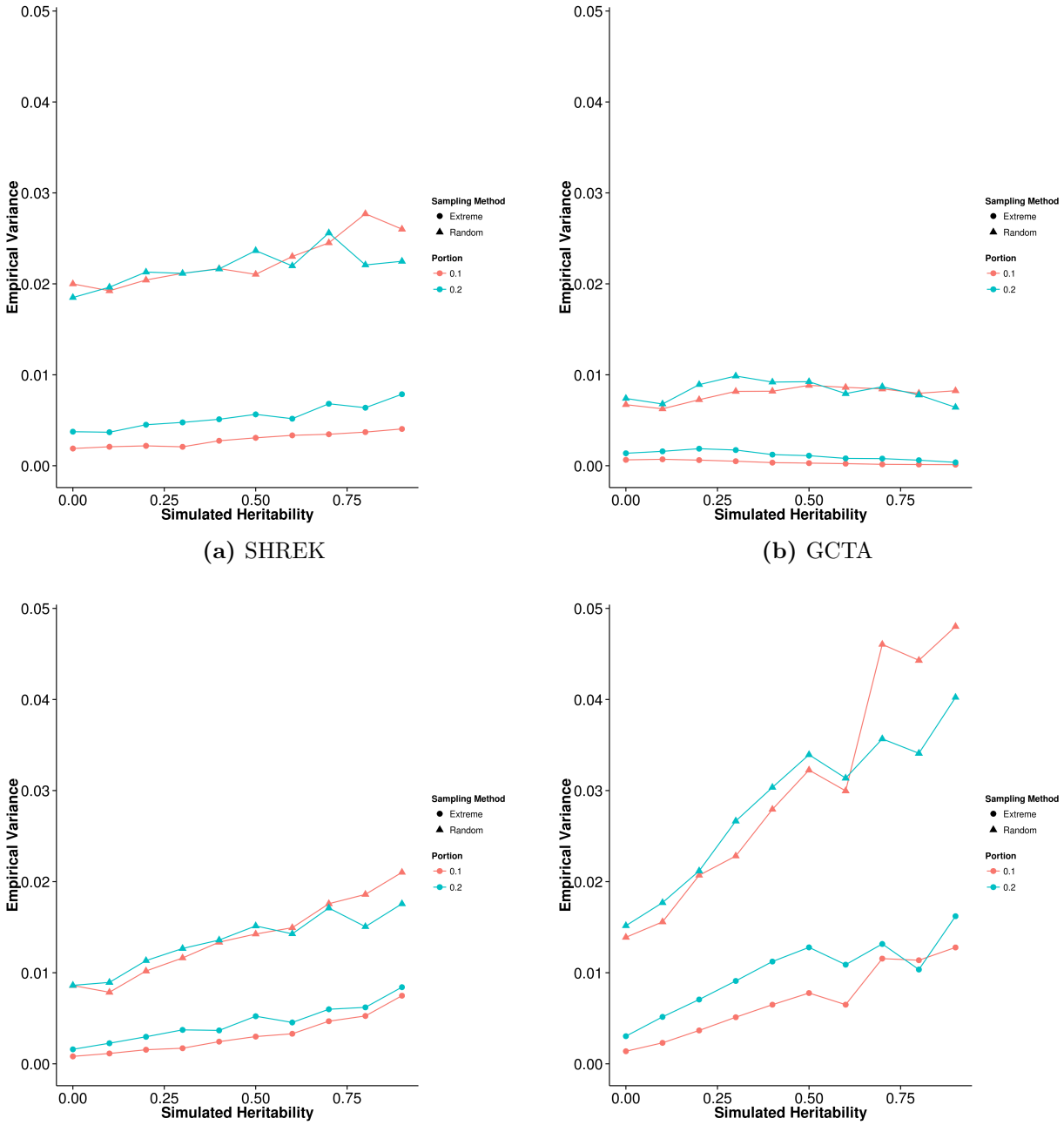
(a) SHREK

(b) GCTA

(c) LDSC with fix intercept

(d) LDSC with intercept estimation

Figure 2.11: 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.



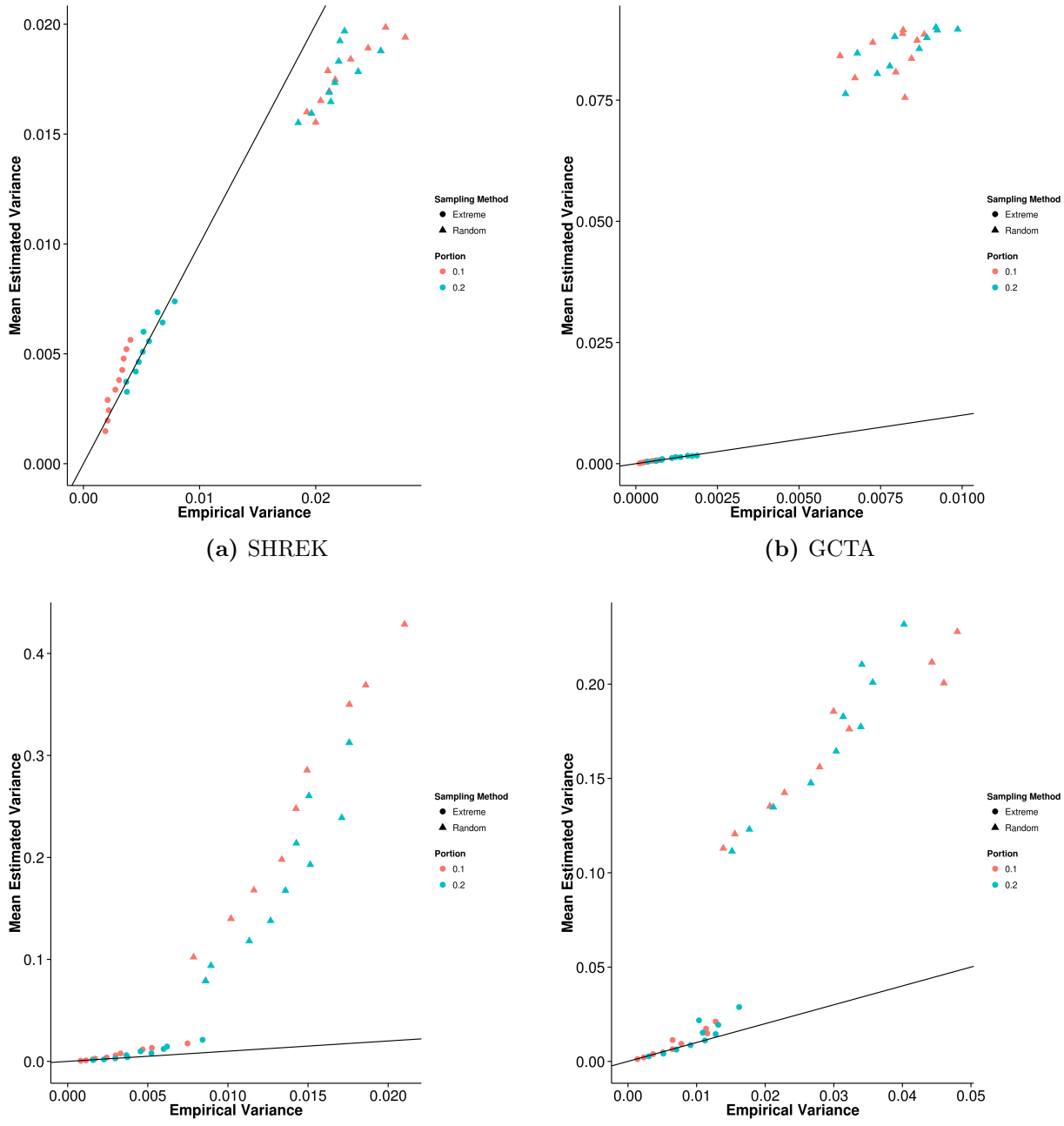
(a) SHREK

(b) GCTA

(c) LDSC with fix intercept

(d) LDSC with intercept estimation

Figure 2.12: 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 2.13: 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.

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: Comparing the MSE of extreme phenotype sampling (Extreme) and random sampling (Rand). With the exception of GCTA, the extreme phenotype sampling will results in a smaller MSE given the same amount of samples.

2.3.3 Application to Real Data

The main purpose of the current chapter is to test whether if the SNP-heritability of schizophrenia estimated by **Bulik-Sullivan2015** is correct. Therefore, we estimated the heritability of schizophrenia, together with major depression disorder and bipolar with SHREK and LDSC. Surprisingly, the estimated SNP-heritability for all disorder investigated are very different from the estimated provided in the **Bulik-Sullivan2015** paper (table 2.5). Contrary to the previous estimate, the common SNPs accounts for no more than 20% of the heritability of schizophrenia, which suggest that other genetic factors such as rare variants and epigenetic changes might also contribute to the heritability of schizophrenia.

	Major Depression Disorder	Bipolar	Schizophrenia
SHREK	0.252 (0.0273)	0.308 (0.0167)	0.185 (0.00450)
LDSC	0.232 (0.0217)	0.265 (0.0152)	0.198 (0.0057)
LDSC-In	0.154 (0.033)	0.181 (0.0203)	0.135 (0.0072)
Bulik-Sullivan2015	0.409 (0.033)	0.531 (0.022)	0.555 (0.008)

Table 2.5: Heritability estimated for PGC data sets. The heritability estimated from LDSC when intercept estimation was performed (LDSC-In) are lower than the estimates from SHREK and LDSC with fixed intercept. As the intercept estimation was used for the correction of confounding effects such as population stratifications or cryptic relatedness, the larger estimates from SHREK and LDSC might be a result of the confounding effects.

2.4 Discussion

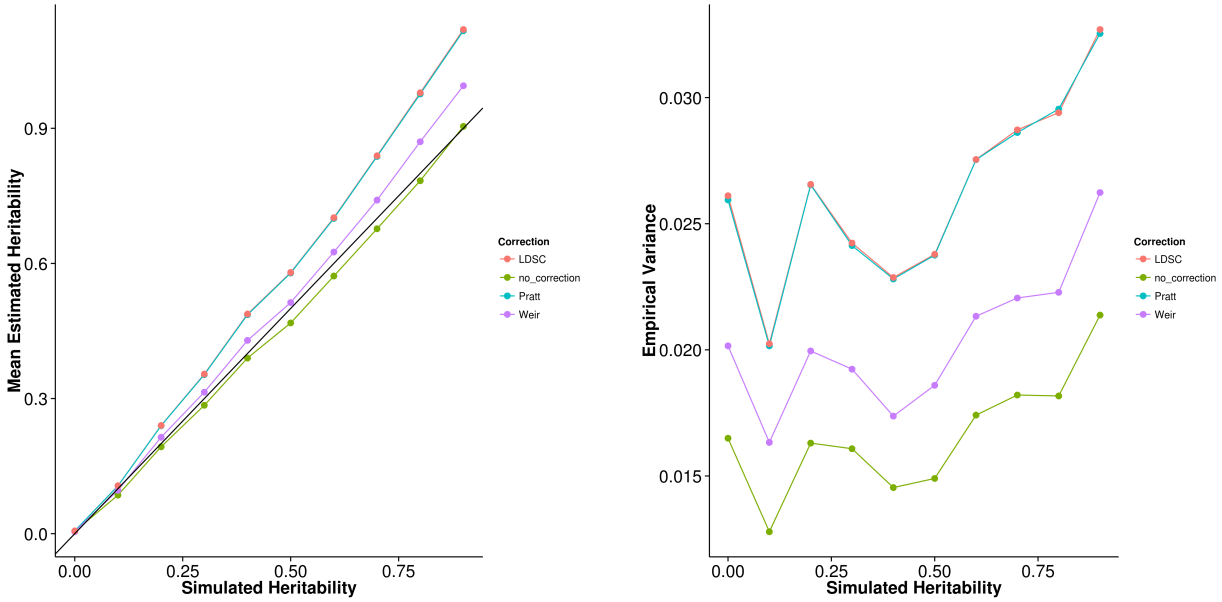
Schizophrenia is a devastating disorder which is of our greatest interest to understand its disease etiology. Recently, in the paper by **Bulik-Sullivan2015** the authors estimated that the SNP-heritability of schizophrenia accounts for majority of the heritability of the disease. However, considering that other genetic variations such as rare mutations and CNV are also associated with schizophrenia, together with the possible presence of the gene-environment interaction, the SNP-heritability of schizophrenia estimated by **Bulik-Sullivan2015** might be too high. Therefore, we set off to develop an alternative algorithm, SHREK for the estimation of SNP-heritability without requiring the individual genotypes, such that we can compare the estimates from the two programme and investigate whether if the estimates from **Bulik-Sullivan2015** is correct. Additionally, we also performed a comprehensive simulation to study the performance of both LDSC and SHREK when handling traits with different genetic architectures.

2.4.1 LD Correction

Similar to LDSC, SHREK relies heavily on the LD information for the estimation of the SNP-heritability. Therefore, it is vital to obtain an accurate estimate of the LD where any sampling errors should be adjusted.

When comparing different bias correction algorithms from section 2.2.9, it was observed that the equation from **Weir1980** eq. (2.51) has the best performance. Therefore, it was selected as the default bias correction algorithm.

By applying the LD correction algorithm, we hope to obtain a more accurate estimate. However, in the subsequent simulations, an upward bias is consistently observed in the estimates from SHREK, which resemble to the inflation observed in the LD correction simulation. Considering that a much smaller num-



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

Figure 2.14: Effect of LD correction to Heritability Estimation when 50,000 SNPs were simulated. It is observed that all LD correction algorithms inflate the estimates when large number of SNPs were simulated. Interestingly, least amount of bias is observed when no LD correction was performed.

ber of SNPs were simulated in the LD correction simulation, it is possible that the inflation was caused by an increased number of SNPs in the genome.

It is also observed that although LDSC also relies on the LD information, the same overestimation is not observed. Upon detail inspection of the algorithm of LDSC, it was found that LDSC employed a different LD correction algorithm:

$$\text{LDSC} : \tilde{R}^2 = \hat{R}^2 - \frac{1 - \hat{R}^2}{n - 2} \tag{2.55}$$

To test our hypothesis and also inspect the performance of the LD correction algorithm from LDSC in SHREK, an additional simulation was performed where a larger number of SNPs were simulated (e.g. 50,000 SNPs on chromosome 1).

From fig. 2.14, it is noted that all LD correction algorithms inflate the

estimates from SHREK, whereas only a small downward bias is observed in the estimates when no LD correction was performed. When inspecting the MSE of the estimates, the MSE is the lowest when no LD correction was performed. This results suggest that when the number of SNPs included in the GWAS increases, the LD correction algorithm might start to have a negative impact to the performance of SHREK.

