Understanding How Genetics and Environments Shape the Development of Schizophrenia

Choi Shing Wan

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

Department of Psychiatry
University of Hong Kong
Hong Kong
January 4, 2016

Abstract

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

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

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

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

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

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

(550 words)

Declaration

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

Signed.....

Choi Shing Wan

Acknowledgements

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

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

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

THANK YOU!

Abbreviations

ARC	neuronal activity-regulated cytoskeleton-associated protein.
bp	base pair.
CEU	Northern Europeans from Utah.
CI	confidence interval.
cM	centiMorgan.
CNV	copy number variation.
DSM	Diagnostic and Statistical Manual of Mental Disorders.
DZ	dizygotic.
ECM	extracellular matrix.
ERCC	External RNA Controls Consortium.
GC	Genomic Control.
GCTA	Genome-wide Complex Trait Analysis.
GD	Gestation Day.
GO	Gene Ontology.
GRM	Genetic Relationship Matrix.
GWAS	Genome Wide Association Study.
IBD	identity by descent.
IL-6	Interleukin-6.
IQ	Intelligence Quotient.
kb	kilobase.
KEGG	Kyoto Encyclopedia of Genes and Genomes.
LD	Linkage Disequilibrium.
LDSC	LD SCore regression.
LRT	likelihood ratio test.
maf	minor allele frequency.
MAPK	mitogen-activated protein kinase.
mb	megabase.
MHC	major histocompatibility complex.
MIA	maternal immune activation.
MLM	mixed linear model.

mRNA	messenger RNA.
MSE	mean squared error.
MZ	monozygotic.
NGS	next generation sequencing.
NMDA	N-methyl-D-aspartate.
PCA	principle component analysis.
PET	positron emission tomography.
PGC	Psychiatric Genomics Consortium.
PI3K	phosphatidylinositol 3-kinase.
PolyI:C	polyriboinosinic-polyribocytidilic acid.
PSD	postsynaptic density.
PUFA	polyunsaturated fatty acid.
QC	quality control.
QQ-plot	quantile-quantile Plot.
REML	restricted maximum likelihood.
RIN	RNA integrity number.
rt-PCR	real time PCR.
SCZ	schizophrenia.
SE	standard error.
SHREK	SNP HeRitability Estimation Kit.
SNP	Single Nucleotide Polymorphism.
SVD	Singular Value Decomposition.
tSVD	Truncated Singular Value Decomposition.
WHO	World Health Organization.
YLD	years lost due to disability.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Schizophrenia

Schizophrenia is a devastating psychiatric disorder affecting approximately 0.3–0.7% of the population worldwide (American Psychiatric Association, 2013). According to the Diagnostic and Statistical Manual of Mental Disorders (DSM)-V, which is 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 (Knapp, Mangalore, and Simon, 2004). Schizophrenia patients also have an increased tendency to commit suicide (Saha, Chant, and Mcgrath, 2007). Based on an World Health Organization (WHO) re-

port, 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

Table 1.1: Top 20 leading causes of YLD calculated by WHO in year 2012. Schizophrenia was considered as one of the top 20 leading causes of YLD (World Health Organization, 2013).

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

severity, schizophrenia has drawn much attention from the research community to delineate disease etiology and mechanisms, and identify risk factors associated with schizophrenia. Ultimately, the goal of schizophrenia research is to identify effective treatment(s) to improve the quality of life of patients.

1.2 Understanding Disease Etiology

An important first step in schizophrenia research is to understand whether if genetic or environmental variation contribute more to the disease etiology. A measure of the relative contribution of genetic and environmental influences to individual differences in the liability to a disorder is *heritability*. There are two definitions of heritability: 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 (Falconer and Mackay, 1996)

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

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:

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

One key feature of heritability is that it is a *ratio* of *population* measure-

ments 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 (Bouchard, 2013). 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 (Bouchard, 2013).

The definition of heritability becomes more complicated when we take into account different forms of genetic effects; this leads to the concept of narrow sense heritability.

1.2.2 Narrow Sense Heritability

The effects of genes are not always additive but can differ depending on the other gene at the same locus (dominance) or genes at different loci (epistasis). As a result, one can partition the total genetic variance into variance due to additive genetic effects (σ_A^2), variance due to dominant genetic effects (σ_D^2), and variance due to other epistatic genetic effects (σ_I^2), as follows:

$$\sigma_G^2 = \sigma_A^2 + \sigma_D^2 + \sigma_I^2$$

As individuals only transmit one copy of each gene at a single genetic locus to their offspring, relatives other than full siblings and identical twins will only share a maximum of one gene for each locus. Considering that dominance and epistatic genetic effects are interactive effect, which usually involve more than one gene, these effects are unlikely to contribute substantially to the resemblance be-

tween relatives other than monozygotic twins and full siblings (Visscher, William G Hill, and Wray, 2008). On the other hand, the additive genetic effects are usually transmitted from parent to offspring, thus it is more useful to consider the narrow sense heritability (h^2) which only includes the additive genetic effects, when predicting parent-offspring resemblance:

$$\begin{aligned} h^2 &= \frac{\sigma_A^2}{\sigma_P^2} \\ h^2 &= \frac{\sigma_A^2}{\sigma_G^2 + \sigma_E^2} \end{aligned} \tag{1.1}$$

To obtain the additive genetic effect, we can first consider the genetic effect of a parent to be $G_p = A + D$. As only half of the additive effect were 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 by random and D' is the non-additive genetic effect in the offspring. If we then consider the parent offspring covariance, we will get

$$\begin{aligned} \text{Cov}_{\text{OP}} &= \sum(\frac{1}{2}A + \frac{1}{2}A' + D')(A + D) \\ &= \frac{1}{2}\sum A^2 + \frac{1}{2}\sum AD + \frac{1}{2}\sum A'(A + D) + D'(A + D) \\ &= \frac{1}{2}V_A + \frac{1}{2}\text{Cov}_{AD} + \frac{1}{2}\text{Cov}_{A'A} + \frac{1}{2}\text{Cov}_{A'D} + \text{Cov}_{D'A} + \text{Cov}_{D'D} \end{aligned} \tag{1.2}$$

Under the assumption of random mating, A' should be independent from A and D. Moreover, as D' was specific to the child, it should be independent from A and D, with the covariance between the additive genetics and non-additive genetics being zero (Falconer and Mackay, 1996). Thus, eq. (1.2) becomes

$$\begin{aligned} \text{Cov}_{\text{OP}} &= \frac{1}{2}V_A + \text{Cov}_{AD} \\ &= \frac{1}{2}V_A \end{aligned} \tag{1.3}$$

Now if we assume the variance of phenotype of the parent and offspring were the

same, then using eq. (1.3), we can obtain the narrow-sense heritability as

$$h^2 = \frac{1}{2}\frac{V_A}{\sigma_P^2} \tag{1.4}$$

In the simple linear regression equation $Y = X\beta + \epsilon$, the regression slope can be calculated as

$$\beta_{XY} = \frac{\mathrm{Cov}_{XY}}{\sigma_X Y} \tag{1.5}$$

which resemble eq. (1.4). Therefore, we can calculate the narrow sense heritability as

$$h^2 = 2\beta_{OP} \tag{1.6}$$

where β_{OP} is the slope of the simple linear regression regressing the phenotype of an offspring to the phenotype of *one* of its parents. We can further generalize eq. (1.6) to all possible relativeness

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

where r is the relativeness of X and Y.

A key assumption in this calculation is that only additive genetic factors are shared among relatives. However, this is very unlikely to be entirely true as relatives do tends 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) and we shall discuss the partitioning of variance in the later sections.

Nonetheless, eq. (1.7) provide a simple example to the calculation of the narrow sense heritability. However, in the case of discontinuous trait (e.g. disease status) the calculation becomes more complicated because the variance of the phenotype was dependent on the population prevalence. As eq. (1.7) does not account for the trait prevalence, it cannot be directly applied to discontinuous traits. In order to perform heritability estimation on discontinuous trait, the concept of liability

threshold model proposed by Falconer, 1965 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 follow a normal distribution as is the case of many quantitative traits (Visscher, William G Hill, and Wray, 2008). The variance of phenotype can therefore be calculated as the variance under the normal distribution. However, such is not the case for disease like schizophrenia where only a dichotomous disease status ("affected" and "normal") are obtained. The variance of these phenotypes are therefore more difficult to obtain.

Falconer (1965) proposed the liability threshold model, which suggests that these discontinuous traits also follow a continuous distribution with an additional parameter called the "liability threshold". Under the liability threshold model, the discontinuous traits are assumed to be affected by combination of multitude of genetics and environmental factors, each with small effects. The main difference is that the phenotype of an individual is determined by whether if the combined effects of these factors ("liability") are above a particular threshold ("liability threshold") (fig. 1.1), e.g. 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 trait by comparing the mean liability of the general population when compared to the relatives of the affected individuals. For example, if we consider a single

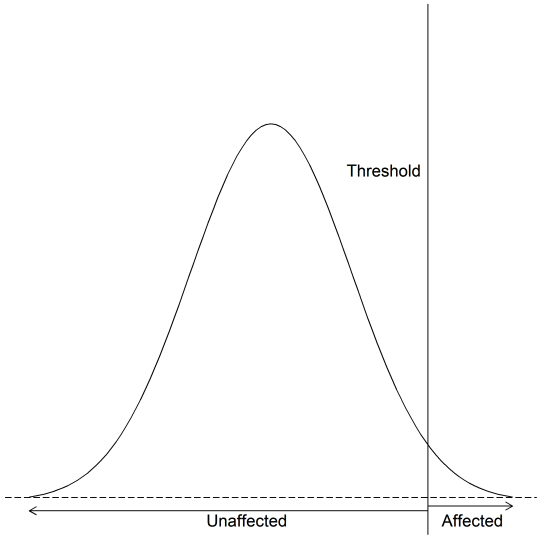


Figure 1.1: The liability threshold model. Only when an individual has a liability above the liability threshold will he/she be affected.

threshold model of a dichotomous trait, where

$$T_G = \text{Liability threshold of the general population}$$

$$T_R = \text{Liability threshold of relatives of the index case}$$

$$q_G = \text{Prevalence in the general population}$$

$$q_R = \text{Prevalence in relatives of the index case}$$

$$L_a = \text{Mean Liability of the index case}$$

by assuming both the liability distribution of the general population and the relative of the index case both follows the standard normal distribution, we can align the two distributions with respect to T_G and T_R. We can then calculate the mean liability of the index case L_a as $L_a = \frac{z_G}{q_G}$ where z_G is the density of the normal distribution at the liability threshold T_G. Then we can express the regression of relatives' liability on the liability of the index case as

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

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

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

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 their additive genetic effect of alleles, they also share some of the environmental factors such as diet and socio-economic status.

In order to discriminate the genetic factors from the shared environmental factors, adoption study can be carried out. An advantage of adoption studies is that if the child was separated from their family early after birth, then the shared environmental factors should be minimized. Any resemblance between the parent and child should be driven mainly by the shared genetic factors.