There are multiple possibilities to the problem. First, as SHREK estimates the SNP-heritability by solving the matrix equation eq. (2.21), which requires the computation of the inverse of the LD matrix. Although in theory, the LD correction algorithm can adjust for the sampling errors in the LD, the adjustment was applied in a per element-wise fashion. It is unclear how this per element-wise correction affects the matrix equation nor is there any methods for the correction to be applied in a matrix form. In fact, of all the LD correction algorithms tested, only eq. (2.51) can be expressed in a matrix form (e.g. $\hat{\boldsymbol{R}}_{\boldsymbol{sq}} - \boldsymbol{N}$ where $\boldsymbol{N}$ is a matrix of $\frac{1}{2n}$) and it has the best performance. Therefore, it is possible that the inflation caused by the LD correction algorithm is because they are incompatible with the matrix formula. However, further researches are required to gain a better understanding of how the LD sampling error can be corrected in the matrix formula.

Another possibility reason for the difference between the results form the two simulation might be because of the change in SNP density. In the LD correction simulation in section 2.2.9, the simulation was performed on chromosome 22, where 5,000 SNPs were randomly selected. This correspond to roughly 1400 SNPs in a 1 mb region. Whereas for the current simulation, where 50,000 SNPs were randomly selected from chromosome 1, which roughly correspond to around 2,000 SNPs in a 1 mb region. Although the difference is small, it is possible that the distribution of LD might differ, where more SNPs in higher LD can be observed when the SNPs are denser together. If the LD algorithms only works for smaller LD, then it is possible

that the LD correction might introduce bias into the final estimates. Again, further study are required.

Nonetheless, it seems that the LD sampling bias are introducing a systematic bias to the estimate of SHREK. By introduce a correct LD correction algorithm, we expect that the systematic bias of SHREK can be corrected and will provide an accurate unbiased estimated of SNP-heritability. However, the complexity and difficulties of this is beyond the scope of this thesis and it should be considered as an important direction for further research.

2.4.2 Simulation Results

One of the main purpose of the current study is to understand how different sampling strategies and different genetic architectures affect the performance of LDSC and SHREK. Therefore, a number of simulations were performed.

2.4.2.1 Quantitative Trait Simulation

First, it is observed that GCTA has the best overall performance in estimating the SNP-heritability for quantitative traits. However, it requires individual genotypes for its estimation. Therefore, when the individual genotypes are unavailable, GCTA cannot be performed. Here, GCTA serves as a reference point for SHREK and LDSC, allowing us to assess whether if we have performed the simulation correctly.

When comparing the performance of SHREK and LDSC, it is observed that the performance of LDSC decreases when the number of causal SNPs decreases whereas SHREK remains relatively robust to the change in number of causal SNPs. Our results agree with those in **Bulik-Sullivan2015** where an increased standard error was observed for LDSC when the number of causal SNPs are small.

Given that the main purpose of LDSC is to delineate polygenicity from the confounding factors, it might not be able to handle the oligogenic traits. On the other hand, SHREK makes no assumption to the genetic architecture of the trait. Therefore, this might be the reason why SHREK has a better performance than LDSC when oligogenic traits were provided.

It is also noted that in the current simulation scheme, the intention of varying the number of causal SNPs was to assess the performance of LDSC and SHREK when traits with different genetic architectures are provided. However, this also unintentionally varied the "density" of the causal SNPs, i.e. by changing the number of causal SNPs without adjusting the overall genome size / SNP distribution, the density of the causal SNPs reduces as the number of causal SNPs decrease. Therefore, it is possible that the performance of LDSC is affected by a reduced causal SNPs density instead of the change in polygenicity. Additional simulation will have to be performed to test this hypothesis where instead of varying the number of causal SNPs, one should varies the density of the causal SNPs. So for example, we can simulate 50,000 SNPs from chromosome 1 where 5 of them are causal. In one scenario, the 5 causal SNPs can all be within a 1 mb region whereas in another scenario, the causal SNPs can be evenly distributed among the 50,000 SNPs. If indeed LDSC is affected by the density of the causal SNPs, we would expect LDSC to have a better performance when the causal SNPs are clustered together.

2.4.2.2 Extreme Effect Size

In additional to the quantitative trait simulation, we have also simulated scenario where a small amount of causal SNP(s) accounts for majority of the effect size. In such scenario, SHREK outperforms LDSC only when 1 of the causal SNP was simulated with large effect size. However, upon re-examination of eq. (2.54), which was used for the simulation of the effect sizes in this simulation, it is noted that

an unnecessary upper bound of (h^2) was imposed. Essentially, when the number of "extreme" causal SNPs increases, the effect size of these "extreme" causal SNPs decreases. When more an more "extreme" causal SNPs were included in the simulation, the effect sizes of these causal SNPs were no longer extreme. Therefore, to better examine the effect of causal SNPs with extreme effect, the effect sizes of the causal SNPs should first be simulated using eq. (2.52). Then, for the m "extreme" SNP(s), their effect sizes should be multiplied with a large constant (e.g 100). Therefore, the results of the current simulation does not represent the scenario where a number of causal SNPs for a trait has extreme effect and the simulation should be repeated with the proper effect size simulation before we can draw any conclusion.

2.4.2.3 Binary Trait Simulation

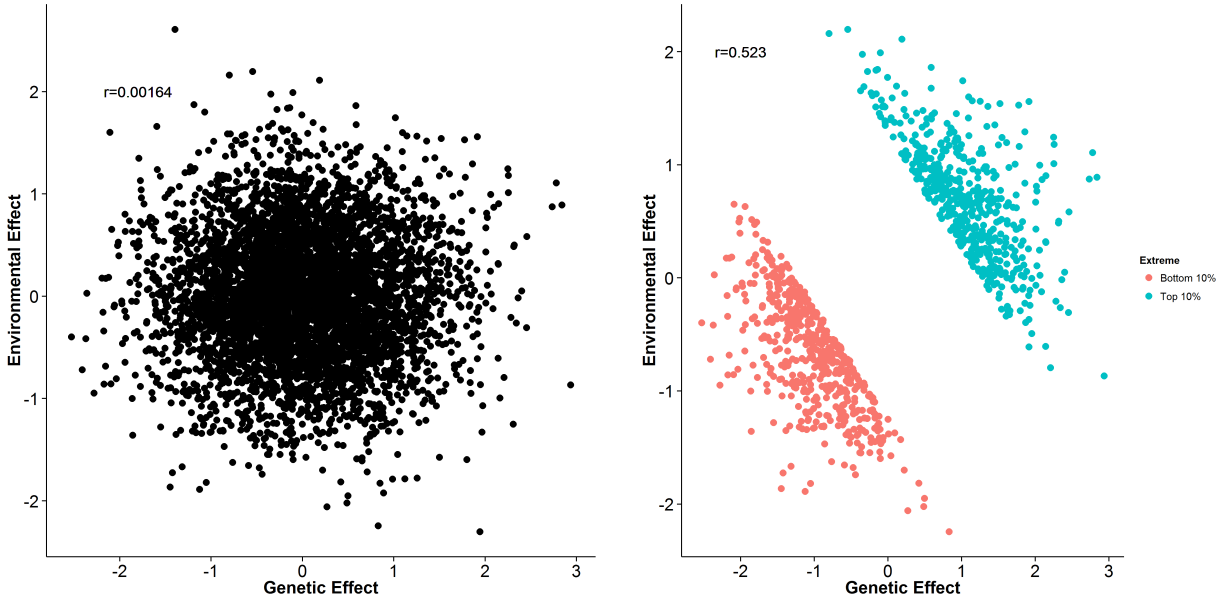
In order to estimate the SNP-heritability for binary traits, it is important to correct for the ascertainment bias. Nevertheless, the correction of ascertainment bias are nontrivial and often introduce bias to the estimates. For example, **Golan2014** observed that GCTA underestimates the heritability explained by common variants for binary 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 (**Golan2014**). According to **Golan2014** there is an oversampling of the cases relative to their prevalence in the population in binary trait studies. 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 (**Golan2014**). This leads to the estimates from GCTA to be strongly downward biased, where the magnitude of bias increases as the population prevalence decreases, when the heritability increases and when the proportion of cases is closer to half.

In our simulation, the same bias is observed for GCTA. Interestingly, it is observed that the bias of the estimates from SHREK and LDSC are also proportional to the population prevalence of the trait. As the population prevalence decreases, the estimates from LDSC and SHREK biased *upwards*. The inflation of the estimates suggest that although the population prevalence also affects the estimates from SHREK and LDSC, the bias might be introduced differently when compared to GCTA. Further investigation are required to understand how the population prevalence influence the estimates from SHREK and LDSC.

On the other hand, a surprising observation is that by applying the intercept estimation, the estimates of LDSC becomes more robust to the change in population prevalence (figs. 2.8d, 2.16d, 2.19d and 2.22d). As no confounding factors were simulated, it was expected that the intercept estimation function should be redundant. However, results suggest that in the estimation of SNP heritability for binary traits, the intercept estimation might be beneficial even when no confounding factors were presented, especially when the population prevalence is small (e.g. < 0.05) table 2.3. Further investigation are required to understand how the intercept estimation improves the performance for binary traits. This might provide insight for the development of a better algorithm for the estimation of the SNP-heritability of binary traits for both LDSC and SHREK.

2.4.2.4 Extreme Phenotype Sampling

When budgets are limited, extreme phenotype sampling might help to increase the power of the association study given the same amount of samples. Compared with the same number of randomly selected individuals, the extreme selection design can increase the power by a factor of $\frac{V_{P'}}{V_P}$ where $V_{P'}$ is variance of the trait of the selected sample and V_P 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



(a) No Extreme sampling

(b) Extreme Sampling

Figure 2.15: Effect of extreme sampling design. Although the genetic and environmental effect were simulated independently, an artificial correlation is observed when extreme phenotype sampling was performed. This lead to a downward bias in the estimates from GCTA (**Golan2014**).

distribution, one can achieve the same power as a study with random sampling design that has 4 times the sample size (**Sham2014**).

Interestingly, in the simulation of extreme phenotype sampling, the estimates from GCTA are also biased downward. The pattern of bias is similar to what was observed in the binary trait simulation. It is observed that when extreme phenotype sampling was performed, an artificial correlation between the genetic and environmental effects was also introduced (fig. 2.15), similar to what was proposed by **Golan2014** Therefore, it is possible that the bias observed in the estimates of GCTA, SHREK when extreme phenotype sampling was performed is the same as that observed in the binary trait simulation.

However, in additional to the bias from the artificial correlation, as the phenotypes from the extreme phenotype sampling was coded in continuous scale instead of 0/1, it might lead to non-normality of residuals in the results of the linear

regression, further complicating the problem. Further investigation are therefore required to allow the algorithms to not only handle the correlated genetic and environmental effects, but also the non-normal residuals in the results of the linear regression of association.

Another important observation in the simulation of extreme phenotype sampling is to allow for the comparison between the random sampling and extreme phenotype sampling strategies. Overall, given the same number of samples, performance of LDSC and SHREK are more than 3 fold better when extreme phenotype sampling was performed, suggesting that extreme phenotype sampling improves not only the power of association studies but also the performance in the estimation of SNP-heritability.

Peculiarly, in the simulation of random sampling, although the empirical variance is the same as what was observed in the quantitative trait simulation, GCTA and LDSC were unable to estimate their empirical variance. The estimated variance from GCTA and LDSC can be more than 10 fold larger than the empirical variance yet the same bias was not observed in the quantitative trait simulation even though all parameters are the same.

It is therefore unclear why a different estimated variance are obtained. One possible explanation is that in the extreme phenotype sampling simulations, a population of samples were first simulated. Samples were than randomly selected from the population. As the normalization of the genotype and phenotypes (normalized according to population, instead of samples) are slightly different from what the quantitative trait simulations, this might be the reason why the two simulations provides a different results. It might be useful for us to repeat the whole simulation to investigate whether if the differnece is indeed due to difference in the normalization of the genotypes/ phenotypes.

Nevertheless, as the sampling were only performed *after* the simulation of

phenotypes, any difference in performance should be a result of different sampling strategies. Thus it is safe to conclude that extreme phenotype sampling can provide more power for not only the association studies, but also the for SNP-heritability estimation given the same amount of samples.

2.4.3 Limitations of the Simulation

In our simulation, we have tried to consider as much parameter as possible yet there are still some parameters that were not tested. For example, in most simulations, only 1,000 samples were simulated (2,000 for binary trait simulation). Although the sample size selected corresponds to the lower quantile of the published GWAS, it is unclear whether if the performance of LDSC and SHREK will linearly improve as the sample size increases. Considering that the sample size of today's GWAS can be as high as 150,000 (e.g. PGC), it is important for us to also test the performance of the algorithms when a big sample size is provided. Unfortunately, to simulate a single data set with 50,000 SNPs and 150,000 samples, 60 gigabyte data will have to be generated and we have not been successful in generating such a large set of data. Therefore, to investigate the effect sample size to the performance of LDSC and SHREK, we can simulate GWAS with a smaller number of SNPs (e.g. 5,000 SNPs on chromosome 22), where we varies the sample size. This should serves as an important follow up to the current study.

Another concern regarding our simulation is that the causal SNPs were always included in the simulated GWAS. In reality, most causal SNPs might be absent from the GWAS and their signals were only detected due to LD. Therefore, it might be useful for us to repeat all of our simulations with all the causal SNPs removed from the final GWAS.

As mentioned in previous sections, when the number of causal variants were

varied, the *density* of the causal variants were unintentionally varied. Therefore, it is uncertain whether if the change in performance was due to the number or the density of the causal variants. A controlled simulation are therefore required to delineate the effect of the number and density of the causal variants to the performance of LDSC and SHREK.

Additionally, in our simulations, we have only simulated SNPs with MAF ≥ 0.05. Although GWAS lack power in detecting SNPs with low MAF, it is still interesting to investigate how the MAF of the GWAS SNPs affect the performance of the algorithms.

2.4.4 SNP-Heritability of Schizophrenia

The main goal of the current study is to investigate whether if the SNP-heritability of schizophrenia estimated by LDSC is accurate. Therefore, we repeated the analysis of **Bulik-Sullivan2015** with SHREK and LDSC.

Surprisingly, all SNP-heritability estimated by LDSC differ significantly to those estimated in the supplementary materials by **Bulik-Sullivan2015** (table 2.5). After communicating with the corresponding author (**Bulik-Sullivan2015c**), it was confirmed that an older implementation of LDSC was used to generate the estimates in the supplementary table. Specifically, in the formula of LDSC

$$\mathrm{E}[\chi^2|l_j] = Nl_j\frac{h^2}{M} + Na + 1 \tag{2.56}$$

$$l_j = \text{LD score of variant } j$$

$$N = \text{Sample Size}$$

$$a = \text{Contribution of confounding biases}$$

$$h^2 = \text{heritability}$$

M was originally defined as the total number of SNPs in the reference panel used

to estimate LD score. However, in the current version of LDSC, M was defined as the number of SNPs with MAF$>$ 5% in the reference panel used to estimate LD score. **Bulik-Sullivan2015** suggested that it is more appropriate based on new data they observed after their original paper was published. From the caption of the supplementary table, it was 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." (**Bulik-Sullivan2015**). This explained the discrepancy between our estimates and the estimates observed in the supplementary table from **Bulik-Sullivan2015**

This result is rather worrying as with a change of M, more than 2 fold difference can be observed between the estimates. As it is uncertain what the "correct" definition of M for LDSC should be, there is also a large uncertainty in the estimated SNP-heritability from LDSC. On the other hand, SHREK doesn't suffer from this ambiguity as one will always only include SNPs found on both the reference panel and the GWAS chip when estimating the SNP-heritability. Therefore it might be more reliable to use SHREK for the estimation of SNP-heritability when compared to LDSC unless one can be certain the definition of M is correct for LDSC.