In the classical adoption study carried out by Heston (1966) 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. It was observed that there was an increased risk of schizophrenia in individuals born to schizophrenic mothers when compared to the control groups even-though they were brought up in a different environment as that of their mother. 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, any prenatal influence such as alcohol abuse and malnutrition

during pregnancy can confound the results. Therefore, an alternative method would be the twin studies, utilizing the genetic relationship between monozygotic (MZ) and dizygotic (DZ) twins.

MZ twins share all their genetic components (both additive (A) and non-additive (D) genetic factors) and common environmental factors (C). The only difference between the MZ twins is the non-shared environmental factors (E). For DZ twins, they share the same common environmental factors. However, only $\frac{1}{2}$ of their additive genetic factors and $\frac{1}{4}$ of their non-additive genetic factors are shared, whereas none of the non-shared environmental factors are shared among the twins (Rijsdijk and Pak C Sham, 2002).

In view of this, Falconer and Mackay, 1996 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.

By combining Falconer's formula and the concept of liability threshold model, Gottesman and Shields (1967a) 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 Sullivan, Kendler, and M. C. Neale (2003). Based on data obtained from 12 published schizophrenia twin studies, Sullivan, Kendler, and M. C. Neale (2003) found a much contribution from genetics on the liability of schizophrenia (81%, confidence interval (CI)=$73\% - 90\%$). Furthermore, in the large scale population based studies peroformed by Lichtenstein et al. (2009), the genetic contribution to schizophrenia was found to be 64%. Together, these results provide strong support for schizophrenia as a genetic disorder.

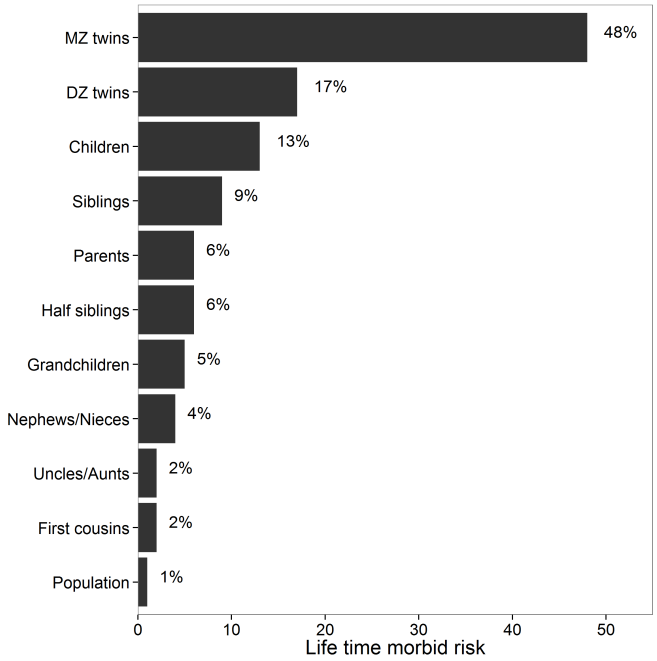


Figure 1.2: Lifetime morbid risks of schizophrenia in various classes of relatives of a proband. It was noted that the morbid risk of MZ twins were only 48%, much lower than one would expect if schizophrenia follows a Mendelian pattern. Reproduced with permission from journal (Riley and Kendler, 2006).

1.3 Schizophrenia Genetics

Although schizophrenia is highly heritable, little is known about the disease mechanism of schizophrenia and the genetic architecture of the disorder. However, 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 (Gottesman and Shields, 1967a; Gottesman and James Shields, 1982; Gottesman, 1991).

In view of this, Gottesman and Shields (1967a) proposed that schizophrenia follows 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, Risch (1990) proposed that the cause 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 (Risch, 1990).

This might explain the early inconsistent findings of linkage studies in schizophrenia (Harrison and Weinberger, 2005). As genetic data from large, multi-generational pedigrees with both affected and unaffected individuals are required, subject recruitment of linkage studies are challenging. With the small sample size and the relatively small effect expected in schizophrenia, it is difficult for linkage studies to identify the susceptibility loci. Other methods are therefore required to identify the susceptibility loci of schizophrenia.

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 (E S Lander et al., 2001). 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 (T. I. H. Consortium, 2005), 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 important for genetic research as it 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 together, a LD block is formed. By performing associ-

ation testing on SNPs representing majority of 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 common 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 methods such as chi-square test or logistic regression.

However, when a large number of SNPs were tested, the frequency of type I error increases (Peters et al., 2010). Multiple testing correction is therefore vital in the analysis of GWAS.

The simplest method for the correction of GWAS is to use the genome wide threshold (p-value $\leq 5 \times 10^{-8}$), 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 (M.-X. X. Li et al., 2011), which reduced the genome-wide threshold according to the LD structure.

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 (S Purcell, Cherny, and P C Sham, 2003). Similar to linkage studies, a larger sample size are

required to identify susceptible loci with a smaller effect.

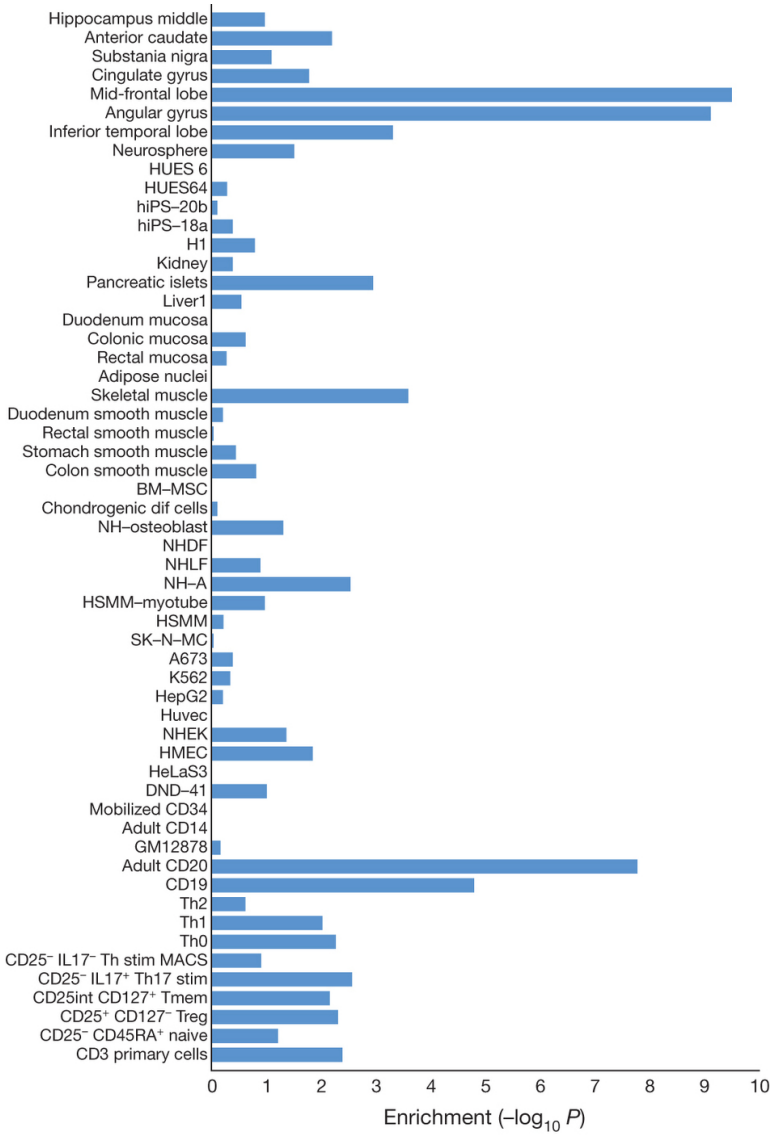
1.3.2.1 The Success of Psychiatric Genomic Consortium

Early GWAS in schizophrenia were largely disappointing, where no robust genetic markers associated with schizophrenia were identified. The main reason for the failure of the early GWAS is the relatively small sample size.

To overcome the problem of sample size, large consortium were formed such that genetic data from different research groups were combined and analyzed. Finally, in 2014, the Schizophrenia Working group of the Psychiatric Genomics Consortium (PGC) has conducted a multi-stage schizophrenia GWAS of up to 36,989 schizophrenia samples and 113,075 controls (Stephan Ripke, B. M. Neale, et al., 2014). A total of 128 linkage-disequilibrium-independent SNPs were found to exceeded the genome-wide significance (p-value $\leq 5 \times 10^{-8}$), corresponding 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, schizophrenia association were significantly enriched at enhancers active in brain and enriched at enhancers active in tissues with important immune functions (fig. 1.3) (Stephan Ripke, B. M. Neale, et al., 2014).

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 is not driven only by the MHC region. Considering the role of immune system in neural development (B. Zhao and Schwartz, 1998; Deverman and Patterson, 2009),

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 (Stephan Ripke, B. M. Neale, et al., 2014).



perturbation in the immune system is likely to disrupts the brain development. Therefore, the immune system might have an important role in the etiology of schizophrenia.

Despite the success of PGC schizophrenia GWAS, it is uncertain whether if all common variants associated with schizophrenia has been captured. With the unknown number of causal loci with moderate-to-small effect size, many SNPs associated with schizophrenia may be left undetected with the current sample size. However, it is also possible that the PGC schizophrenia GWAS has already captured all or near most of the SNPs associated with the disease. Therefore, estimating the contribution of these common SNPs to schizophrenia has important implications for future research strategy.

1.3.3 Contribution of Common SNPs

GWAS usually imposes a stringent genome wide significant threshold to avoid false positive findings. However, if individual SNPs have a small effect on the trait, the real association might be filtered. Therefore, to estimate the true contribution of common SNPs to a disease (SNP-heritability), it is important 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) (J Yang et al., 2011). The GRM represents the "genetic distance" between all individuals within the GWAS. Genetic relationship

between individual j and k is 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), they follow 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, J Yang et al. (2011) fitted the effects of all the SNPs as random effects by a mixed linear model (MLM)

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

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

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.

The main concept of GCTA is that instead of testing the associations for individual SNPs, effects of *all* SNPs are fit as random effects in a MLM, such that a single parameter can be estimated, i.e. the variance explained by all SNPs or SNP-heritability. Given the information of the GRM, J Yang et al. (2011) implemented the restricted maximum likelihood (REML) using the average information algorithm to estimate the σ_g^2 and σ_ϵ^2, where the REML is a form of maximum likelihood estimation that allows unbiased estimates of variance and covariance parameters. The SNP-heritability of the trait is then defined as $\frac{\sigma_g^2}{\sigma_g^2+\sigma_e^2}$.

Jian Yang, Benyamin, et al. (2010) were able to estimate the variance in height explained by SNPs from the height GWAS to be around 45%, much larger than previously reported 5%. Although the estimates was still less than 80%, which is the expected heritability of height, Jian Yang, Benyamin, et al. (2010) was able to demonstrated that one possible source of "missing heritability" might were due to incomplete LD. By considering incomplete LD, Jian Yang, Benyamin, et al. (2010) estimated that the proportion of variance explained by causal variants can be as high as 0.84 with standard error (SE) of 0.16, which is closer to the heritability of height.

An limitation of GCTA is that genotype data are required to calculate the GRM. For studies which only summary statistics are available, for example, the meta-analysis of schizophrenia, GCTA analysis cannot be performed.

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 (Zheng, Freidlin, and Gastwirth, 2006).

However, the assumption of a small number of causal SNPs might not be true, especially for complex disease such as schizophrenia. Through careful simulation, Jian Yang, Weedon, et al. (2011) 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.

Based on the work of Jian Yang, Weedon, et al. (2011), B. K. Bulik-Sullivan et al. (2015) hypothesized that the probability of a SNP to be a neighbor of the causal variant is higher if the SNP is in LD with a larger number of SNPs. However, this probability should be independent of the confounding factors such as population stratification and cryptic relatedness. Therefore, B. K. Bulik-Sullivan et al. (2015) developed the LD score. 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^2 of 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):

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

When confounding factors present in the study (e.g. population stratification), eq. (1.15) can instead 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 confounding bias.

By considering eq. (1.16) as a regression model, B. K. Bulik-Sullivan et al. (2015) observed that the contribution of common variants (the SNP heritability h^2) will be the slope of the regression, whereas the intercept minus one will represent the mean contribution of the confounding bias, such as population stratification. The LD SCore regression (LDSC) was then implemented using eq. (1.16) to delineate the contribution from confounding factors and common genetic variants B. K. Bulik-Sullivan et al. (2015).

To test whether LDSC can truly delineate the contribution from confound-

ing factors and common genetic variants, B. K. Bulik-Sullivan et al. (2015) performed a series of simulation. When the simulated traits are polygenic and no confounding factors are presented, the average LDSC intercept was close to one and the estimates were unbiased in all situation. Only when the number of causal variants was small will the standard error of the estimates become very large. On the other hand, when the GWAS was simulated with only the confounding factors such as population stratification, the intercept estimated was approximately equal to the GC inflation factor, with only a small positive bias in the regression slope.

Furthermore, when polygenic traits were simulated with confounding factors, the intercept of LDSC was approximately equal to the mean χ^2 statistic among the null SNPs, providing strong evidence that LDSC can partition the inflation in test statistic, even in the presence of both bias and polygenicity.

B. K. Bulik-Sullivan et al. (2015) then estimated the SNP heritability of schizophrenia, using the summary statistics from the PGC schizophrenia GWAS (Stephan Ripke, B. M. Neale, et al., 2014). It is estimated that the SNP heritability of schizophrenia is 0.555 with SE of 0.008 after adjusting for ascertainment bias. The estimated SNP heritability was lower than the heritability estimated from population based study (64% (Lichtenstein et al., 2009)) and twin studies (81% (Sullivan, Kendler, and M. C. Neale, 2003)). Therefore, it is possible for variants other than common SNPs to account for the heritability of schizophrenia.

1.3.3.3 Partitioning of Heritability

Another implication of LDSC is that it allows the partitioning of heritability, which helps to identify pathways that are associated with a trait.

Traditionally, functional enrichment analysis in GWAS only consider SNPs that passed the genome wide significance threshold. However, for complex traits

such as schizophrenia, it is possible that some of the SNPs with small effect size do not reach genome wide significance threshold at the current sample size. For example, in 2013, only 13 risk loci were detected using 13,833 schizophrenia samples and 18,310 controls (S Ripke et al., 2013). When the sample size increased to 34,241 schizophrenia samples and 45,604 controls in 2014, 108 risk loci were identified (Stephan Ripke, B. M. Neale, et al., 2014). Thus, by only selecting SNPs passing the genome wide significant thresholds, some of the risk loci might be ignored from the functional enrichment analysis.

To estimate whether a functional category is associated with the trait, LDSC utilize the summary statistic of all the SNPs included in the GWAS. The partitioning of the heritability is then calculated as

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

The main difference between eq. (1.17) and eq. (1.16) is that $\frac{h^2}{M} l_j$ is 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.

Using data from Stephan Ripke, B. M. Neale, et al. (2014), and functional categories derived from the ENCODE annotation (ENCODE Project Consortium, 2012), the NIH Roadmap Epigenomics Mapping Consortium annotation (Bernstein et al., 2010) and other studies, Finucane et al. (2015) attempted to identify functional categories that were most enriched in schizophrenia. Finucane et al. (2015) found that brain cell types and immune related cell types were most enriched in schizophrenia. Among the functional categories, the most enriched category in schizophrenia was the H3K4me3 mark in the fetal brain (table 1.2). As H3K4me3 is mostly linked to active promoters, it is possible that genes activated in fetal brain (e.g. genes related to brain development) are associated with schizophrenia, supporting the idea of schizophrenia as 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.(Finucane et al., 2015)

1.3.4 Rare Variants in Schizophrenia

The heritability estimates from B. K. Bulik-Sullivan et al. (2015) using the PGC GWAS data suggested that common SNPs have relatively less contribution ot the genetic predisposition of individuals to schizophrenia. Therefore, it is possible that rare variants might also contributes to the heritability of schizophrenia.

1.3.4.1 Copy Number Variation

A possible source of rare variants is copy number variations (CNVs). 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 (Feuk, Carson, and Scherer, 2006). Due to the length of these variants, the CNV might contain the entire genes and their regulatory regions, which might in turn contribute to significant phenotypic differences (Feuk, Carson, and Scherer, 2006).

Recently, Szatkiewicz et al. (2014) conducted a GWAS for CNV association with schizophrenia using the Swedish national sample (4,719 schizophrenia samples and 5,917 controls). Szatkiewicz et al. (2014) were able to identify association between schizophrenia and CNVs such as 16p11.2 duplications, 22q11.2 deletions, 3q29 deletions and 17q12 duplications. Through the gene set association analysis, calcium channel signaling and binding partners of the fragile X mental retardation protein were found to be enriched by CNV observed in schizophrenic samples (Szatkiewicz et al., 2014). Interestingly, in the PGC schizophrenia GWAS, association was observed genes encoding the voltage-gated calcium channel subunits. The results indicated that both common variants and CNV may be affecting the same set of pathways or gene sets in the etiology of schizophrenia.

Similarly, Walsh et al. (2008) also found that genes disrupted by structure variants in schizophrenic samples were significantly overrepresented in 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.

In general, CNVs associated with schizophrenia were rare ($\leq$ 12 in 4,719

samples (Szatkiewicz et al., 2014)) and have a relative large effect (e.g. odd ratio > 2 (Szatkiewicz et al., 2014; Walsh et al., 2008)).

1.3.4.2 Rare Single Nucleotide Mutation

Unlike CNV, which affects a large region, rare SNPs cannot be captured using current genotyping chips. Therefore, large scale association of rare SNPs was unavailable until the development of the next generation sequencing (NGS) technology. Recent progress in NGS has enabled the sequencing of the whole genome or exome, providing a per-base resolution, therefore allow for the identification of rare genetic variants without requiring "tagging".

Using exome sequencing, S M Purcell et al. (2014) sequenced the exome of 2,536 schizophrenia cases and 2,543 normal controls. S M Purcell et al. (2014) identified a common missense allele on *CCHCR1* in the MHC 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 (Fromer et al., 2014) and the postsynaptic density, all of which have been reported to be associated with schizophrenia in previous genetic studies (Stephan Ripke, B. M. Neale, et al., 2014).

The overlaps between the rare variant studies and the common variant studies suggest that both rare and common variants are likely to be acting upon the same pathway and are complementary to each other.

1.4 Environmental Risk Factors of Schizophrenia

Apart from genetic variants, another possible source of the "missing" heritability can come from interaction between the genetic and environmental risk factors. Although previous studies (Gottesman and Shields, 1967b) suggested that the non-additive genetic factors were unlikely to contribute to schizophrenia, the possibility of the involvement of gene-environmental interaction ($G \times E$) were not ruled out. Indeed, in the adoption study conducted by Tienari et al. (2004), it was 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 (Tienari et al., 2004). Moreover, using the national registers in Finland, Clarke et al. (2009) 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. Together, these findings support a mechanism of gene-environment interaction in the causation of schizophrenia.

Many environmental factors have been associated with schizophrenia, including prenatal infection (A S Brown and Derkits, 2010), winter birth (O'Callaghan et al., 1991), tobacco consumption (Kelly and McCreadie, 1999) and socio economic status (McGrath et al., 2008). These environmental factors are therefore potential targets for the study of $G \times E$ interaction. However, by and large, the prenatal infection is the largest environmental risk factor of schizophrenia. Furthermore, existing evidence indicated that there are an interaction between prenatal infection and genetic variations (Clarke et al., 2009). Investigation on how prenatal infection trigger schizophrenia and how it interacts with genetic variations in the development of schizophrenia are therefore important.

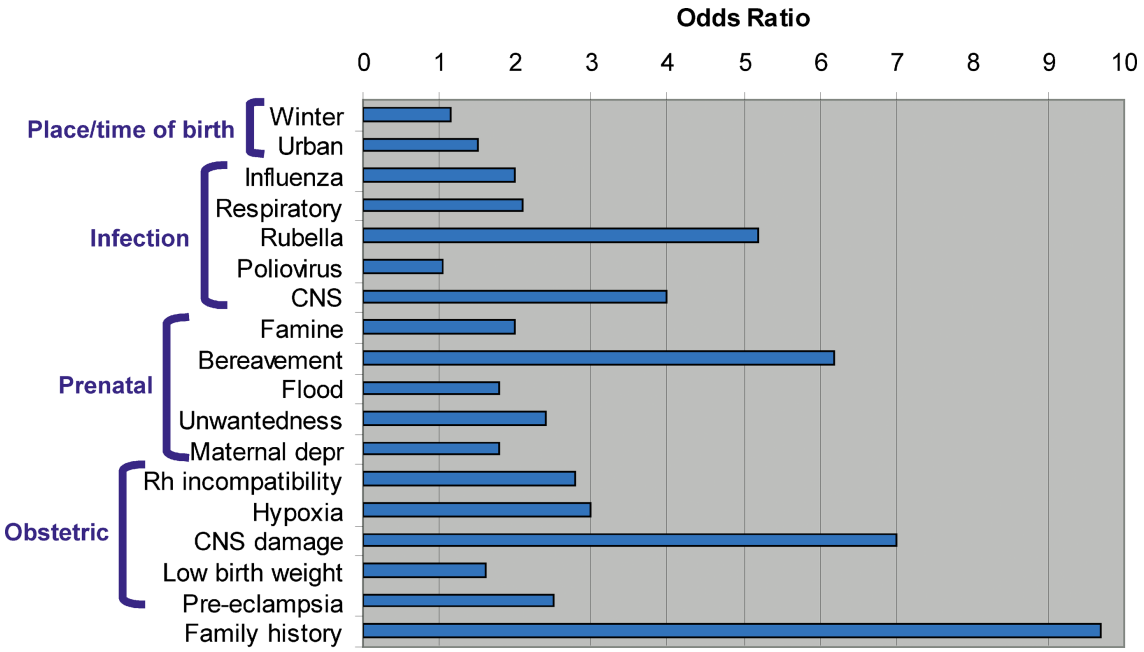


Figure 1.4: Risk factors of schizophrenia. It was observed that family history of schizophrenia was the largest risk factors. Risk of schizophrenia can be more than 9 times higher than the general population for individual with a family history of schizophrenia

1.4.1 Prenatal Infection

Prenatal infection is the single largest non-genetic risk factor of schizophrenia (fig. 1.4) (Sullivan, 2005). Initial clues of an involvement of prenatal infection in the etiology of schizophrenia comes from the observations that births during the winter and spring months; and births in urban areas were related to an increased risk of the disorder (A S Brown and Derkits, 2010). It was also observed that there was an increased risk of schizophrenia in individuals who were fetuses during the 1957 influenza epidemic (Mednick, 1988). As the chance of getting infectious diseases varies by season, and infectious diseases spread faster in urban regions due to higher population density, these evidences suggested that prenatal infection might be associated with schizophrenia.