That said, SHREK is not without its own problem. Because of its reliance on the tSVD for solving the equation, it is generally slower when compared to LDSC. When the SNP density is high $> 6,000$ SNPs in a 3 mb region), SHREK will start to require a significant amount of time for the estimation of SNP-heritability because of the O(n^3) time complexity of the algorithm. So for example, when applying SHREK to the PGC schizophrenia GWAS data, because of the exceptionally high density of the SNPs, we have to lower the block size in order to complete the estimation. This will lead to a small inflation in the estimates of SHREK. To handle this problem, we have updated the implementation of SHREK to utilize the Armadillo library

(**Sanderson2010**). Armadillo library can be 3 times faster than EIGEN C++ library by utilizing the multi-threading high performance OpenBLAS library for the computation of SVD. We also remove SNPs with low imputation info score and SNPs in perfect LD with each other in the new implementation. This is expected to not only increase the speed of SHREK, but also allow for a more accurate and robust estimate.

Another observation is that the estimates from SHREK and LDSC with fixed intercept are generally higher than those provided by LDSC with intercept estimation (LDSC-In). Based on the results from our binary trait simulation, for traits with a population prevalence less than 0.5, the estimates from LDSC-In biased downward whereas estimates from LDSC and SHREK were biased upward. Therefore estimates from LDSC-In and SHREK can be served as the lower and upper bound of the true SNP-heritability respectively.

Although our simulations do provide some insight as to how to interpret the estimates, it is important to remember that no confounding effects were simulated. In the presence of confounding factors, spurious associations might be observed, leading to an inflated summary statistics (**Zheng2006**), which might inflates the estimates from SHREK and LDSC. Therefore it is possible that the SNP-heritability of schizophrenia estimated by SHREK and LDSC are overestimated.

For LDSC, **Bulik-Sullivan2015** argues that by estimating the intercept of the LD score regression model, one can separate the confounding effects from the polygenicity. Most importantly, **Bulik-Sullivan2015** hypothesized that by adjusting for the principal components in the association testing, the remaining population stratification might be small. They further hypothesized that after the adjustment of the population stratification in the association of the non-admixed population, the LD score between the populations should be approximately equal, which leads to eq. (1.16). Thus, in theory, by performing LDSC with the intercept estimation

function, one can estimate the SNP-heritability of a trait without worrying about the confounding factors.

On the other hand, in the derivation of SHREK, it was assumed that the correlation between a SNP and the phenotype is close to 0 (eq. (2.11)), which leads to the -1 in eq. (2.18). Therefore, if the summary statistics are inflated by confounding factors, the estimates from SHREK might be inflated. However, we argue if the association were performed with the population stratification and other covariates adjusted, the resulting summary statistics might be only mildly affected by the confounding effects. Also, if the presence of confounding factors leads to difference of LD score between the samples, which violates the assumption of LDSC, then both SHREK and LDSC might suffer. Given the superior performance of SHREK in the estimation of the SNP-heritability for binary traits when compared to LDSC, it might still be preferable to use SHREK for the estimation of SNP-heritability for binary traits. In order to investigate how different confounding factors affect the performance of SHREK and LDSC, additional simulations must be performed.

Additionally, when admixed population is used for the estimation of SNP-heritability, we hypothesized that both LDSC and SHREK might be underperformed. This is because when working with admixed population, the LD within the samples might not be accurately represented by the reference panel. Because SHREK and LDSC both relies heavily on the LD information to estimate the SNP-heritability, it might not be possible to accurately estimate the SNP-heritability for admixed population unless there is a reference panel that is representative of the sample population.

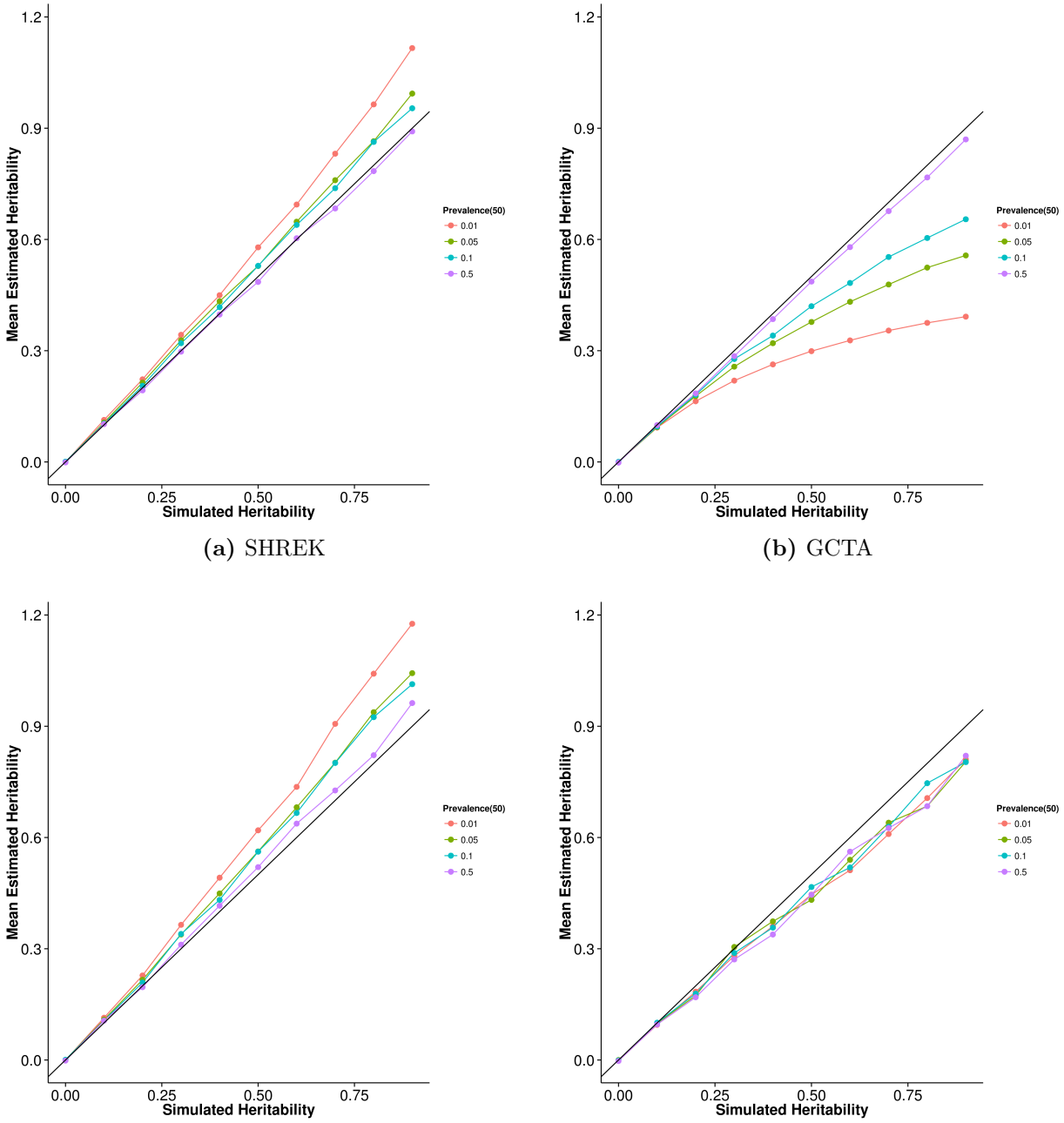
Given the new estimates from SHREK and LDSC, the SNP-heritability of schizophrenia can be no more than 20%. When compared to the heritability estimated from twin studies, $40\% \sim 60\%$ of the heritability remains unaccounted for. One possible source of the "missing heritability" might be residing on the sex

chromosome. When estimating the SNP-heritability, SNPs on the sex chromosomes were discard. This is because when performing the association analysis on the sex chromosomes, different association test can be performed, which might result in a different summary statistics (**Wong2014**) (Supplementary materials). Therefore, it is difficult to estimate the SNP-heritability from the sex chromosomes with on the summary statistics. Further investigation are required before LDSC and SHREK can be applied to the estimation of the contribution of SNPs resides on the sex chromosomes.

The relatively small SNP-heritability of schizophrenia might also suggest that other genetic variants such as the rare variants and epigenetic changes might be another possible source of heritability. However, as LD calculated from rare variants usually have a large standard error, SHREK and LDSC might not perform as well when handling rare variants. For example, **Bulik-Sullivan2015** reported that when all causal variants of a trait are rare (MAF $< 1\%$), LDSC will often generate a negative slope, with the intercept exceeding the mean χ^2 statistic. It is unclear whether if SHREK and GCTA will also suffer in the same way when most causal variants are rare. Therefore, it might be beneficial to perform additional simulations to study the effect of rare causal variants to the performance of the tools in the estimation of SNP-heritability.

To conclude, as the SNP-heritability of schizophrenia less than 20%, it might be better for the field to invest more time into the identification of rare variants that are associated with schizophrenia using methods such as the next generation sequencing technology. Additionally, it might also be interesting to investigate the contribution of epigenetic factors, such as methylation, to the etiology of schizophrenia. Only then can we gain better understanding of how genetics contribute to the risk of schizophrenia, and hopefully allow us to develop better treatment for schizophrenia.

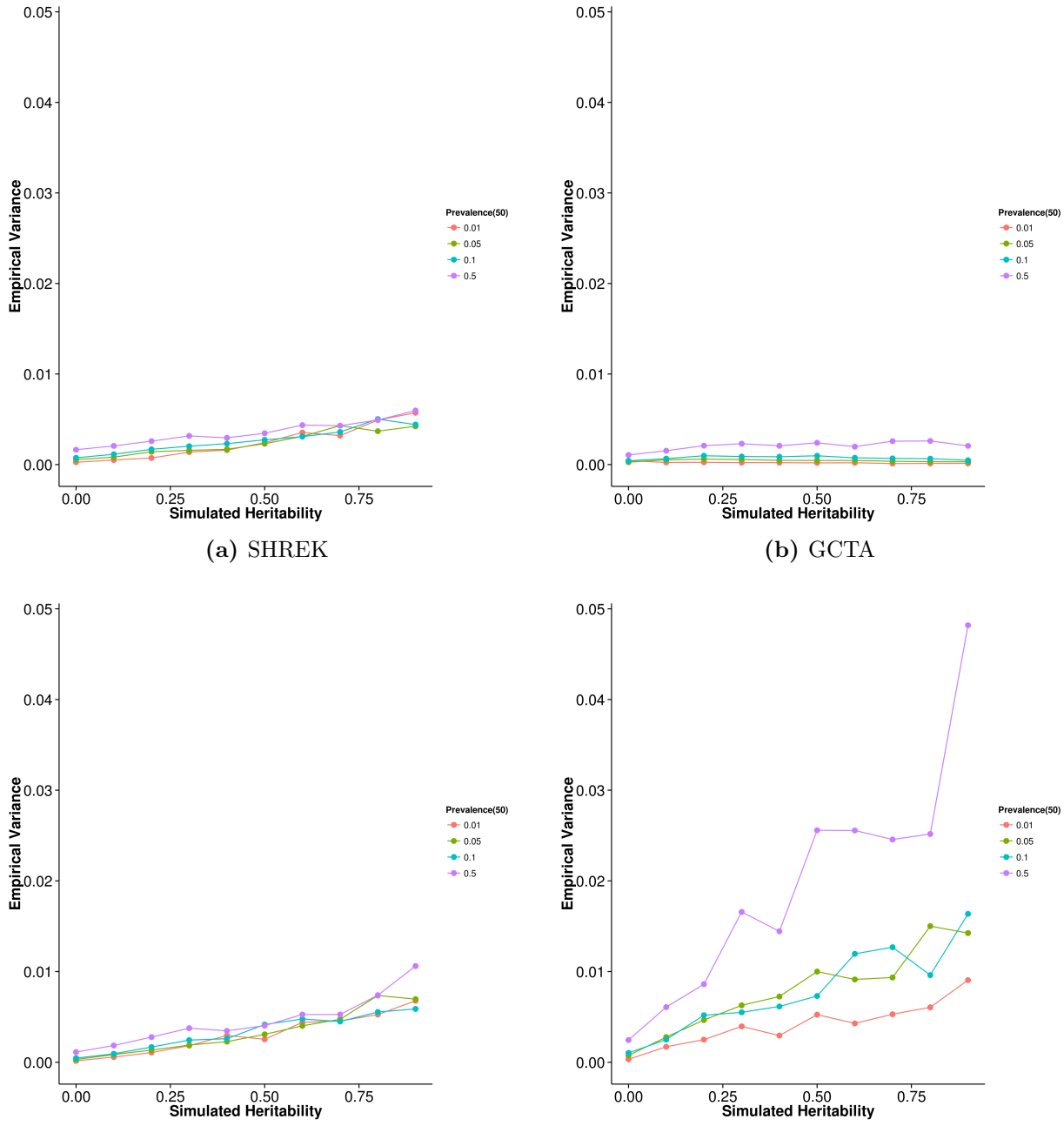
2.5 Supplementary



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

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

Figure 2.16: 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.