Early studies of prenatal infection in schizophrenia mainly relies on ecological data such as influenza epidemics in the population to define the exposure status

(A S Brown and Derkits, 2010). The problem of these studies was that the exposure status was based solely on whether an individual was in gestation at the time of the epidemic without any confirmation of maternal infection during pregnancy. Therefore, the exposure status might be inaccurate and unreliable, leading to difficulties in replication of the findings from these epidemiological studies (A S Brown and Derkits, 2010). 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 (A S Brown and Derkits, 2010). It was found that the risk of schizophrenia increases as long as an individual's mother was infected by any form of infectious agents such as influenza, HSV-2 and *T.gondii* during gestation (A S Brown and Derkits, 2010). As various infectious agents increase the risk of schizophrenia, it leads to the hypothesis that maternal immune activation (MIA) (A S Brown and Derkits, 2010) rather than a particular infectious agent, is the source of risk factor. It was suggested that the maternal immune response to the infection might have disrupted the brain development in the fetus, thus leading to an elevated risk of schizophrenia (Garbett et al., 2012).

A great challenge in the study of MIA is that it is not possible to carry out controlled experiment on human samples due to ethical concerns. Thus, a popular alternative is to employ rodent models. 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 (American Psychiatric Association, 2013), which are not readily detectable in rodents. This raises challenge in diagnosing whether the rodent is schizophrenic or not. Therefore, the phenotypes of rodent samples are usually defined by the expression "schizophrenia-like" behaviours such as impaired prepulse inhibition, impaired working memory and reduced social interaction (U Meyer, Yee, and J Feldon, 2007).

However, the behavioral abnormality is not unique to schizophrenia, but

can also be observed in autistic samples. Moreover, risk of autism is also increased by MIA (Alan S Brown, 2012). As a result, studies using these rodent models were non-specific to schizophrenia or autism. However, the discussion of the etiology of autism and the similarity and difference between autism and schizophrenia is beyond the scope of the current thesis. Therefore, we will limit our discussion to schizophrenia.

A common rodent model in the study of effect of MIA is to use the viral analogue polyriboinosinic-polyribocytidilic acid (PolyI:C) to induce the maternal immune response during pregnancy. It was found that offspring exposed to PolyI:C displays phenotypes mirrors phenotypes observed in schizophrenia (Q. Li, C. Cheung, Wei, Hui, et al., 2009; Urs Meyer, Joram Feldon, and Fatemi, 2009; Q. Li, C. Cheung, Wei, V. Cheung, et al., 2010), such as deficiency in prepulse inhibition (Cadenhead et al., 2000). Because PolyI:C only induce the MIA without infecting the fetuses, the PolyI:C model provide strong evidence that MIA, instead of the specific infection, contributes to the increased risk of schizophrenia.

Smith et al. (2007) were able to demonstrate that a single injection of Interleukin-6 (IL-6) to the pregnant mouse can induce schizophrenia-like behaviour in the adult offspring. By eliminating the IL-6 from the maternal immune response using either genetic methods (IL-6 knock out) or with blocking antibodies, the behaviour deficits associated with MIA were not present in the adult offspring (Smith et al., 2007). The results indicate 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 (Oskviga et al., 2012; Garbett et al., 2012) 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 (Garbett et al., 2012). For example, genes with neuro-

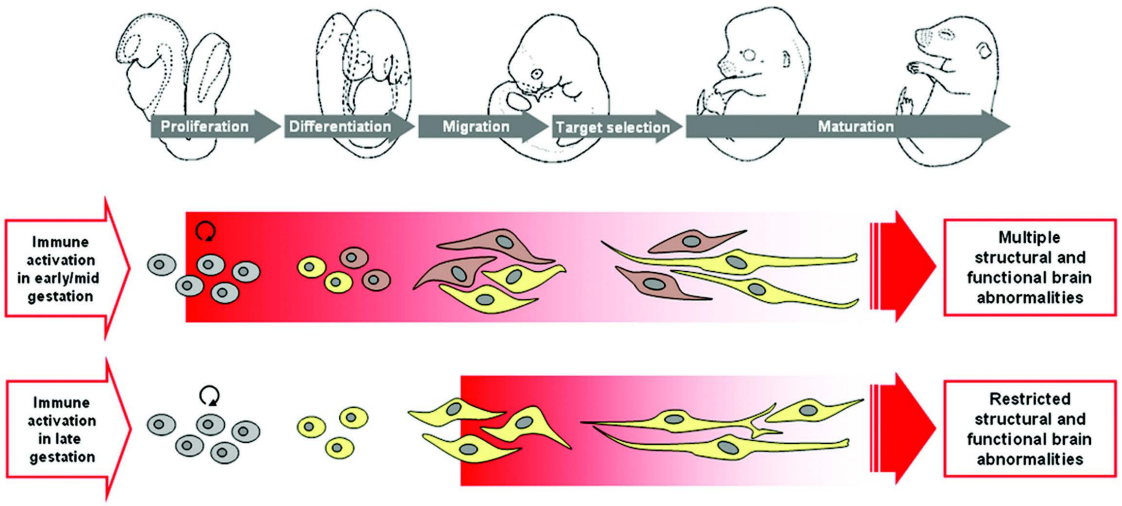


Figure 1.5: Hypothesized model of the impact of prenatal immune challenge on fetal brain development. Maternal infection in early/mid pregnancy may affect early neurodevelopmental events in the fetal brain, thereby influencing the differentiation of neural precursor cells (grey) into particular neuronal 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 (U Meyer, Yee, and J Feldon, 2007)

protective function such as crystallins might also have additional roles in neuronal differentiation and axonal growth (Garbett et al., 2012). 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, Garbett et al. (2012) 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 (Vuillermot et al., 2010). Specifically, MIA induces an early abnormality in specific dopaminergic systems such as those in the striatum and midbrian region (Vuillermot et al., 2010). Based on these observations, U Meyer, Yee, and J Feldon (2007) hypothesized that inflammation in the fetal brain during early gestation not

only can disrupt neurodavelopmental processes such as cell proliferation and differentiation, it also predispose the developing nervous system to additional failures in subsequent cell migration, target selection, and synapse maturation (fig. 1.5) (U Meyer, Yee, and J Feldon, 2007).

In a separate study by Giovanoli et al. (2013), mice were exposed to a lower dosage of PolyI:C during early gestation. A low dose of PolyI:C was selected as it only leads to restricted behavioral abnormalities in adulthood, thereby avoiding possible ceiling effects of the prenatal immunological manipulation on long-term brain and behavioral functions (Giovanoli et al., 2013). Offspring born were then left undisturbed or exposed to unpredictable stress during peripubertal development.

It was observed that offspring exposed to PolyI:C has an increased level of dopamine in the nucleus accumbens independent postnatal stress exposure. On the other hand, serotonin (5-HT) were 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 (Giovanoli et al., 2013). Giovanoli et al. (2013) therefore suggest that the prenatal insult serves as a "disease primer" that increase offspring's vulnerability to subsequent insults.

Together, these results supports the involvement of MIA in the development of schizophrenia. A S Brown and Derkits (2010) even estimated that one third of all schizophrenia cases could have been prevented shall all infection were prevented from the entire pregnant population.

One of the critical consideration in the study of MIA is the specific gestation period of vulnerability to infection-mediated disturbance (U Meyer, Yee, and J Feldon, 2007). 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, can be as early as the first trimester (U Meyer, Yee, and J Feldon, 2007). By reviewing existing MIA studies, U Meyer, Yee, and J Feldon (2007) suggested that 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 (U Meyer, Yee, and J Feldon, 2007). Subsequent MIA studies using the PolyI:C mouse model also support the hypothesis proposed by U Meyer, Yee, and J Feldon (2007), where it was observed that MIA early in gestation event exert a more extensive impact on the phenotype of offspring (Q. Li, C. Cheung, Wei, Hui, et al., 2009; Q. Li, C. Cheung, Wei, V. Cheung, et al., 2010).

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. As technology advances, RNA Sequencing technique can now be employed to examine the global messenger RNA (mRNA) expression changes in the brain of the adult offspring exposed to MIA during early gestation.

1.4.2 RNA Sequencing

Before the development of the NGS, the global 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.

When compared to microarray, RNA Sequencing has a number of advantages, most notably, because RNA Sequencing does not rely on specific probe hy-

bridization, it does not suffer from bias introduced by probe performances such as signal saturation, cross-hybridization, background noises and non-specific hybridization (S. Zhao et al., 2014).

Furthermore, in addition to differential expression analysis, alternative splicing analysis and de-novo transcript assembly can be readily performed on the same set of RNA Sequencing data. For microarray, de-novo assembly cannot be detected and specialized chips are required in order to perform alternative splicing analysis.

However, the analysis of RNA Sequencing is more complicated than microarray. The first consideration for the analysis of RNA Sequencing data is the sequence alignment. RNA sequencing generates sequence reads from the mRNA transcripts. Alignments have to be performed in order to quantify the expression level of the genes, where the sequence reads can either be aligned to the transcriptomes or the genome.

Sequence reads from RNA sequencing can be directly aligned to the transcriptomes as the reads are originated from the transcripts. However, multiple isoform can share the same exon. This leads to problem of mapping uncertainties, e.g. a single read can be aligned to multiple transcripts (B. Li and Dewey, 2011). The alignment uncertainly will have a negative impact to downstream analysis.

Another alignment method is to align the reads directly to the genome. However, as reads are originated from the mRNA transcripts, the intronic regions might be spliced out. Therefore, the alignment algorithm will have to "split" the reads in order to align them onto the corresponding exons. Some of the splice aware aligner includes TopHat2 (D. Kim et al., 2013), STAR (Dobin et al., 2013) and MapSplice (K. Wang et al., 2010).

Another difficulty in the analysis of RNA Sequencing data is the differential

expression analysis. In RNA Sequencing, the expression of a gene is represented by the number of reads aligned to the gene. Differential expression analysis then aims to identify statistical significant difference in the gene expression level between different conditions. However, unlike microarray, where the signal usually follows a normal distribution (Hoyle et al., 2002; Giles and Kipling, 2003), the distribution of the RNA Sequencing count data are more complicated.

Early RNA Sequencing experiment assumes the gene expression counts follows the Poisson distribution (Marioni et al., 2008), 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 (S Anders and W Huber, 2010). 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 (S Anders and W Huber, 2010; Robinson, McCarthy, and G K Smyth, 2010) or the beta negative binomial distribution (Trapnell et al., 2012), 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.

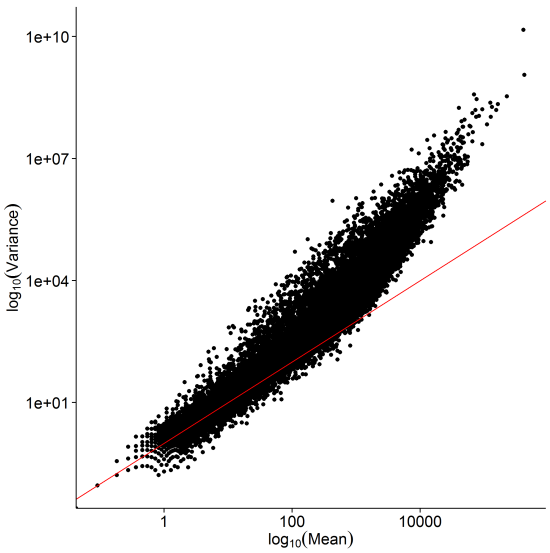


Figure 1.6: Over-dispersion observed in RNA Sequencing Count Data. If the RNA Sequencing count data follows the Poisson distribution, then the mean and variance of the data should be equal (follow the diagonal). However, it was observed that as the mean increases, the variance increases even more, suggesting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

1.5 Summary

In this thesis, we would like to perform extensive simulations to investigate the effect of different genetic architectures and sampling strategies in GWAS to the performance of LDSC, for example, the effect of extreme phenotype samplings. On the other hand, as LDSC might under-perform in certain scenarios (e.g. oligogenic traits) (B. K. Bulik-Sullivan et al., 2015), there is a need of developing an alternative algorithm that can be applied for various genetic architecture models with equal performance. Thus we developed an alternative algorithm for the estimation of SNP heritability using summary statistics from GWAS. Ultimately, we would like to re-estimate the SNP heritability of schizophrenia.

Currently, evidences suggest there might be an interaction between prenatal infection and genetic variations in the development of schizophrenia (Tienari et al., 2004; Clarke et al., 2009). Therefore, we hypothesize that the differential gene expression induced by MIA and genetic mutation might be acting upon the same

functional pathways/gene sets. To test our hypothesis, a RNA Sequencing study was performed to capture gene expression changes induced by early MIA events (Gestation Day (GD)9) in the mouse cerebellum using the PolyI:C mouse model. Enrichment analysis was then performed to investigate whether the differential expressed genes were enriched in gene sets associated with schizophrenia. Partitioning of heritability were also performed to investigate whether these gene sets contribute disproportionately to the SNP heritability of schizophrenia.

Furthermore, recent study from our lab suggested that n-3 polyunsaturated fatty acid (PUFA) rich diet might help to reduce the schizophrenia-like behaviour in mice exposed to early MIA insults (Q. Li, Leung, et al., 2015). Therefore, we also investigated the effect of n-3 PUFA rich diet in the gene expression pattern in the brain of the MIA samples.

To summarize, this thesis will be divided into three parts. The main focus of Chapter 2 is to investigate the effect of different sampling strategies and genetic architecture to the performance of LDSC. At the same time, SNP HeRitability Estimation Kit (SHREK), an alternative algorithm for the estimation of SNP heritability is also introduced. We also re-examined the SNP heritability of schizophrenia and other psychiatric disorders.

Chapter 3 describes our the RNA sequencing study on the effect of MIA and n-3 PUFA diet on gene expression of the mouse cerebellum.

Lastly, we summarize and conclude all findings in Chapter 4 and give future perspectives on genetic studies of schizophrenia.

2 Heritability Estimation

2.1 Introduction

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

The estimation of heritability on binary trait has always been complicated, as ascertainment bias correction is required when transforming the estimate on the binary scale to that on the liability scale. Nevertheless, the correction of ascertainment bias are nontrivial and can introduce bias to the estimates. For example, Golan, Eric S Lander, and Rosset (2014) observed that GCTA underestimates the heritability explained by common variants for case control studies. The magnitude of this bias is affected by a number of factors, including the population prevalence of the trait, the observed prevalence, the true underlying heritability and the number of genotyped SNPs (Golan, Eric S Lander, and Rosset, 2014). As for case control samples, B. K. Bulik-Sullivan et al. (2015) only investigated the performance of LDSC for traits with heritability of 0.8 and a population prevalence of either 0.1 or 0.01, more simulates are required in order to investigate whether if LDSC suffers

from the same bias as GCTA.

Finally, as noted by B. K. Bulik-Sullivan et al. (2015), when the number of causal variants is low, the standard error of the LDSC estimate become very large, indicating that LDSC is best suited to polygenic traits. This led to the need of developing an alternative algorithm that can be applied for various genetic architecture models.

Herein, SHREK, an alternative algorithm to LDSC for the estimation of SNP heritability using GWAS summary statistics was introduced. To examine the effect of different sampling strategies and genetic architectures on the performance of LDSC and SHREK, extensive simulation were performed. Moreover, to demonstrate that SHREK also works outside of simulated data, we also estimated the SNP heritability of schizophrenia and other psychiatric disorders using SHREK.

The work in this chapter were done in collaboration with my colleagues who have kindly provided their support and knowledges to make this piece of work possible. Dr Johnny Kwan, Dr Miaxin Li and Professor Sham have helped to lay the foundation of this study. Dr Timothy Mak has derived the mathematical proof for our heritability estimation method. Miss Yiming Li, Dr Johnny Kwan, Dr Miaxin Li, Dr Desmond Campbell, Dr Timothy Mak and Professor Sham have helped with the derivation of the standard error of the heritability estimation. Dr Henry Leung has provided critical suggestions on the implementation of the algorithm.

2.2 Methodology

2.2.1 Heritability Estimation

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

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

In a typical GWAS, regression is performed to test for an association between each SNP and the phenotype of interest:

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

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

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

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

A challenge arises when calculating the heritability from GWAS which only the summary statistic or p-value are available. When the genotype data is not accessible, estimation of $\mathrm{Var}(\boldsymbol{x})$ becomes infeasible, thus eq. (2.2) cannot be used.

In order to estimate the SNP heritability using GWAS summary statistics, the estimation of the relative effect size of each individual SNPs is necessary. It was observed that when $\boldsymbol{x}$ and $\boldsymbol{y}$ are standardized, β^2 equals to the coefficient

of determination (r^2). Thus, based on properties of the Pearson product-moment correlation coefficient:

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

where t follows the student-t distribution under the null and n is the number of samples. The r^2 can then be obtained by taking the square of eq. (2.3)

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

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

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

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

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

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

However, when there are LD between SNPs, the situation will become more complicated as the observed effect of each SNP is influenced by the neighboring SNPs in LD with it:

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

To account for the LD structure, we first assume our phenotype $\boldsymbol{y}$ and genotype $\boldsymbol{x} = (x_1, x_2, \ldots, x_m)^t$ are standardized such that

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

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

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

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

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

$$\boldsymbol{y} = \boldsymbol{\beta}^t \boldsymbol{x} + \epsilon \tag{2.8}$$

Since the phenotype is standardized with variance of 1, the SNP heritability can then be expressed as

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

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

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

where $\boldsymbol{H}$ is the variance of the "true" effect. Heritability can then be expressed as

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

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

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

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

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

In reality, the *observed* correlation is usually errors-prone. Therefore we define the *observed* correlation between SNP_i and the phenotype to be:

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

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

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

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

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

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

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

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

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

We have

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

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

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

If we define

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

we get

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

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

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

the SNP heritability is then defined as

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

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

2.2.2 Calculating the Standard error

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

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

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

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

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

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

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

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

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

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

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

By definition,

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

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

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

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

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

After some tedious algebra, we can get

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

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

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

χ^2 distribution will yield

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

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

2.2.3 Case Control Studies

The estimation of heritability in case control studies requires the correction of ascertainment bias. Therefore, the adjustment of estimates from eq. (2.17) is necessary to obtain an accurate estimation of SNP heritability. Under the liability threshold model, the disease status is determined by whether a standard normal liability is above the threshold (affected) or below threshold (unaffected). The mean values of disease liability in affected and unaffected individuals are given 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, respectively. Let v be the proportion of affected individuals in a sample, then the expected mean of liability in the sample is:

$$\begin{aligned} \mathrm{E}(L) &= v\frac{z}{K} + (1-v)(-\frac{z}{1-K}) \\ &= \frac{z(v-K)}{K(1-K)} \end{aligned} \tag{2.22}$$

and the square of mean liability is:

$$\mathrm{E}(L^2) = \frac{z^2(K^2 - 2vK - v)}{K^2(1-K)^2} \tag{2.23}$$

Then we get

$$\begin{aligned} \mathrm{Var}(L) &= \mathrm{E}(L^2) - (\mathrm{E}(L))^2 \\ &= \frac{z^2(v(1-v))}{K^2(1-K)^2} \end{aligned} \tag{2.24}$$

Therefore, in case-control sample, the summary statistic is attenuated from the standard scenario by a factor of $\mathrm{Var}(L)$. To correct for this bias, the following can be performed

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

2.2.4 Extreme Phenotype Sampling

GWAS provides unprecedented power to perform hypothesis-free 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 increases the statistical power (Guey et al., 2011). For example, by including only the samples from the top 5% and bottom 5% of the phenotype distribution, the power of the detection is the same as a study with random sampling design that has 4 times the sample size (Pak C Sham and Shaun M Purcell, 2014).

The extreme phenotype selection design increases the summary statistics by a factor of $\frac{V_{P'}}{V_P}$ where $V_{P'}$ is the trait variance of the selected sample and V_P is the trait variance of the general population (Pak C Sham and Shaun M Purcell, 2014). Thus, to adjust for the inflation, $\frac{V_P}{V_{P'}}$ is multiplied to eq. (2.14)

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

2.2.5 Inverse of the Linkage Disequilibrium matrix

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

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

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

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

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

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

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

Now consider the vector $\boldsymbol{B}$ is collected with some error $\boldsymbol{\epsilon}$ attached to it.

The solution to $\boldsymbol{Ax} = \boldsymbol{B}$ becomes:

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

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

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

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

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

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

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

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

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

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

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

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

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

By selecting an appropriate t, tSVD can effectively regularize the ill-conditioned matrix and help to find a reasonable approximation to $\boldsymbol{x}$. A problem with tSVD however is that it only works when matrix $\boldsymbol{A}$ has a well determined numeric rank (Hansen, 1987). That is, the performance of tSVD is optimal when there is a large gap between σ_k and σ_{k+1}. On the other hand, 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.