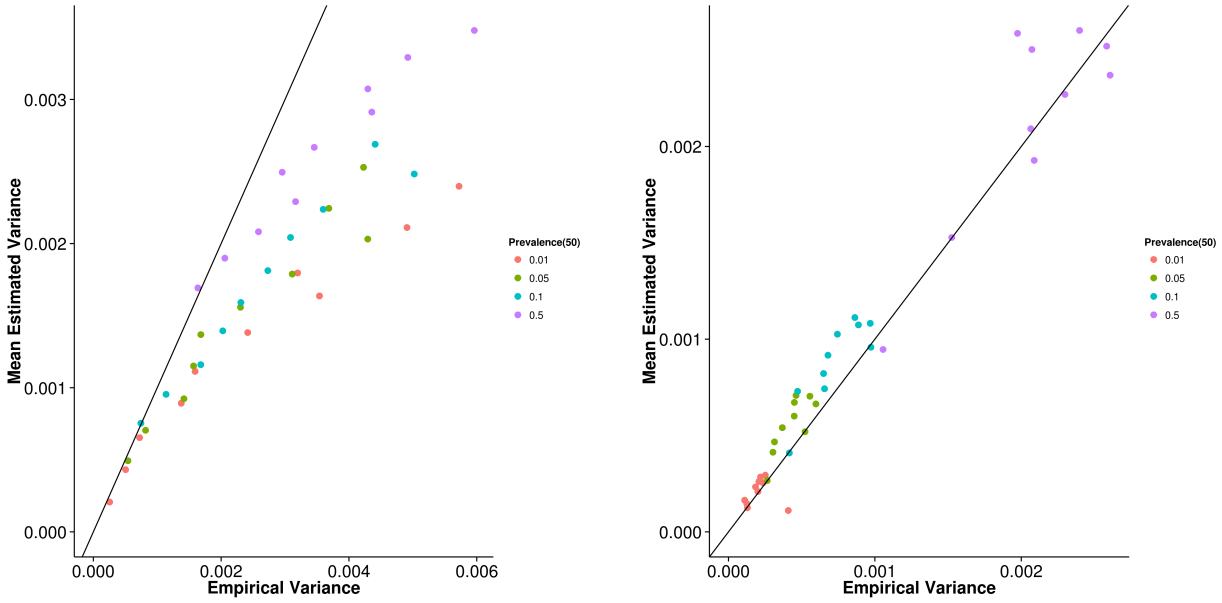
(a) SHREK

(b) GCTA

(c) LDSC with fix intercept

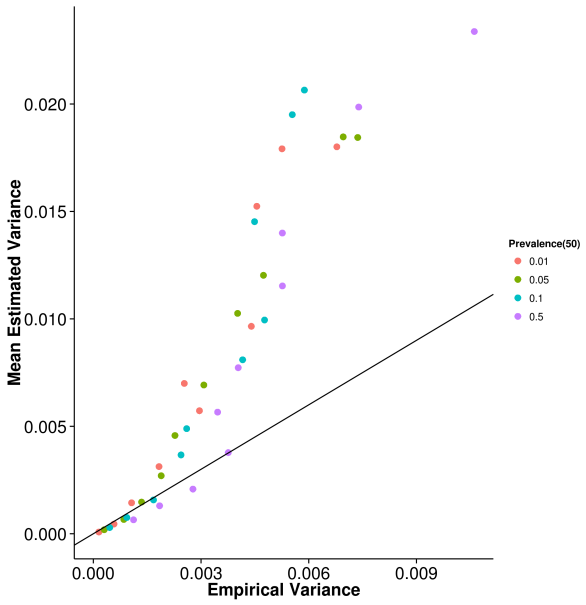
(d) LDSC with intercept estimation

Figure 2.17: 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.

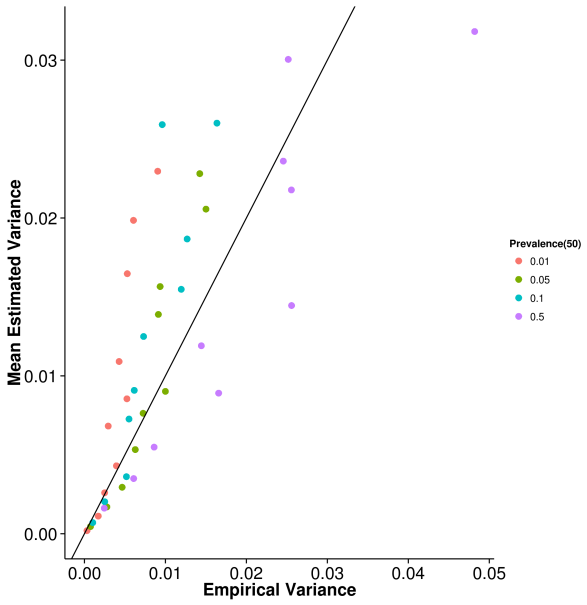


(a) SHREK

(b) GCTA

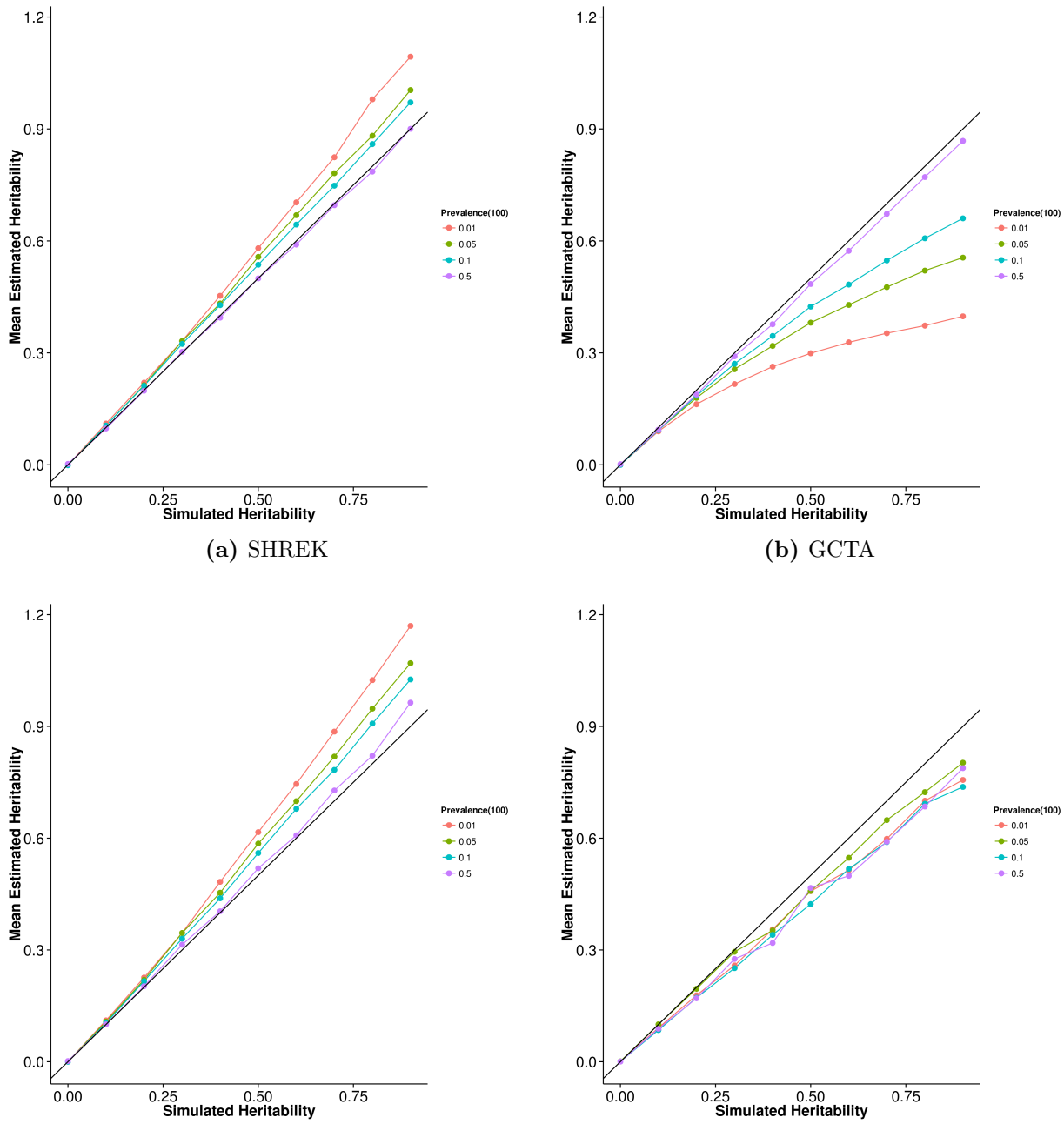


(c) LDSC with fix intercept



(d) LDSC with intercept estimation

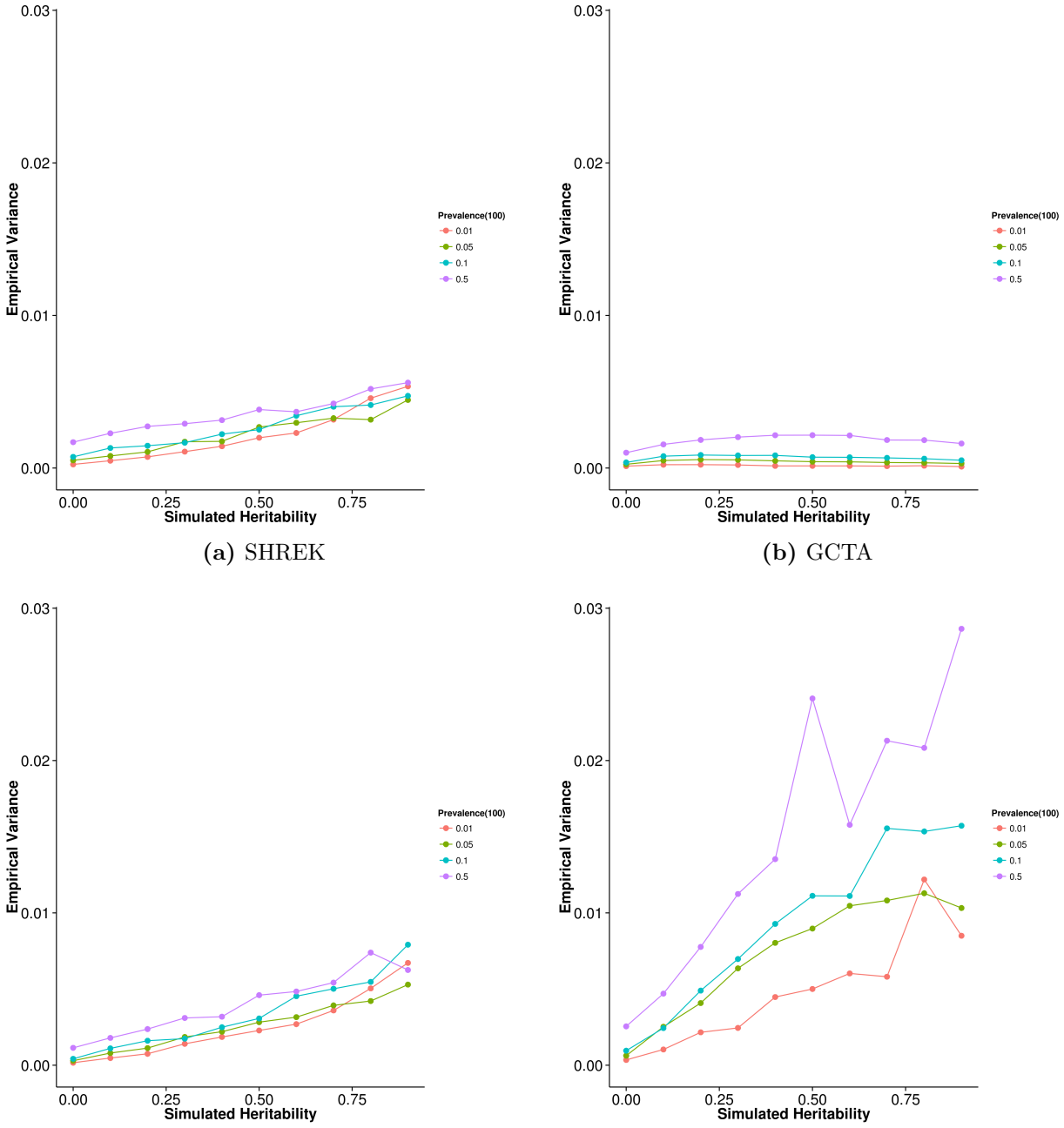
Figure 2.18: Estimated variance of results from case control simulation with random effect size simulation when compared to empirical variance when 50 causal SNPs was simulated. Again, the estimation of variance from SHREK tends to be downwardly biased and LDSC with fixed intercept tends to be upwardly 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 2.19: 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.



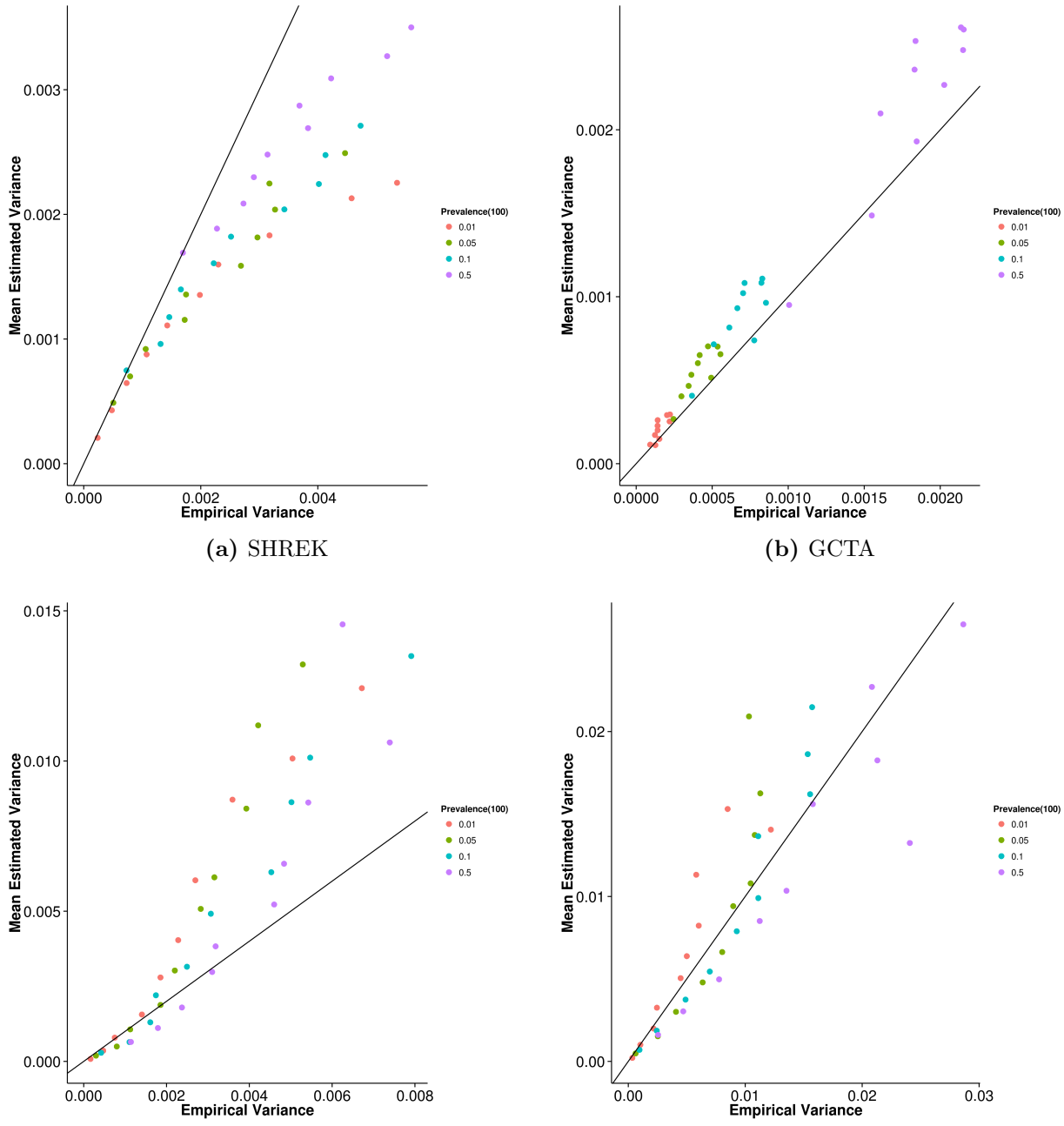
(a) SHREK

(b) GCTA

(c) LDSC with fix intercept

(d) LDSC with intercept estimation

Figure 2.20: 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.