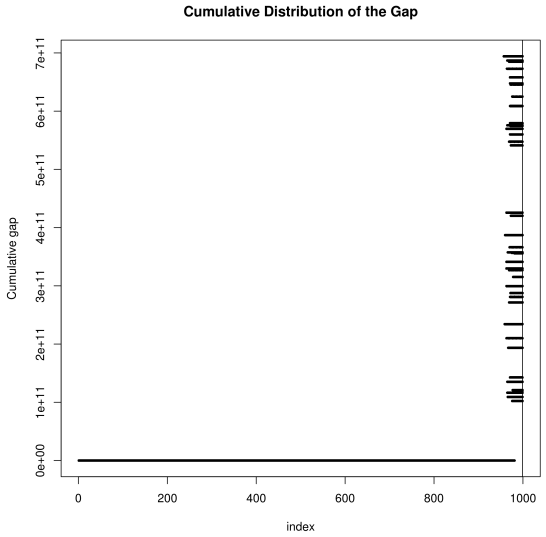
The easiest way to test if the matrix $\boldsymbol{A}$ has well-defined rank by calculating the "gap" in the singular values:

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

a large value (e.g "gap") usually indicates that the matrix has a well-defined rank.

Hence, simulation was carried out to investigate whether if LD matrix has a well-defined rank. 1,000 samples were randomly simulated from the HapMap (Altshuler et al., 2010) Northern Europeans from Utah (CEU) population with 1,000 SNPs randomly select from chromosome 22 using HAPGEN2 (Su, Marchini, and Donnelly, 2011). HAPGEN2 allow for the simulation of samples with LD structure similar to that observed in the reference panel. The LD matrix and its corresponding singular values were computed. The whole process were repeated 50 times and the cumulative distribution of the "gap" of singular values were plotted (fig. 2.1). It is

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



clearly shown that the LD matrix has a well-defined rank with a mean maximum "gap" of 466,198,939,298. Therefore the choice of tSVD for the regularization is appropriate.

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

2.2.6 Comparing Different LD correction Algorithms

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

estimated sample LD. The sample LD can then be represented as:

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

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

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

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

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

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

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

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

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

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

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

Again, simulations were performed to assess the performance of each individual correction methods. Firstly, 5,000 SNPs with maf ≥ 0.1 were randomly selected from chromosome 22 from the 1000 genome CEU haplotypes and were used as an input to HAPGEN2 (Su, Marchini, and Donnelly, 2011) to simulate 1,000 individuals. HAPGEN2 is a simulation tools which simulates new haplotypes as an

imperfect mosaic of haplotpyes from a reference panel and the haplotypes that have already been simulated using the *Li and Stephens* (LS) model of LD (N. Li and Stephens, 2003). This allows for the simulation of genotypes with LD structures comparable to those observed in CEU population. Of those 5,000 SNPs, 100 of them were randomly selected as the causal variants. Orr (1998) suggested that the exponential distribution could be used to approximate the genetic architecture of adaptation. As a result, effect sizes were simulated with an exponential distribution with $\lambda = 1$:

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

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

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

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

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

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

size of each super population form the 1,000 genome project, were simulated as a reference panel for the calculation of LD matrix. This is because in reality, the raw sample genotypes were unavailable and has to rely on an independent reference panel for the calculation of LD matrix. Thus, this simulation procedure should provide a realistic representation of the common usage of the algorithm.

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

1. Randomly select 5,000 SNPs with maf$>$ 0.1 from chromosome 22
2. Simulate 500 samples using HAPGEN2 and used as the reference panel
3. Randomly generate 100 effect sizes with eq. (2.41)
4. Randomly assign the effect sizes to 100 SNPs with heritability from 0 to 0.9 (increment of 0.1)
5. Simulate 1,000 samples using HAPGEN2 and calculate their phenotype according to eq. (2.42)
6. Perform heritability estimation using our algorithm with different LD correction algorithm
7. Repeat step 5-6 50 times

2.2.7 Comparison with Other Algorithms

After identifying the optimal LD correction algorithm, it is important to compare the performance of our algorithm to the other existing methods. Therefore, simulations were performed where quantitative and binary traits were simulated under different genetic architectures. The effect of sampling strategies, such as random

sampling and extreme phenotype selection, on the heritability estimation were also investigated.

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

2.2.7.1 Sample Size

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

2.2.7.2 Number of SNPs in Simulation

To reduce the computational cost for simulation, only 50,000 SNPs from chromosome 1 were used for simulation, which correspond to 200 SNPs within a 1 megabase

(mb) region.

2.2.7.3 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 (Su, Marchini, and Donnelly, 2011), using the 1000 genome CEU haplotypes as an input. As GWAS usually lack power in detecting rare variants (e.g. maf < 0.05), SNPs with maf < 0.05 were excluded.

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

For GCTA, the sample genotypes were provided to calculate the genetic relationship matrix. Sample phenotypes were also provided for GCTA to estimate the SNP heritability.

On the other hand, for LDSC and our algorithm, an independent 500 samples were simulated as the reference panel for the calculation of LD scores and LD matrix. The association between the genotype and phenotype were calculated using PLINK (Shaun Purcell et al., 2007). The summary statistics and the reference panel were then provided for LDSC and our algorithm to estimate the SNP heritability.

This simulation procedure should provide a realistic representation of the common usage of the algorithms.

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

1. Randomly select 50,000 SNPs with maf>0.05 from chromosome 1
2. Simulate 500 samples using HAPGEN2 to be served as a reference panel
3. Randomly generate k effect size with $k \in \{5, 10, 50, 100, 500\}$ following eq. (2.41), with heritability ranging from 0 to 0.9 (increment of 0.1)
4. Randomly assign the effect size to k SNPs
5. Simulate 1,000 samples using HAPGEN2 and calculate their phenotype according to eq. (2.42)
6. Perform heritability estimation using our algorithm, GCTA, LDSC with fixed intercept and LDSC with intercept estimation.
7. Repeat step 5-6 50 times
8. Repeat step 1-7 10 times

2.2.7.4 Extreme Effect Size

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

To simulate extreme effect size, 100 causal SNPs were simulated where m

of those account for 50% of all the effect sizes with $m \in \{1, 5, 10\}$. The effect sizes were then calculated as

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

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

1. Randomly select 50,000 SNPs with maf > 0.05 from chromosome 1
2. Simulate 500 samples using HAPGEN2 and used as the reference panel
3. Randomly generate 100 effect size where m has extreme effect, following eq. (2.43), with $m \in \{1, 5, 10\}$
4. Randomly assign the effect size to 100 SNPs
5. Simulate 1,000 samples using HAPGEN2 and calculate their phenotype according to eq. (2.42)
6. Perform heritability estimation using our algorithm, LDSC with fixed intercept, LDSC with intercept estimation and GCTA
7. Repeat step 5-6 50 times
8. Repeat step 1-7 10 times

2.2.7.5 Case Control Studies

In the simulation of case control samples, two additional parameters, the population prevalence (p) and observer prevalence (q), have to be taken into consideration. The

liability threshold model were used to model the samples. 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. (2.41). $\frac{1{,}000}{p}$ samples were then simulated and their phenotype were calculated using eq. (2.42). The phenotypes were then standardized and cases were defined as sample with phenotype passing the liability threshold with respect to p. An equal amount of controls were then randomly selected from samples with phenotype below the liability threshold.

In summary, the case control 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. (2.41) where $k \in \{10, 50, 100, 500\}$
4. Randomly assign the effect size to k SNPs
5. Simulate $\frac{1{,}000}{p}$ samples using HAPGEN2 and calculate their phenotype according to eq. (2.42)
6. Define case control status using the liability threshold and randomly select the same number of case and controls for statistic analysis
7. Perform heritability estimation using our algorithm, LDSC with fixed intercept, LDSC with intercept estimation and GCTA
8. Repeat step 5-7 50 times
9. Repeat step 1-8 10 times

2.2.7.6 Extreme Phenotype Sampling

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

From the 50,000 SNPs, 100 SNPs were randomly selected as the causal SNPs and their effect sizes were simulated using eq. (2.41). 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. (2.42). 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. 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 Pak C Sham and Shaun M Purcell (2014) were applied to the estimates from LDSC and GCTA. Finally, the heritability estimated based on different sampling strategies were compared. For each population, the whole process were repeated 50 times. In total, 10 independent populations were simulated. In summary, the following simulation procedures were used:

1. Randomly select 50,000 SNPs with maf$>$ 0.05 from chromosome 1
2. Simulate 500 samples using HAPGEN2 and used as the reference panel
3. Randomly generate 100 effect size following eq. (2.41), with heritability ranging from 0 to 0.9 (increment of 0.1)
4. Randomly assign the effect sizes to 100 SNPs
5. Simulate $\frac{500}{K}$ 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. (2.42) and were standardized
7. Top 500 and bottom 500 samples (ranked by phenotype) were selected, representing the extreme phenotype sample selection strategy
8. 1,000 samples were also randomly selected to represent the general random sampling strategy
9. Perform heritability estimation using our algorithm, GCTA, LDSC with fixed intercept and LDSC with intercept estimation.
10. Adjust the estimation from LDSC and GCTA by the extreme phenotype ad-

justment factor as proposed by Pak C Sham and Shaun M Purcell (2014)

11. Repeat step 5-10 50 times

12. Repeat step 1-11 10 times

2.2.8 Application to Real Data

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

The reference genome were downloaded from 1000 genome (hg19) (Project et al., 2012) and were converted to PLINK binaries using the PLINK --vcf function. The European super population was extracted which contains a total of 503 samples. Singleton and multi-allelic SNPs were filtered out from the reference panel. Cryptic relatedness between samples can inflate the LD due to increased allele sharing amongst relatives. It is therefore important to filter out related samples. Genotypes were first pruned, then the identity by descent (IBD) between samples were calculate using the PLINK option --genome. Sample pairs with relatedness ≥ 0.125 ($\approx$ third degree relatedness) were removed. In total, 446 samples remained after quality control. The LD score was calculated based on the 446 samples using a 1 mb window size. SNPs with maf < 0.1 were filtered out by default. 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 (Hinrichs et al., 2006) were performed to convert the genomic coordinates