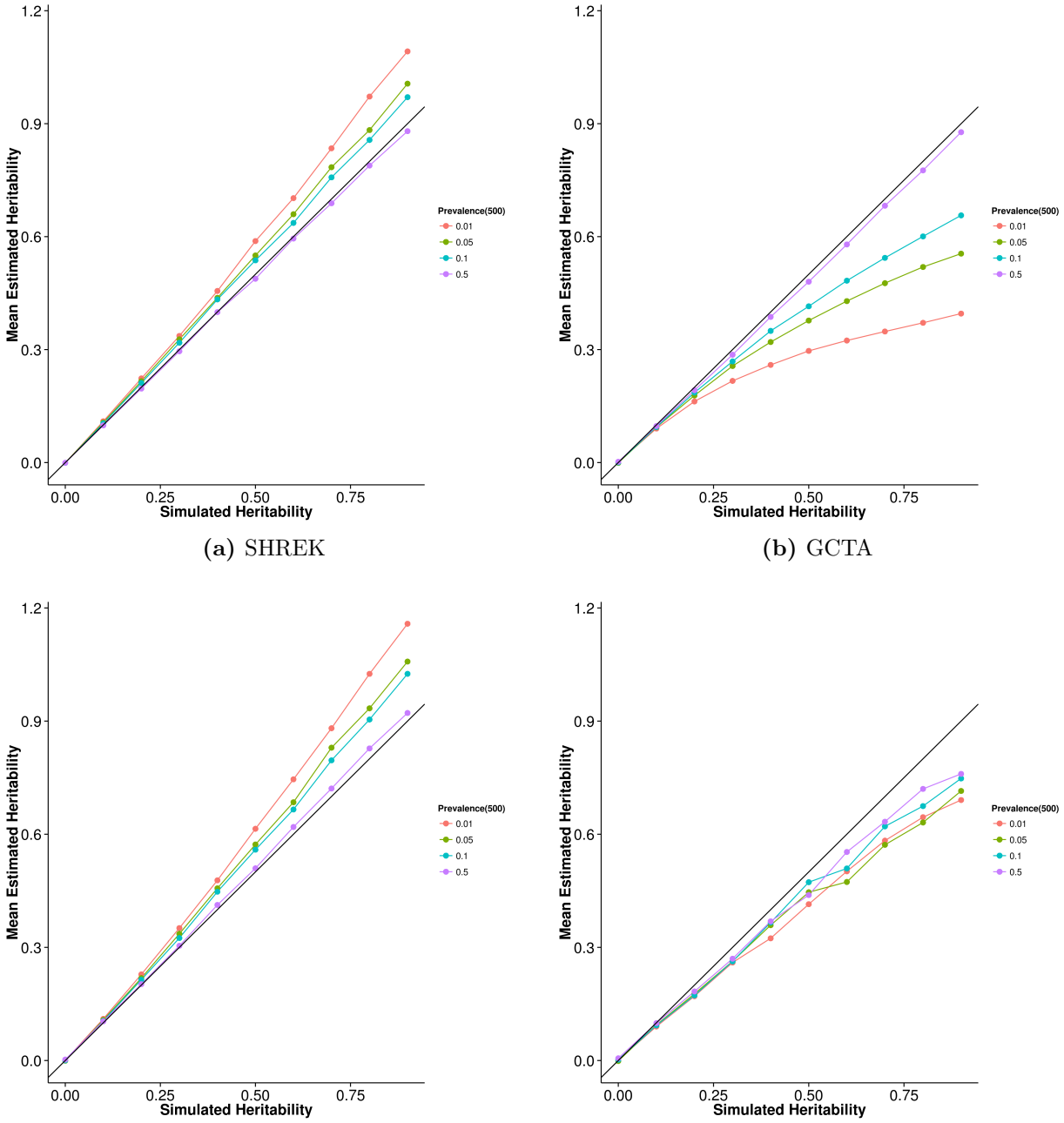
(a) SHREK

(b) GCTA

(c) LDSC with fix intercept

(d) LDSC with intercept estimation

Figure 2.21: 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.



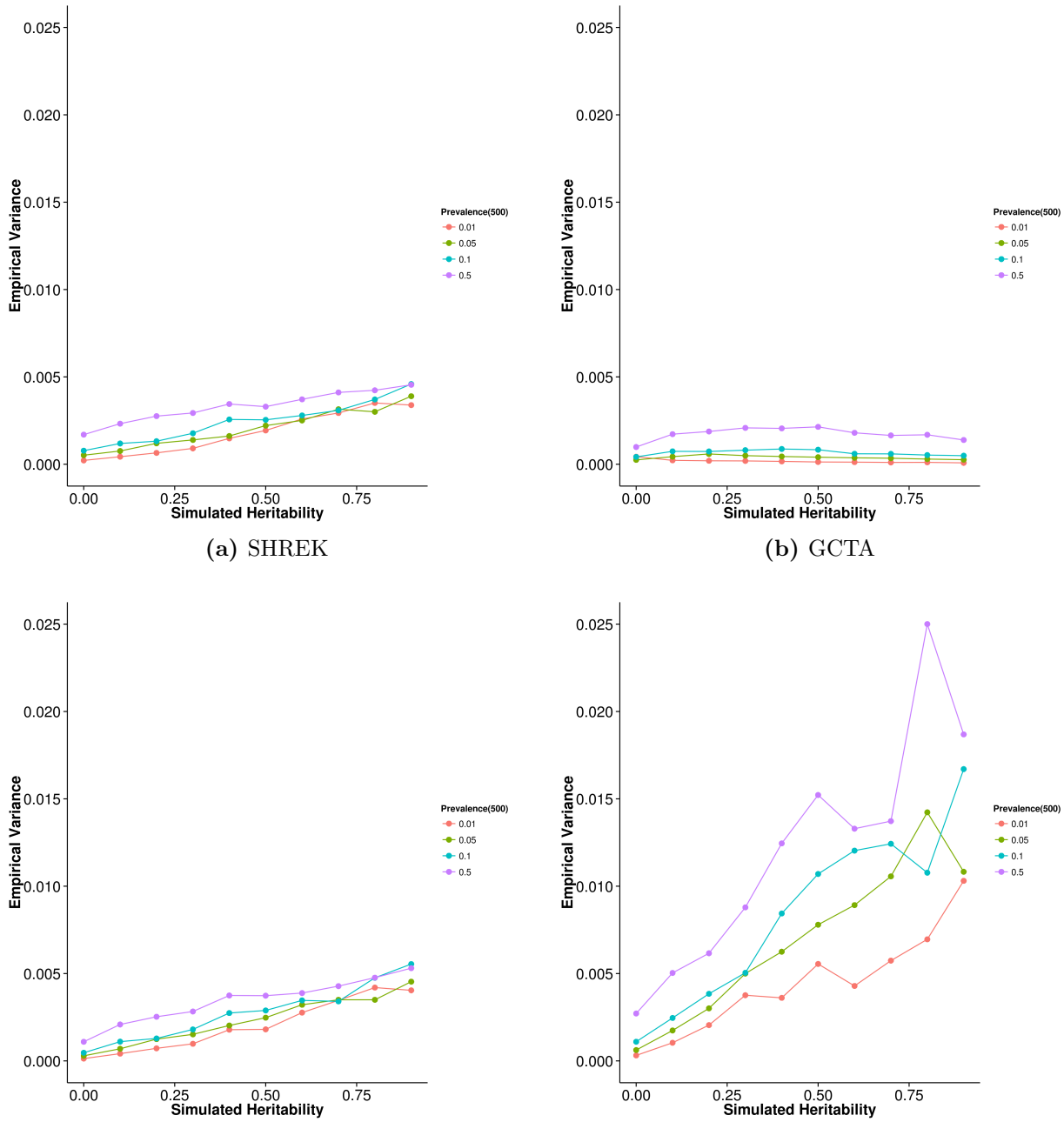
(a) SHREK

(b) GCTA

(c) LDSC with fix intercept

(d) LDSC with intercept estimation

Figure 2.22: 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.



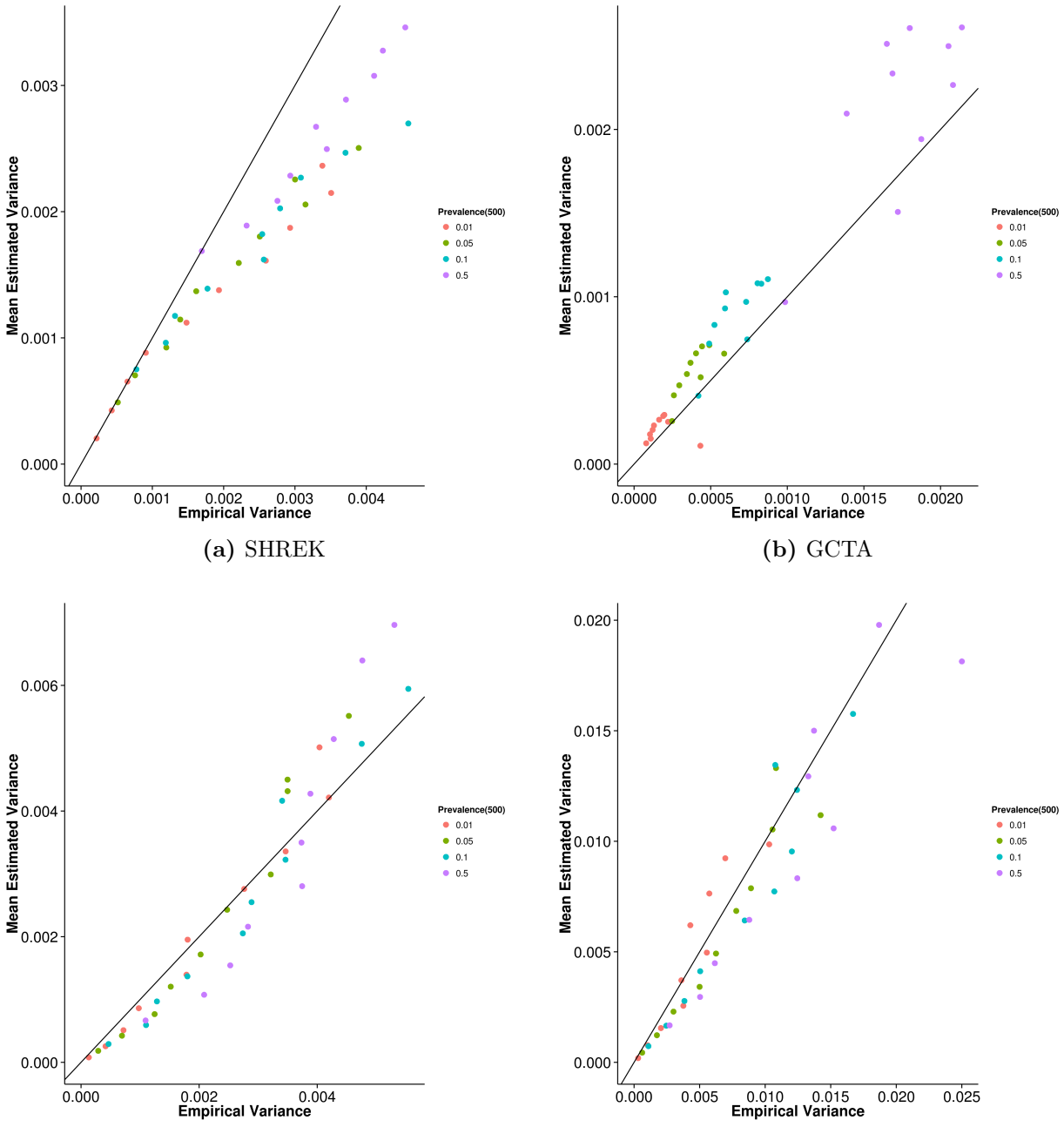
(a) SHREK

(b) GCTA

(c) LDSC with fix intercept

(d) LDSC with intercept estimation

Figure 2.23: 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 2.24: 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.

3 n-3 Polyunsaturated Fatty Acid Rich Diet in Schizophrenia

3.1 Introduction

Because schizophrenia is a genetic disorder with heritability of $\sim$ 80%, much of the research efforts have been spend on the genetic studies of schizophrenia. However, environmental factors also have some influence to schizophrenia. From the meta-analysis of the twin studies, **sullivan2003schizophrenia** estimated that the shared or common environmental influences on liability to schizophrenia is around 11% with a confidence interval ranging from 3%-19%. In addition, it was observed that there are some genetic-environmental interaction influencing schizophrenia (**Tienari2004**; **Clarke2009**), which may lead to overestimation of the heritability **zuk2012mystery** Therefore, environmental factors might also be an important target for the study of schizophrenia.

Among all possible environmental influence, we are most interested to investigate the effect of prenatal infection to schizophrenia because of previous identification of interaction between prenatal infection and genetic variations (**Clarke2009**).

Slowly but steadily, progress has been made in the study of the effect of prenatal infection in schizophrenia using the rodent models (**Oskvig2012**; **Smith2007**;

Garbett2012a). Current evidences suggest that maternal immune activation (MIA) as the main mechanisms that leads to the increased risk of schizophrenia in individual whose mother were exposed to infection during pregnancy. It is suggested that MIA serves as a "disease primer" that increase the offspring's vulnerability to subsequent insults (**Giovanoli2013**) or are causing a developmental "lag" in the nervous system (**Meyer2007a**; **Garbett2012a**).

However, most of these studies have been focusing on the effect of MIA during the mid-gestation period, yet it was suggested that MIA during early gestation have a larger impact when compared to MIA during mid/late-gestation (**Meyer2007a**; **Li2010a**)

Furthermore, it has been suggested that n-3 polyunsaturated fatty acid (PUFA) diet might have potential in alleviating the symptoms of schizophrenia (**Li2015**; **Trebble2003**). In mouse, it was found that n-3 PUFA can inhibits the production of IL-6 (**Trebble2003**) - a major mediator in MIA model (**Smith2007**). Apart from its anti-inflammatory property, n-3 PUFA such as docosahexaenoic acid (DHA) also plays a critical role in the development of central nervous system (**Clandinin1999**; **Kitajka2002**). Given its strong implication in neuronal functioning, it is possible that n-3 PUFA rich diet may reduce the symptoms of schizophrenia, as reported by a recent study (**Li2015**).

Given these information, we are interested to not only investigate the effect of MIA during early gestation, but also the effect of n-3 PUFA rich diet to the adult offspring exposed to early MIA insults. Herein, we perform a RNA Sequencing study to investigate the gene expression changes induced by early MIA exposure in the brain of the adult offspring, and also expression changes induced by n-3 PUFA rich diet, using the polyriboinosinic-polyribocytidilic acid (PolyI:C) mouse model.

Additionally, it is interesting to investigate whether if gene sets found to be associated with schizophrenia are also enriched by genes differentially expressed

in the MIA condition. It is also interesting to investigate the relative contribution of these gene sets to the liability risk of schizophrenia. Therefore, LDSC analysis was performed to investigate the relative contribution of these candidate gene sets to the heritability of schizophrenia. SHREK was not used for this analysis because we have not tested for its performance in the partitioning of heritability, whereas **Finucane2015** has conducted simulations to show that LDSC can perform the partitioning of heritability.

The work in this chapter were done in collaboration with my colleagues who have kindly provide their support and knowledges to make this piece of work possible. Dr Li Qi and Dr Basil Paul were responsible for generating the animal model and providing the sample for our study; Dr Li Qi and Dr Desmond Campbell helped with the experimental design; Vicki Lin has helped with the RNA extraction; Tikky Leung for her high quality sequencing service; Nick Lin for his help in tackling problems encountered during sequencing quality control; Dr Johnny Kwan, Dr Desmond Campbell, Dr Timothy Mak and Professor Sham for their guidance in the statistical analysis.

3.2 Methodology

3.2.1 Sample Preparation

Female and male C57BL6/N mice were bred and mated by The University of Hong Kong, Laboratory Animal Unit. Timed-pregnant mice were held in a normal light–dark cycle (light on at 0700 hours), and temperature and humidity-controlled animal vivarium. All animal procedures were approved by the Committee on the Use of Live Animals in Teaching and Research (CULATR) at The University of Hong Kong.

The MIA model was generated following procedures previously reported (**Li2009c**). A dose of 5mg kg^{-1} PolyI:C in an injection volume 5ml kg^{-1}, prepared on the day of injection was administered to pregnant mice on Gestation Day (GD) 9 via the tail vein under mild physical constraint. GD 9 is when the neural tube close and previous studies suggested that this is the critical period where MIA during this period have a larger effect than MIA in mid-gestation period (**Meyer2007a**; **Li2010a**). Therefore it was selected as the MIA time point in our study.

Control animals received an injection of 5ml kg^{-1} 0.9% saline. The animals were returned to the home cage after the injection and were not disturbed, except for weekly cage cleaning. The resulting offspring were weaned and sexed at postnatal day 21. The pups were weighed and littermates of the same sex were caged separately, with three to four animal per cage. Half of the animal were fed on diets enriched with n-3 PUFAs and half were fed a standard lab diet until the end of the study. The latter 'n-6 PUFA' control diet had the same calorific value and total fat content as the n-3 PUFA diet. The diets were custom prepared and supplied by Harlan Laboratories (Madison, WI, USA). The n-6 and n-3 PUFA were derived from corn oil or menhaden fish oil, respectively. The n-6 PUFA control diet, was based on the standard AIN-93G rodent laboratory diet (**Reeves1993**), and contained 65 g kg^{-1} corn oil and 5 g kg^{-1} fish oil with an approximate (n6)/(n3) ratio of 13:1. The n-3 PUFA diet contained 35 g kg^{-1} corn oil and 35 g kg^{-1} fish oil with an approximate (n6)/(n3) ratio of 1:1 (**Olivo2005**). To avoid being confounded by sex difference, we only use the male offspring for our analysis. The male offspring were sacrificed by cervical dislocation on postnatal week 12, which roughly correspond to adulthood in human, and the cerebellum was extracted and stored in -80°C until RNA extraction.

Although generally, hippocampus (**Velakoulis2006**; **Nugent2007**) and prefrontal cortex (**Knable1997**; **Perlstein2001**) are the two most interesting brain

regions to study in schizophrenia, cerebellum dysfunction has also been reported in schizophrenia (**Yeganeh-Doost2011**; **Andreasen2008**). Specifically, positron emission tomography (PET) studies have shown that a dysfunction in the cortico-cerebellar-thalamic-cortical neuronal circuit, which contributes to "cognitive dysmetria", e.g. impaired cognition, and other symptoms of schizophrenia (**Yeganeh-Doost2011**). Taken together, cerebellum might plays an important part in the etiology of schizophrenia and are therefore suitable for the current study.

3.2.2 RNA Extraction, Quality Control and Sequencing

Total RNA was extracted from each cerebellum tissue using RNeasy midi kit (Qiagen) following the manufacturer's instructions. RNA quality was assayed using the Agilent 2100 Bioanalyzer and RNA was quantified using Qubit 1.0 Flurometer. Samples with RNA integrity number (RIN) < 7 were not included in our study as the RNA are most likely degraded. As a hypothesis generation study, we select a minimum of 3 samples per group and each samples must come from a different litter to control for littering effect. The RNA Sequencing library was performed at the Centre for Genomic Sciences, the University of Hong Kong, using the KAPA Stranded mRNA-Seq Kit. All samples were sequenced using Illumina HiSeq 1500 at 2 lanes (2×101 bp paired end reads). We distribute the samples such that each lane contain roughly the same amount of samples from different conditions.

3.2.3 Sequencing Quality Control

Quality control (QC) of the RNA Sequencing read data was assessed by FastQC (**Andrews2010**), which reports the overall quality of the high throughput sequence, and allow the identification of any potential problems and biases.

From the FastQC report, it was noted that some adapter sequences re-

SampleID	Litter	Diet	Condition	Lane	Batch	Rin
B1	3	O3	POL	1	B	7.7
B2	6	O3	POL	2	B	7.7
F1	4	O3	POL	1	F	7.6
F4	1	O3	SAL	2	F	8.1
B4	5	O3	SAL	1	B	7.8
B5	14	O3	SAL	2	B	7.7
F2	2	O6	POL	1	F	7.5
E3	11	O6	POL	2	E	7.8
C2	7	O6	POL	2	C	7.9
B6	13	O6	SAL	2	B	7.4
E6	14	O6	SAL	1	E	8
C6	1	O6	SAL	1	C	7.8

Table 3.1: Sample information. O3 = n-3 PUFA diet; O6 = n-6 PUFA diet; POL = PolyI:C exposed; SAL = Saline exposed. We have tried to separate the samples into different lane and batch to control for the lane and batch effect. Samples from different litters were also used with the exception of F4 and C6 which came from the same litter but were given a different diet.

mained in the final sequence. By using trim_glore, a wrapper for cutadapt (version 1.9.1) (**Martin2011**), the adapter sequences were removed from the sequence reads and only reads that were at least 75 bp long were retained for subsequent alignment.

3.2.4 Alignment

In a recent review by **Engstrom2013** it was demonstrated that STAR (**Dobin2013**) has the best performance in term of accuracy and speed among all the aligners investigated. Thus STAR aligner was used in our study. The RNA sequencing reads were mapped to the *Mus musculus* reference genome (mm10, Ensembl GRCm38.82) using the STAR aligner (version 2.5.0a) (**Dobin2013**). And the quantification of the gene expression levels were conducted using featureCounts (version 1.5.0) (**Liao2014**).

3.2.5 Data Quality Assessment

Data quality assessment and quality control are essential steps of any data analysis. In order to assess the quality of the count data, unsupervised clustering was performed using the functions provided by the DESeq2 (version 2.1.4.5) package. Sample with abnormal count data was removed from the analysis.

3.2.6 Differential Expression Analysis

There are many statistical tools available for the differential gene expression analysis. Based on the review of **Seyednasrollah2015** it was suggested that DESeq2 and limma are the most robust statistical packages for analyzing RNA Sequencing data. As the authors of DESeq2 are very active in providing supports for the package, DESeq2 (version 2.1.4.5) (**Love2014**) was used as the statistic package for the differential gene expression analysis.

One of the most controversial RNA sequencing study in RNA Sequencing was the mouse ENCODE study by **Yue2014** where most of the findings reported were found to be confounded by lane and batch effect **Gilad2015** This highlights the importance of lane and batch effect in the design of RNA Sequencing. To avoid batch and lane effect, the whole sampling collection procedure and sequencing was performed in a way where we minimize the batch and lane difference between conditions (table 3.1). However, because of the sample quality differed across different batches, we were unable to fully balance out the batch effect. Therefore, it was necessary to control for batch effect in the analyzes.

The following statistical comparisons were performed:

1. Saline exposed samples with n-3 PUFA rich diet vs Saline exposed samples with n-6 PUFA rich diet

2. PolyI:C exposed samples with n-3 PUFA rich diet vs PolyI:C exposed samples with n-6 PUFA rich diet

3. Saline exposed samples with n-6 PUFA rich diet vs PolyI:C exposed samples with n-6 PUFA rich diet

We used $\sim Batch + Condition + Diet + Condition : Diet$ as our model of statistical analysis where Condition is the MIA exposure status. RIN was not included in the statistical model as suggested by the author.

In our analysis, genes with base mean count < 10 were removed to reduce noise associated with low expression and the Benjamini and Hochberg method was then used to correct for multiple testing.

3.2.7 Gene Set Analysis

The main goal of the current study is to investigate whether the effect of MIA or diet act upon the same functional gene sets as the genetic variants associated with schizophrenia in the development of the disease. Specifically, as genes related to postsynaptic density (PSD) (**purcell2014polygenic**; **Consortium2015a**) and calcium ion channel (**purcell2014polygenic**; **Ripke2014**; **Szatkiewicz2014**) has been implicated to be involved in the etiology of schizophrenia, it is interesting to investigate whether these gene sets were also enriched by genes perturbed by MIA or diet.

To compile a list of relevant gene-sets, significant gene sets from **purcell2014polygenic**; **Consortium2015a** were retrieved, which include gene set related to calcium ion channel and PSD. Gene sets and pathway related to PSD and calcium ion channel were also retrieved from Kyoto Encyclopedia of Genes and Genomes (KEGG) (**Kanehisa2000**) and Gene Ontology (GO) (**Consortium2015b**), both have been widely used for systematic genetic analysis. In addition, the schizophrenia GWAS

gene set, constructed based on associated GWAS LD-intervals from **Ripke2013** by **purcell2014polygenic** was also included.

The Wilcoxon Rank Sum test was performed to assess whether the gene sets were enriched by genes affected by either MIA or diet. Pathways with adjusted p-value < 0.05 (using Benjamini and Hochberg adjustment) were considered as significant. To ensure the same significance are not observed in neutral gene sets, we also performed permutation analysis where 1,000 pseudo gene sets with the same number of genes as the candidate gene sets were created at random. We then performed the same gene set enrichment analysis on these gene sets to obtain an empirical distribution of the null. Based on the empirical distribution, we were able to derive the empirical p-value of the gene sets.

3.2.8 Partitioning of Heritability

In order to identify the relative contribution of the significant gene sets to the heritability of schizophrenia, partitioning of heritability of schizophrenia was performed using LDSC.

Firstly, SNPs were assigned to genes based on human genome hg19 positions if they lay within 35 kb upstream or 10 kb downstream of the gene. If SNPs mapped within more than one gene, they were assigned to all such genes, following the procedure employed by **Consortium2015a**

Then, the partitioning of heritability was performed using LDSC (**Bulik-Sullivan2015**) --annot and --overlap-annot options, with window size of 1000kb window size and the LD score generated in section 2.2.11. The MHC region (chr6:25,000,000-35,000,000) was removed from the analysis due to its unusual LD and genetic architecture (**Finucane2015**).

3.2.9 Designing the Replication Study

The sample size of the current study is relatively small and therefore only serves as a pilot study. It is therefore important to utilize the information from the current study to design a more powerful follow-up study.

In order to estimate the required sample size for the follow-up studies, power estimation was performed using Scotty (**Busby2013**). Based on the current count data, Scotty can estimate the required sample size of the follow up study in order to detect at least 90% of the differentially expressed genes with least 2× difference, and for at least 80% of genes to reach 80% of the maximum power.

3.3 Results

3.3.1 Sample Quality

On average, 87 million reads were generated for each sample of which more than 90% of the read bases has quality score > 30. More than 97% of the sequence reads remains after adapter trimming was performed. Over 90% of the trimmed reads were uniquely mapped to the *Mus musculus* reference genome (mm10, Ensembl GRCm38.82) using the STAR aligner (version 2.5.0a) (**Dobin2013**).

Unsupervised clustering was performed to assess the count data quality. It is observed that none of the samples are clustered by lane or by batch (fig. 3.1). However, one of the sample in the n3-PolyI:C group is found to be substantially different from all other samples. It is unclear whether the difference was a result of sample contamination from other sources, or was a result of sample mis-label. The sample was therefore excluded from subsequent analyses.

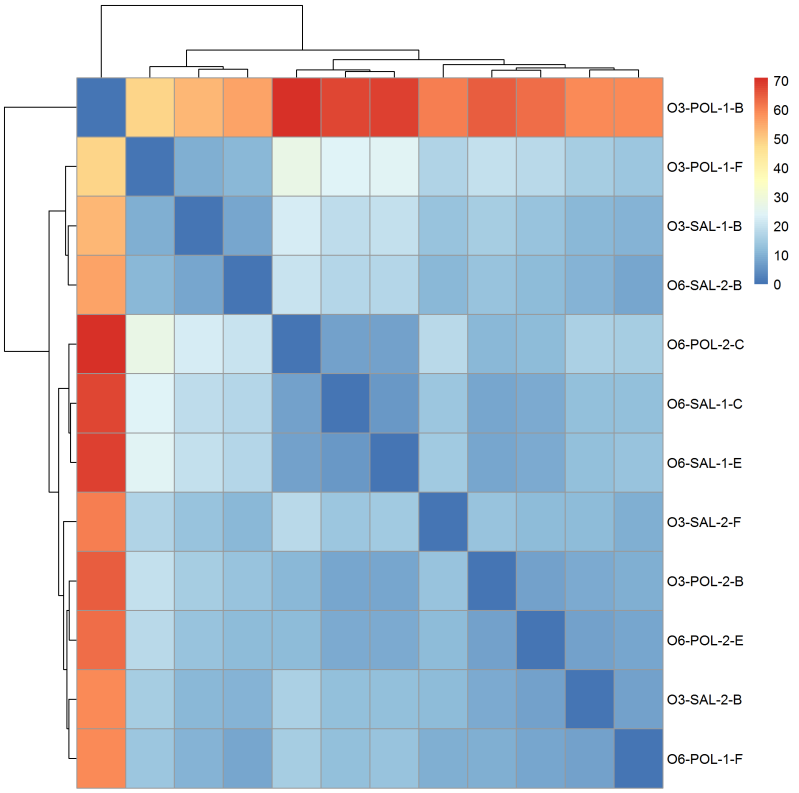


Figure 3.1: Sample Clustering results. Samples were labeled as <Diet>-<Condition>-<Lane>-<Batch> where O3 = n-3 PUFA rich diet; O6 = n-6 PUFA rich diet; POL = PolyI:C; SAL = Saline.
No clear clustering for lane or batch effects are observed. However, one sample from the n3-PUFA-PolyI:C group is found to be substantially different from all other samples. It is unclear whether the difference is due to sample contaminations or sample mis-label. To avoid problems in down-stream analysis, we excluded this sample from subsequent analyses

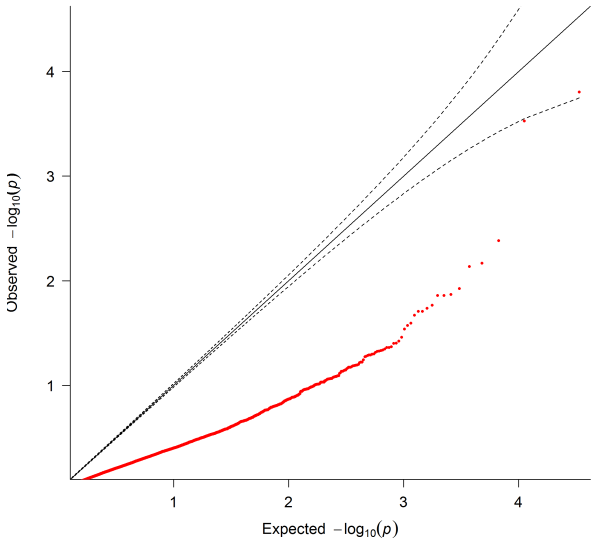
3.3.2 Differential Expression Analysis

DESeq2 analysis was performed after excluding the problematic sample. Of the 16,747 genes that passed through quality control, only *Sgk1* (p-adjusted=0.00186) was found to be significantly differentially when comparing the effect of n-3 PUFA rich diet in PolyI:C exposed mice (fig. 3.2c). On the other hand, no significant differentiation is observed in all other comparisons (figs. 3.2a and 3.2b).