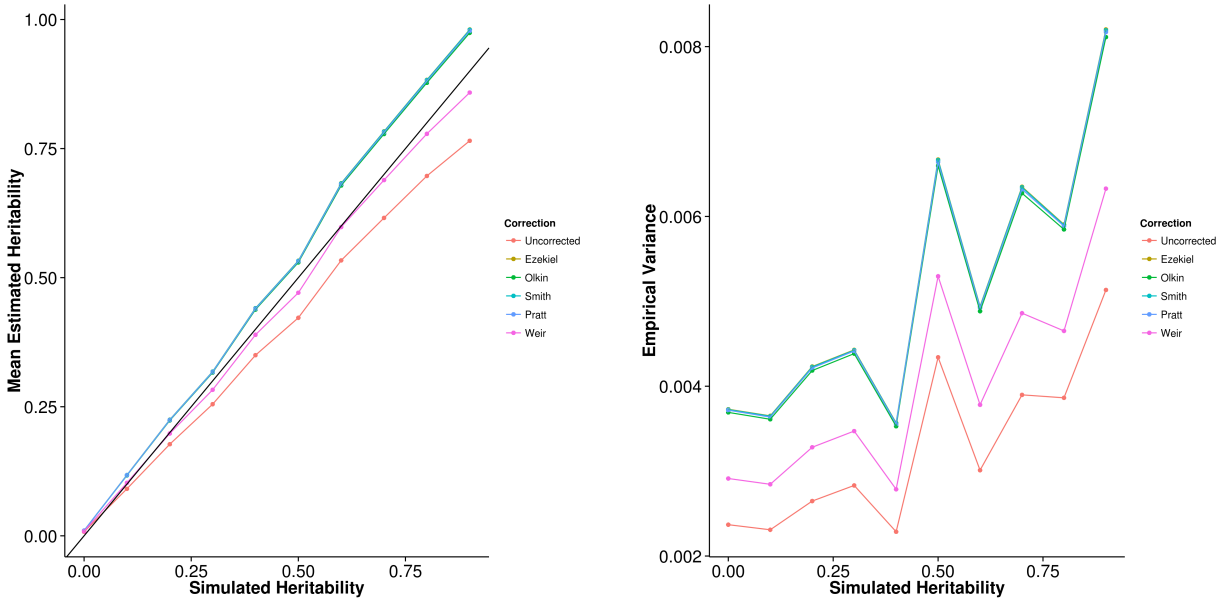
to genome version hg19. Due to difference in composition of the sex chromosome in male and female (e.g. XY in male, XX in female) and the lack of information on the male to female ratio, it is difficult to estimate the SNPs heritability on the sex chromosomes. Special care are required for the estimation of SNP heritability on the sex chromosome and this function has not been implemented in either SHREK or LDSC. Therefore, the SNP heritability 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 (B. K. Bulik-Sullivan et al., 2015).

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

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

2.3 Results

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



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

Figure 2.2: 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. (2.40) algorithms leads to an overestimation of heritability.

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. 2.2a). It is observed that when no bias correction was applied, the mean estimates biased downward, as expected.

On the other hand, all the correction methods, with the exception of the formula proposed by Weir and W G Hill (1980) (eq. (2.40)), result in upwardly biased estimates. This suggests that most of the bias correction algorithms have "over-adjusted" the sampling error, therefore leads to an inflation in the estimates. Based on the simulation results, it is concluded that the formula proposed by Weir and

W G Hill (1980) works best and are therefore selected as the default LD correction algorithm for SHREK.

2.3.2 Comparing with Other Algorithms

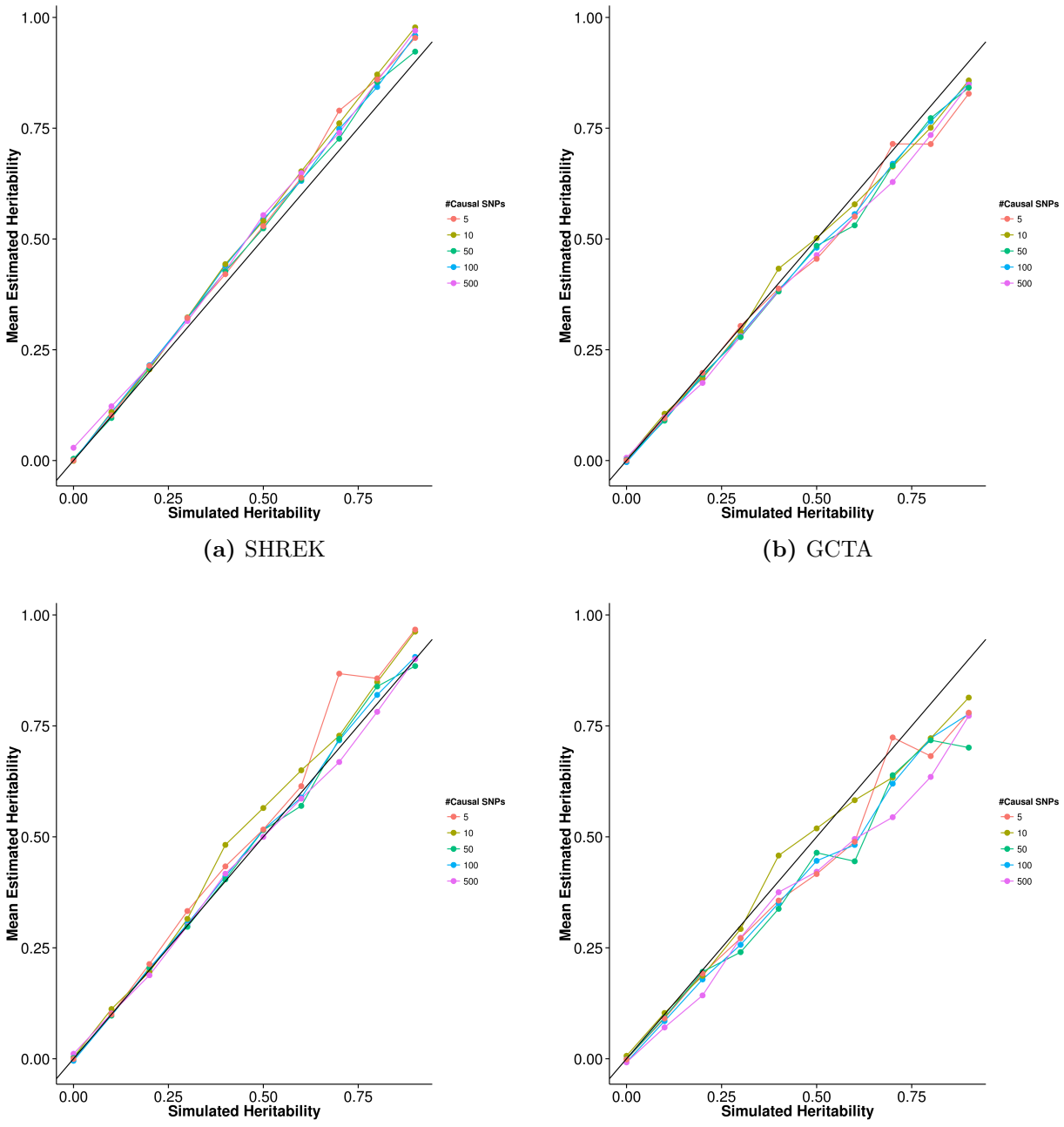
After the selection of LD correction algorithm, it is important to compare the performance of SHREK with the existing algorithms. Another aim of the current study is to investigate how different sampling strategies and genetic architectures influence the performance of LDSC. Therefore a series of extensive simulation analyses were performed.

2.3.2.1 Quantitative Trait Simulation

By varying the number of causal SNPs and the heritability of the trait, the performance of LDSC, SHREK and GCTA in the SNP heritability estimation of quantitative traits were investigated.

First, when comparing the mean estimates to the simulated heritability, a small upward bias is observed in the estimates from SHREK (fig. 2.3a). On the other hand, estimates from GCTA are moderately biased downward (fig. 2.3b), similar to the estimates from LDSC with intercept estimation (fig. 2.3d), but with a smaller variability. When the intercept was fixed, LDSC can accurately estimate the SNP heritability. And an upward bias can only be observed when the number of SNPs is small.

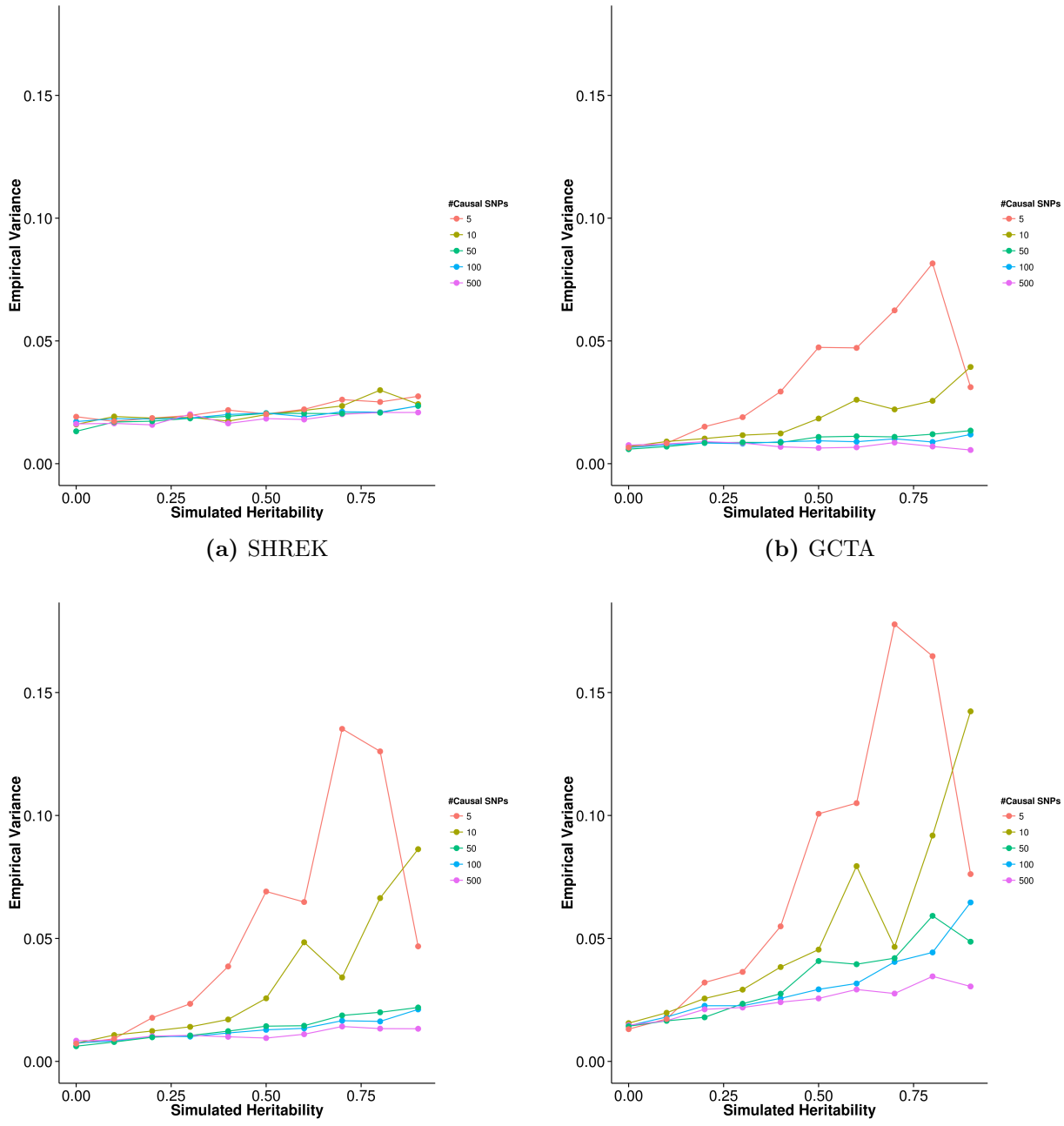
Secondly, the empirical variance of the estimates were also an important indicator of the performance of the algorithms. It is clear that empirical variance of the estimates from LDSC are sensitive to the number of causal SNPs (figs. 2.4c and 2.4d). When the number of causal SNPs decreases, the variance of the estimates increases, as reported by B. K. Bulik-Sullivan et al. (2015). Moreover, consistent