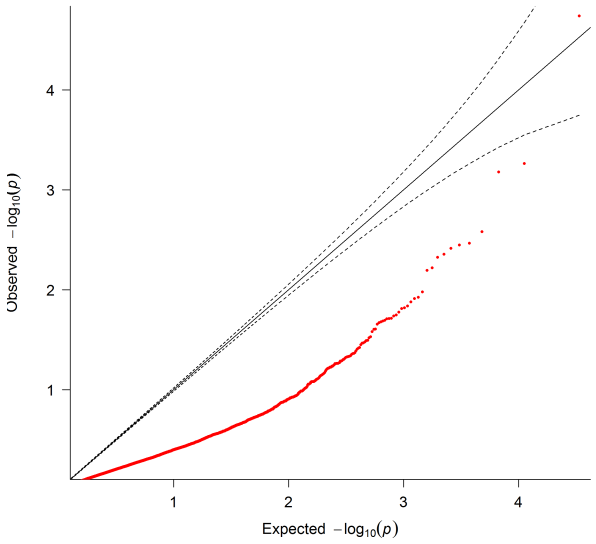
3.3.3 Gene Set Analysis

In total, 7 gene sets were included for the gene set analysis (table 3.2). Of the 7 gene sets tested, 6 are significantly enriched in MIA, whereas only the PSD gene set from GO are significantly enriched in PolyI:C exposed mice given the n-3 PUFA rich diet. None of the gene sets are significant in Saline exposed mice given the n-3 PUFA rich diet. All significant gene sets remain significant after performing the permutation analysis.

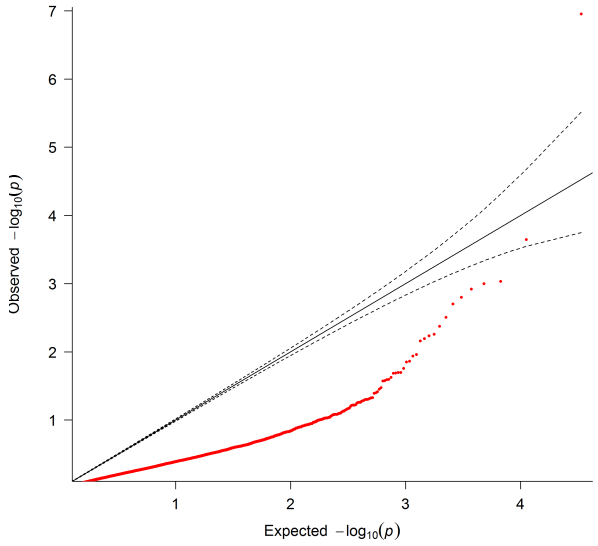
For all the gene sets related to PSD, the PSD gene set from **purcell2014polygenic** is the only one that is not found to be significant in all conditions. Upon further investigation, the PSD gene set from **purcell2014polygenic** is found to be based on the work of **Kirov2012** which includes not only the PSD, but also neuronal activity-regulated cytoskeleton-associated protein (ARC), N-methyl-D-aspartate (NMDA) receptor complex and metabotropic glutamate receptor 5 (mGluR5) subsets.



(a) O6-Saline mice vs O6-PolyI:C mice



(b) O6-Saline mice vs O3-Saline mice



(c) O6-PolyI:C mice vs O3-PolyI:C mice

Figure 3.2: QQ Plot of statistic results. From the quantile-quantile Plot (QQ-plot), it is observed that most of the observed p-values are less than expected. Because the sample size is relatively small, it is likely for the current study to lack detection power, therefore leads to an deflation in p-values.

Table 3.2: Results of gene set analysis. In total, 7 gene sets were retrieved from **purcell2014polygenic** KEGG and GO. Firstly, Wilcoxon Rank sum test was performed. Except for the PSD gene set obtained from **purcell2014polygenic** all pathways are enriched in MIA. On the other hand, the PSD gene set obtained from GO is the only gene set that are significantly enriched in PolyI:C exposed mice receiving the n-3 PUFA rich diet, whereas none of the gene sets are significantly enriched in Saline exposed mice receiving the n-3 PUFA rich diet. Upon further investigation, the PSD gene set from **purcell2014polygenic** was found to be based on the work of **Kirov2012** which includes not only the PSD, but also ARC, NMDA receptor complex and metabotropic glutamate receptor 5 (mGluR5) subsets. The broader definition of the PSD gene set form **purcell2014polygenic** might explain the difference observed between the PSD set from **purcell2014polygenic** and PSD set from GO.

Gene Set	Source	Size	Category	Diet in PolyIC Mice	MIA Effect	Diet in Saline Mice	Proportion of h^2 explained
Calcium Ion Signaling Pathway	KEGG (hsa04020)	180	Calcium Ion	0.0402	4.40×10^{-7}	0.231	0.0135
Glutamatergic synapse	KEGG (hsa04724)	114	PSD	0.118	0.00490	0.123	0.0134
Voltage-Gated Calcium Channel Activity	GO (GO:05245)	44	Calcium Ion	0.0262	3.45×10^{-6}	0.137	0.00771
Calcium Channel Activity	GO (GO:05262)	111	Calcium Ion	0.0942	0.00209	0.0880	0.0119
PSD	GO (GO:14069)	194	PSD	4.86×10^{-3}	6.31×10^{-9}	0.0383	0.0352
PSD	Purcell	685	PSD	0.113	0.328	0.977	0.0486
Schizophrenia GWAS	Purcell	479	GWAS	0.3048	6.91×10^{-3}	0.551	0.0998

3.3.4 Partitioning of Heritability

It is observed that all calcium ion channel gene sets accounts for around 0.77-1.35% of the SNP heritability of schizophrenia, whereas the PSD gene sets contribute 1.34% to 4.86%.

Finally, the schizophrenia GWAS gene set contribute most to the SNP heritability of schizophrenia, contributing 9.9% of the SNP heritability.

3.3.5 Designing the Replication Study

Using Scotty (**Busby2013**), it is estimated that a minimal of 10 samples per group are required for the follow-up study in order to obtain the desirable power.

3.4 Discussion

3.4.1 Serine/threonine-protein kinase

Our results demonstrated that the expression of Serine/threonine-protein kinase *Sgk1* in the cerebellum of PolyI:C exposed mice might have been affected by n-3 PUFA rich diet. *Sgk1* is a serine/threonine kinase activated by phosphatidylinositol 3-kinase (PI3K)/Akt signaling. Studies have reported that the expression of *Sgk1* is associated with spatial learning, fear-conditioning learning and recognition learning in rat (**Tsai2002**; **Lee2003**). For example, **Tsai2002** observed a 4 fold increase of *Sgk1* in the hippocampus of fast learners when compared to slow learners. Furthermore, the transfection of *Sgk1* mutant DNA impairs the water maze performance in rat (**Tsai2002**).

On the other hand, it was found that *Sgk1* can regulates the AMPA

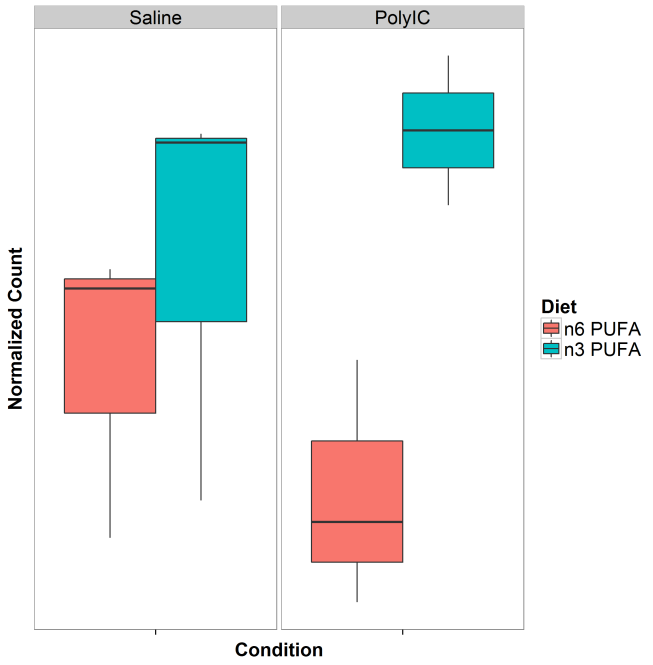


Figure 3.3: Normalized Expression of *Sgk1*. It was observed that the expression level of *Sgk1* increases after the mice was given a n3-PUFA rich diet where a significant increase was observed in mice exposed to PolyI:C.

and kainate glutamate receptors, especially GluR6, which is encoded by *Grik2* (**Lang2006**; **Lang2010**). The kainate receptors contribute to the excitatory post-synaptic current and are important to the synaptic transmission and plasticity in the hippocampus (**Lang2006**). The upregulation of AMPA and kainate receptors are therefore expected to enhance the excitatory effects of glutamate (**Lang2010**).

Furthermore, *Sgk1* can up-regulates the glutamate transporters such as EAAT4 (**Bohmer2004**), which are vital for the clearance of glutamate from the synaptic cleft. This prevents excessive glutamate accumulation, thus help to prevent the neurotoxic effects of glutamate (**Lang2010**). In addition, **Schoenebeck2005** demonstrated that *Sgk1* has a neuroprotective role in oxidative stress situations. Together, the evidences suggest that *Sgk1* has an important role in the regulation of the glutamatergic system. An increase in expression of *Sgk1* might help to improve normal functioning of the glutamatergic system.

Interestingly although the expression of *Sgk1* is lower in the PolyI:C ex-

posed samples when compared to the Saline exposed samples (fig. 3.3), the difference is insignificant (unadjusted p-value=0.0254, q-value = 0.999). A significant difference is only observed when comparing the effect of n-3 PUFA rich diet and the control diet in PolyI:C exposed mice. The expression of *Sgk1* is significantly higher in PolyI:C exposed samples who received the n-3 PUFA rich diet. Although there is no direct evidence linking n-3 PUFA diet with the expression of *Sgk1*, **Zhang2015** demonstrated that n-3 PUFA diet can activates the Akt prosurvival pathway, therefore protecting the neurons from brain damage. Most importantly, the Akt signaling pathway is responsible for the activation of *Skg1* (**Lang2010**).

Therefore, we speculate that the n-3 PUFA rich diet might have indirectly enhanced the expression of *Sgk1* in PolyI:C exposed mice, therefore reduces the schizophrenia-like behaviours. Consider the important role of *Sgk1* in the regulation of the glutamatergic system, the role of *Sgk1* in the effects of n-3 PUFA rich diet on behaviour in the MIA mouse model will be an interesting line of further investigation.

However, previous studies have been focusing on the effect of *Sgk1* in the hippocampus instead of the cerebellum. It is uncertain whether *Sgk1* has the same function in the cerebellum. Therefore, further researches are required to investigate the role of *Sgk1* in the regulation of development of cerebellum.

Additionally, although none of the genes passed the bonferroni threshold in other comparison, it is noted that there are a number of genes that are more significant than all other genes (figs. 3.2a and 3.2b). Upon further investigation, it was found that when comparing the effect of MIA to mouse receiving the control diet, two genes, *Ptgds* and *Capn11*, have higher significance when compared to others. Interestingly, *Ptdgds* catalyzes the conversion of prostaglandin H2 (PGH2) to postaglandin D2 (PGD2), a neuromodulator in the central nervous system, whereas *Capn11*, which encodes for intracellular calcium-dependent cysteine proteases, which

are both related to schizophrenia. Given the interesting function of these genes, it is possible that when more samples are included into the study, there will be sufficient power of association that these genes will pass the significant threshold.

On the other hand, when inspecting the effect of n-3 PUFA rich diet in the saline mouse, it was observed that of the top 10 most significant genes, only 2, *Bub1* and *Sgk1* were also among the top 10 most significant genes when comparing the effect of n-3 PUFA rich diet in the PolyI:C samples. It is therefore possible that the significance of *Bub1* and *Sgk1* are additive effect of the difference in diet and cautions must be taken in the interpretation of the significance of *Sgk1* in the PolyI:C samples.

3.4.2 Gene Set Analysis

In total, 7 gene sets were included in the analysis (table 3.2). All gene sets related to calcium ion channel are found to be significant when comparing the gene expression in MIA samples. Previous studies in schizophrenia have reported the association of genes participating in the calcium ion channel signaling with schizophrenia (**Lidow2003**; **purcell2014polygenic**; **Ripke2014**). For example, in exome sequencing study of schizophrenia conducted by **purcell2014polygenic** an enrichment of non-synonymous variants within the voltage gate calcium ion channel genes was observed in the schizophrenia cases. Similar findings were also obtained in the PGC schizophrenia GWAS (**Ripke2014**).

Calcium ion channel signaling is a key component for normal neural functioning. For example, calcium ion signaling can regulates neuronal gene transcription, neuronal excitability, synaptic plasticity responsible for learning and memory, as well as the release of neurotransmitters from presynaptic endings (**Berridge2014**). Although it is unclear the exact role of the calcium signaling pathway in the etiology

of schizophrenia, it is likely for the disruption of expression or structures of proteins related to the calcium signaling pathway can affect the normal functioning of the neuronal system.

On the other hand, gene sets related to PSD are also found to be significant when comparing the gene expression in MIA samples. PSD genes are highly conserved and have critical roles in excitatory neural signalling components, as well as dendrite and spine plasticity. PSD abnormalities are therefore thought to alter the balance of excitation and inhibition, and variations in this balance might change, not only local circuit function, but also connectivity patterns between brain regions, leading to developmental and behavioral deficits (**Cline2005**).

Most importantly, it is observed that the schizophrenia GWAS gene set, constructed based on associated GWAS LD-intervals from **Ripke2013** by **purcell2014polygenic** is also found to be significant in MIA. This indicates that the genes contain genetic variants associated with schizophrenia are also likely to be affected by early MIA events in the cerebellum. Thus, genetic variants associated with schizophrenia and differential expression induced by early MIA might be affecting similar genetic pathways.

Last but not least, it is observed that a significant difference in PolyI:C exposed mice receiving different diet is only observed in the PSD gene set from GO.

It has been reported that a n-3 PUFA deficiency has a negative impact to normal brain functioning (**Bazinet2014**; **Calon2005**). Subsequent research shown that the expression of PSD proteins are significantly down-regulated in n-3 PUFA depleted mouse brains (**Sidhu2011**). **Sidhu2011** therefore speculated that the reduction of PSD proteins might be an important mechanism for the suboptimal brain functioning associated with n-3 PUFA deficiency.

Given the interaction between the n-3 PUFA diet and expression of the

PSD proteins, it is possible that the n-3 PUFA rich diet can increase the expression of the PSD proteins in the PolyI:C exposed mice, therefore compensating for the reduced neural functioning, leading to reduction of schizophrenia-like behaviours. Further investigation are required in order to obtain direct evidence of how n-3 PUFA diet reduce the schizophrenia-like behaviour. However, it is likely that the PSD and the *Sgk1* gene will play an important role in the underlaying mechanism.

3.4.3 Partitioning of Heritability

To estimate the relative contribution of common variants in the gene sets to the heritability of schizophrenia, partitioning of heritability was performed using LDSC **Bulik-Sullivan2015**

Not surprisingly, the schizophreniaGWAS gene set contributes most to the SNP heritability of schizophrenia, accounting for 9.98% of the SNP heritability.

On the other hand, the relative contribution of the calcium ion channel to the SNP heritability is much smaller, contributing only 0.77% to 1.35% of the heritability. Similarly, the PSD gene sets also only contribute to 5% of the SNP heritability. The relatively smaller contribution only suggest that the contribution of *common* variants in these gene sets contribute for a small portion to the heritability of schizophrenia. Considering these gene sets were also found to be differentially expressed in MIA, it is plausible for $G \times E$ interaction to act upon these gene sets. If $G \times E$ interaction exists between the MIA and common variants observed in these gene sets, the total contribution of these common variants to the heritability of schizophreniamay be higher.

Nonetheless, our results suggest that the differential expression induced by early MIA events in the mouse cerebellum might be affecting the same functional gene sets as genetic variants associated with schizophreniain the etiology of

schizophrenia. Consider the converging evidence of the involvement of the calcium ion channel signalling and PSD, disruption of the calcium ion channel signalling or the PSD complex might have an important role in the disease etiology of schizophrenia. Therefore, calcium ion channel signalling and the PSD should be served as the focus of further research in schizophrenia.

3.4.4 Future Perspective

Despite the small sample size of the current study, it serves as an important starting point for further studies. Given the current data, it is estimated that a minimum of 10 samples are needed per group in order to have adequate power in the replication study. If there are sufficient resources and time, there are also a number of factors that can be incorporated into the study.

First and foremost, it might be interesting to investigate the sex difference in response to different diet and MIA. Changes common to both male and female might provide important insight to the disease etiology of schizophrenia. Additionally, to our knowledge, no one has studied the effect of n-3 PUFA rich diet in MIA condition other than **Li2015** Therefore, it is uncertain whether if the n-3 PUFA rich diet has any effect to samples exposed to MIA in mid-gestation. Given the different effect of MIA in different gestation, it might be interesting to investigate the effect of n-3 PUFA rich diet to samples exposed to MIA during different gestation day.

Finally, we have been focusing on the diet of the offspring, however, it is possible that the maternal diet during pregnancy might also have an effect. In view of this, one might want to study the effect of maternal diet in combination of MIA to the risk of schizophrenia.

3.4.5 Limitations

We first acknowledge that the sample size of the current study is moderate and might be underpowered. This is reflected in the QQ-plots (fig. 3.2) where the observed p-values are generally smaller than expected. An increased sample size is therefore required in order to obtain a larger detection power.

Secondly, only the male brains were examined in the current study. The decision to direct experimental resources to males was made because there are evidences that the male fetus is more vulnerable to environmental exposures such as inflammation in prenatal life (**Bergeron2013**; **Lein2007**). An interesting follow up study would be to investigate the gender difference in response to MIA and dietary change.

Thirdly, although RNA Sequencing was performed, analysis on alternative splicing or de-novo transcript assembly were not performed. It is because with the current sample size, there are insufficient information for de-novo transcript assembly to be performed. Most importantly, as we lack the resource for the functional analysis of de-novo transcripts, we cannot verify our findings, therefore the de-novo transcript assembly was not performed.

On the other hand, to investigate possible alternative splicing events, analysis has to be performed on transcript level instead of gene level. This increases the possible candidates from 47,400 genes to 114,083 transcripts. Therefore, a much larger detection power is required. Furthermore, the functional annotation of transcripts is difficult. While there are a lot of information for the annotation of genes, information on functional difference between isoforms of the same gene are generally lacking. It is therefore difficult to understand the functional impact of the differential expression of different isoforms.

In view of this, although alternative splicing and de-novo transcripts might

play an important role in response to MIA or dietary changes, de-novo transcript assembly and alternative splicing analysis were not performed. Nevertheless, as RNA Sequencing was performed, de-novo transcript assembly and alternative splicing analysis can be performed when sufficient samples are collected in the future.

Fourthly, a high RNA expression level does not guarantee a high protein concentration (**Vogel2012**). Post transcriptional, translational and degradation regulation can all affect the rates of protein production and turnover, therefore contributes to the determination of protein concentrations, at least as much as transcription itself (**Vogel2012**). The RNA Sequencing thus only provide an approximation to the concentration of a particular protein in the samples. Results from the RNA Sequencing study should serve as a candidates for further functional analysis protein assays in order to obtain a better understanding of the condition.

Finally, at the time of this thesis, real time PCR (RT-PCR) and functional studies have not been performed to validate our findings. As RNA Sequencing does not provide any causal linkage between the phenotype and the differential expression functional studies must be carried out in order to validate the functional impact of the differential expression. Moreover, it is also important to validate the expression counts from RNA Sequencing using RT-PCR. Currently, the RT-PCR on *Sgk1* are in progress. Shall the results be validated, subsequent functional studies can be performed.

3.5 Supplementary

Litter	Condition	Diet	Cage	Batch	Lane
1	PolyIC	n-3 PUFA	1	1	1
1	PolyIC	n-6 PUFA	2	5	1
2	PolyIC	n-3 PUFA	3	4	2
2	PolyIC	n-6 PUFA	4	3	3
3	PolyIC	n-3 PUFA	5	2	4
3	PolyIC	n-6 PUFA	6	1	1
4	PolyIC	n-3 PUFA	7	5	1
4	PolyIC	n-6 PUFA	8	4	2
5	PolyIC	n-3 PUFA	9	3	3
5	PolyIC	n-6 PUFA	10	2	4
6	PolyIC	n-3 PUFA	1	2	1
6	PolyIC	n-6 PUFA	2	1	2
7	PolyIC	n-3 PUFA	3	5	2
7	PolyIC	n-6 PUFA	4	4	3
8	PolyIC	n-3 PUFA	5	3	4
8	PolyIC	n-6 PUFA	6	2	1
9	PolyIC	n-3 PUFA	7	1	2
9	PolyIC	n-6 PUFA	8	5	2
10	PolyIC	n-3 PUFA	9	4	3
10	PolyIC	n-6 PUFA	10	3	4
11	Saline	n-3 PUFA	1	3	1
11	Saline	n-6 PUFA	2	2	2
12	Saline	n-3 PUFA	3	1	3
12	Saline	n-6 PUFA	4	5	3

Continued

Litter	Condition	Diet	Cage	Batch	Lane
13	Saline	n-3 PUFA	5	4	4
13	Saline	n-6 PUFA	6	3	1
14	Saline	n-3 PUFA	7	2	2
14	Saline	n-6 PUFA	8	1	3
15	Saline	n-3 PUFA	9	5	3
15	Saline	n-6 PUFA	10	4	4
16	Saline	n-3 PUFA	1	4	1
16	Saline	n-6 PUFA	2	3	2
17	Saline	n-3 PUFA	3	2	3
17	Saline	n-6 PUFA	4	1	4
18	Saline	n-3 PUFA	5	5	4
18	Saline	n-6 PUFA	6	4	1
19	Saline	n-3 PUFA	7	3	2
19	Saline	n-6 PUFA	8	2	3
20	Saline	n-3 PUFA	9	1	4
20	Saline	n-6 PUFA	10	5	4

Table 3.3: Design for follow up study. This design will allow one to balanced out litter effect, cage effect, batch effect and lane effects such that the confounding effects were minimized. One can also include the External RNA Controls Consortium (ERCC) spike in control to serves as an internal standard for additional level of control (**Jiang2011a**).

4 Conclusion

Schizophrenia has long been recognized as a genetic disorder. With the recent advancement in technology, 108 independent genetic loci associated with schizophrenia has finally been identified by the Schizophrenia Working group of Psychiatric Genomics Consortium (PGC) (**Ripke2014**). Based on these results, **Bulik-Sullivan2015** estimated that the common Single Nucleotide Polymorphisms (SNPs) account for 55.5% of the risk of schizophrenia, which is closed to the estimate from twins and family studies ($\sim 64 - 81\%$) (**sullivan2003schizophrenia**; **Lichtenstein2009**). However, in the presence of genetic-environment interaction, the heritability of schizophrenia might be overestimated (**zuk2012mystery**), which might suggest that the SNP-heritability estimated by **Bulik-Sullivan2015** might be too high. Additionally, as rare mutations (**purcell2014polygenic**), copy number variation (CNV) (**Szatkiewicz2014**) and structural variants (**Walsh2008**) all seems to contribute to risk of schizophrenia, it is more likely that **Bulik-Sullivan2015** has overestimated the SNP-heritability of schizophrenia.

Considering that **Bulik-Sullivan2015** only considered a limit number of condition in their binary trait simulation, it is critical to perform a comprehensive simulation to investigate the performance of LD SCore regression (LDSC). It will also be beneficial to have another SNP-heritability estimation tool to see if the same estimation can be reached.

Herein, we have developed SNP HeRitability Estimation Kit (SHREK), a

programme to estimate the SNP-heritability from summary statistics of Genome-Wide Association Study (GWAS). Our simulation results suggest that SHREK provided a more robust estimate for oligogenic traits and for binary traits when no confounding variables was present. Most importantly, when applying SHREK and LDSC to estimate the SNP-heritability of schizophrenia, much smaller estimates (table 2.5) are obtained. The LDSC estimation is more than 2 fold smaller than the one reported by **Bulik-Sullivan2015** Upon contacting the author (**Bulik-Sullivan2015c**), it is found that there is change in definition for one of the parameter in LDSC, which caused a drastic difference in the estimates produced by the two different version of LDSC.

First and foremost, this suggested that without the "correct" definition for the parameters, the LDSC estimates might be relatively volatile. In comparison, there is no such ambiguity for SHREK. However, if there are unadjusted confounding factors within the data, then the performance of SHREK, which does not have a function to handle confounding factors, might also be affected. Therefore, further development for both LDSC and SHREK are required before an accurate and robust estimates of SNP-heritability can be obtained when individual genotypes are unavailable. Meanwhile, if the confounding effect are adjusted, then SHREK might be superior to LDSC in the estimation of SNP-heritability for binary traits or oligogenic traits.

Finally, the new estimates from LDSC and SHREK indicate that the common SNPs included in the PGC GWAS can account for no more than 20% of the liability risk of schizophrenia. Hence, to fully understand the genetic influence in the etiology of schizophrenia, the field should invest more in sequencing and epigenetic studies for the identification of rare mutations and epigenetic changes associated with schizophrenia instead of increasing the sample size of schizophrenia GWAS studies.

Although schizophrenia is a genetic disorder, it was estimated that the shared or common environmental influences on liability to schizophrenia is around 11% with a confidence interval ranging from 3%-19% (**sullivan2003schizophrenia**), suggesting that they also have an important role in the etiology of schizophrenia. In view of this, we performed a RNA Sequencing analysis to study the effect of maternal immune activation (MIA), an environmental influence that is associated with schizophrenia, and the effect of n-3 polyunsaturated fatty acid (PUFA) rich diet in the gene expression pattern of mouse cerebellum.

Although no significant gene is detected when comparing the effect of MIA in mouse receiving the control diet, it is observed that MIA can also affect candidate gene sets that were reported to be associated with schizophrenia. The converging evidence suggest that both the genetic and environmental factors might be acting upon the same functional pathway. Given this information, the study of environmental factors might also provide insight to how genetic influence the risk of schizophrenia. Therefore, it might also be beneficial for one to not only focus on the genetic studies of schizophrenia, but also to study the effect of environmental influence in schizophrenia.

Finally, although we have identified *Sgk1* as a possible mediator of the effect of n-3 PUFA rich diet in the reduction of schizophrenia-like phenotype, more functional analysis and additional samples are required before we be certain of the role of *Sgk1* in the reduction of schizophrenia-like behavior. However, the current study do serves as an important stepping stone for future studies.

4.1 Schizophrenia: Future Perspectives

We are now entering a new era of sequencing, where a wide variety of tools are at our disposal to identify different genetic variations associating with a disease. It

is foresaw that the ability to investigate the whole genome/exome at a per base resolution will allow for the identification of more genetic variations such as the rare mutations, that are associated with schizophrenia.

With the sophistication of technologies, whole genome sequencing can now be performed with the HiSeq X Ten system, costing less than $1,000, allowing for more data to be generated. While the technology advancement have allowed a massive increase in available sequencing data, the bioinformatic analysis are now becoming the bottleneck of genetic research. For example, the alignment of sequence read to low complexity sequence or low-degeneracy repeats remains challenging and error prone, having a negative impact to the quality of the results (**Sims2014**). New sequencing technology such as Oxford Nanopore, which can provide extra long-reads, might help to make alignment easier by providing extra information for each individual reads. However, the Oxford Nanopore is still under development and has a relatively high error rate (**Mikheyev2014**). 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. For complex disease such as schizophrenia, a large amount of susceptibility loci can be observed throughout the genome, yet estimation of the functional impact can only be performed on variants located within the exomic regions. The development of ENCODE project (**ENCODEProjectConsortium2012**) and Genotype-Tissue Expression (GTEx) project (**Consortium2015**) have helped to provide reference point for the annotation of genetic variations in the intergenic regions. However, the functional impact of many genetic variation in the genome remains unknown. Only through the tireless effort of the molecular biologist can sufficient information be gained to allow for the detail annotation of the data.

Moreover, epigenetic studies in schizophrenia (**Wockner2014**; **Nishioka2012**) have identified genes with differential DNA methylation patterns associated with

schizophrenia, whereas etiology studies have reported the possibility of interaction between prenatal infection and genetic variation in risk of developing schizophrenia (**Tienari2004**; **Clarke2009**), suggesting the possible involvement of epigenetics and $G \times E$ interaction in the etiology of schizophrenia. Therefore, it is important to combine information of multiple dimensions – including genetic, CNV, epigenetic changes, genetic and environmental interaction, expression, structural properties and spatial organization of chromosomes – can we understand the etiology of schizophrenia.