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

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

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



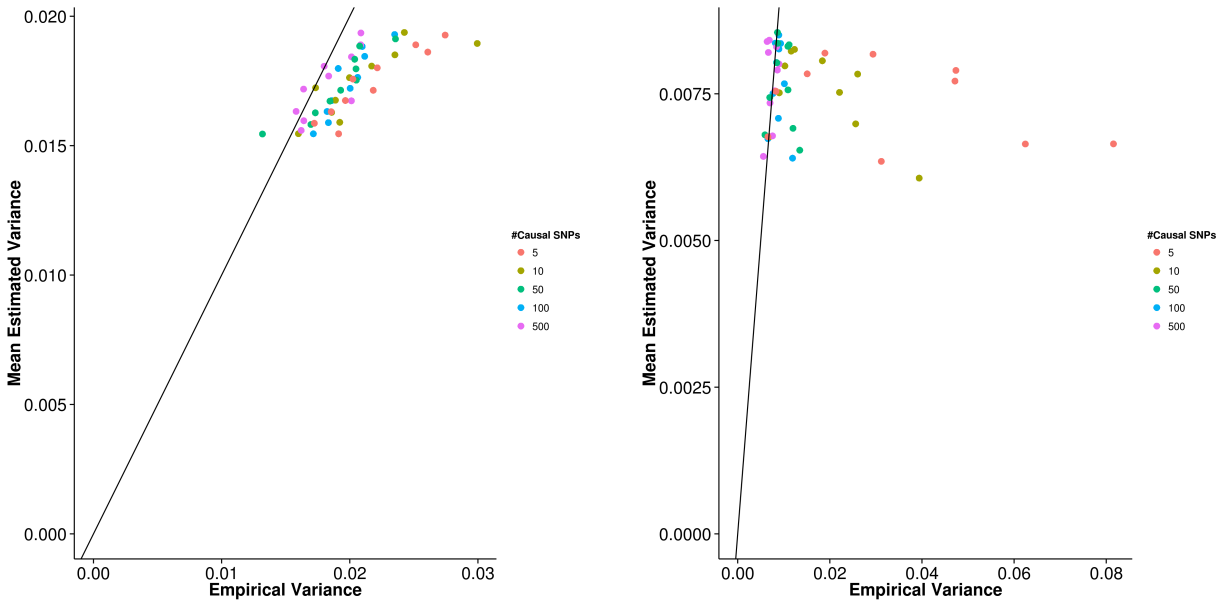
(a) SHREK

(b) GCTA

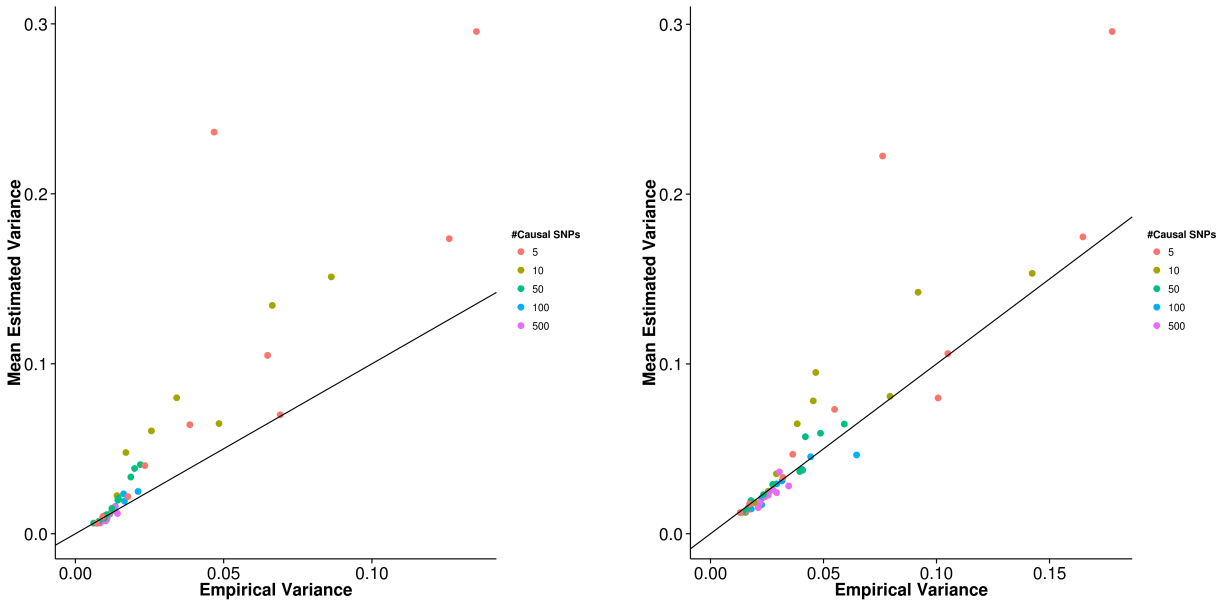
(c) LDSC with fix intercept

(d) LDSC with intercept estimation

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



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



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

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

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

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

with the results from B. K. Bulik-Sullivan et al. (2015), the intercept estimation increases the variance of the estimates from LDSC. Similarly, although GCTA has the lowest variance in its estimates, the variance increases when the number of causal SNPs decreases (fig. 2.4b). On the other hand, estimates from SHREK are relatively insensitive to the number of causal SNPs.

Finally, it is equally important for the algorithms to be able to estimate the variance of its estimates. It is observed that when the number of causal SNPs is large, GCTA can accurately estimates its variance (fig. 2.5b). However, when the number of causal SNPs decreases, GCTA underestimates the variance of its estimates. On the other hand, SHREK consistently underestimate the variance of its estimates (fig. 2.5a). 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. 2.5c and 2.5d), 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.

Taking into account of the bias and variance of the estimates, GCTA has

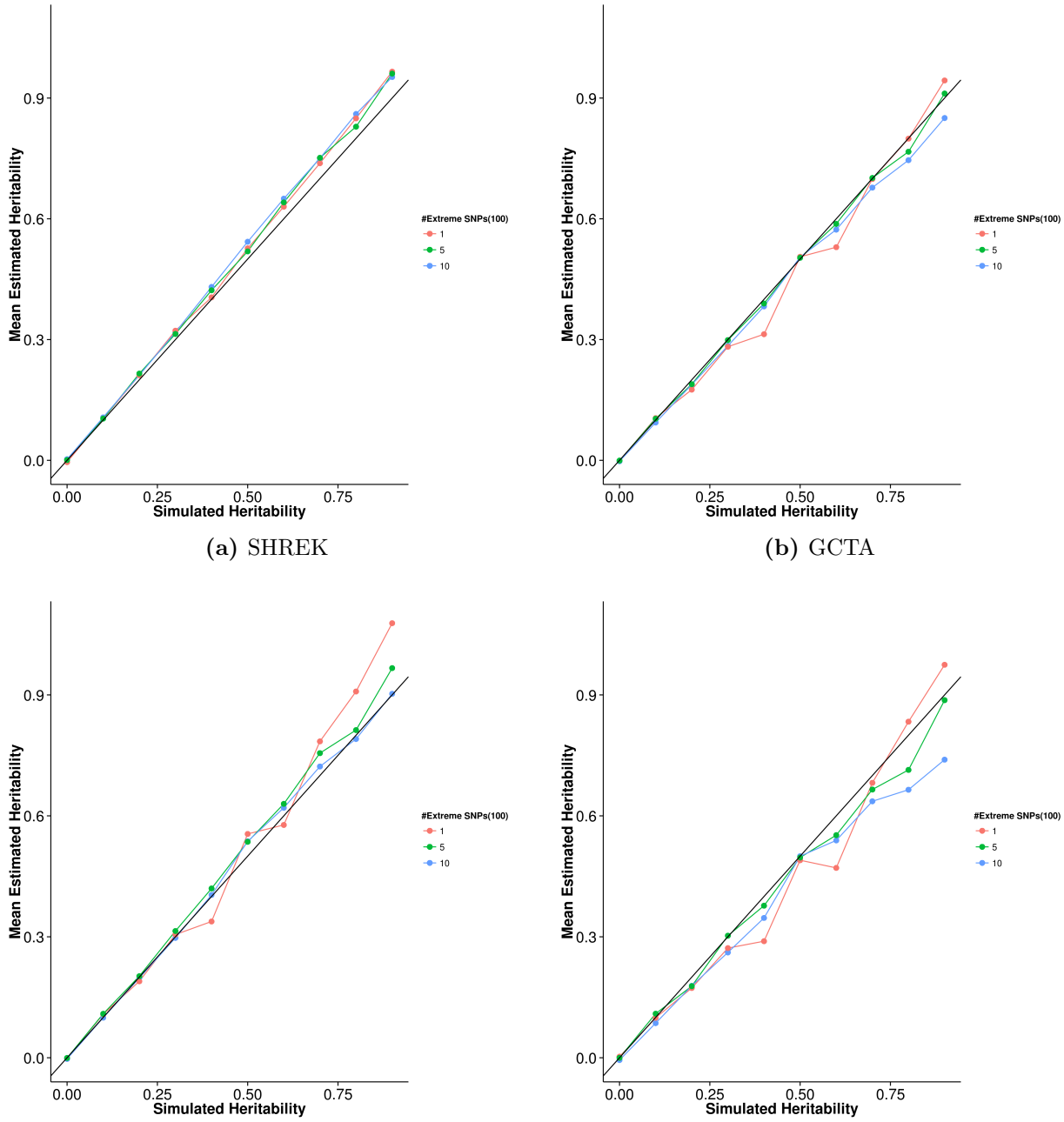
the best overall performance. On the other hand, when the number of causal SNPs is small, SHREK has a better performance when compared to LDSC, whereas LDSC performs better under polygenic condition. It is also observed that estimates from SHREK is the least sensitive to changes in the genetic architecture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

2.3.2.2 Quantitative Trait Simulation with Extreme Effect Size

It is possible for a polygenic trait to have some portion of causal variants with much larger effect size when compared to other causal variants. In order to investigate how extreme effect size in a small number of causal SNPs affect the SNP heritability estimation, 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. 2.6). However, when 1 of the causal SNPs was simulated with large effect, the mean estimates from LDSC and GCTA fluctuate (figs. 2.6b to 2.6d). The same fluctuation is not observed in SHREK (fig. 2.6a). Similarly, the empirical variance of the estimates (fig. 2.7) from GCTA and LDSC increase and fluctuate when only 1 of the causal SNPs was simulated 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.



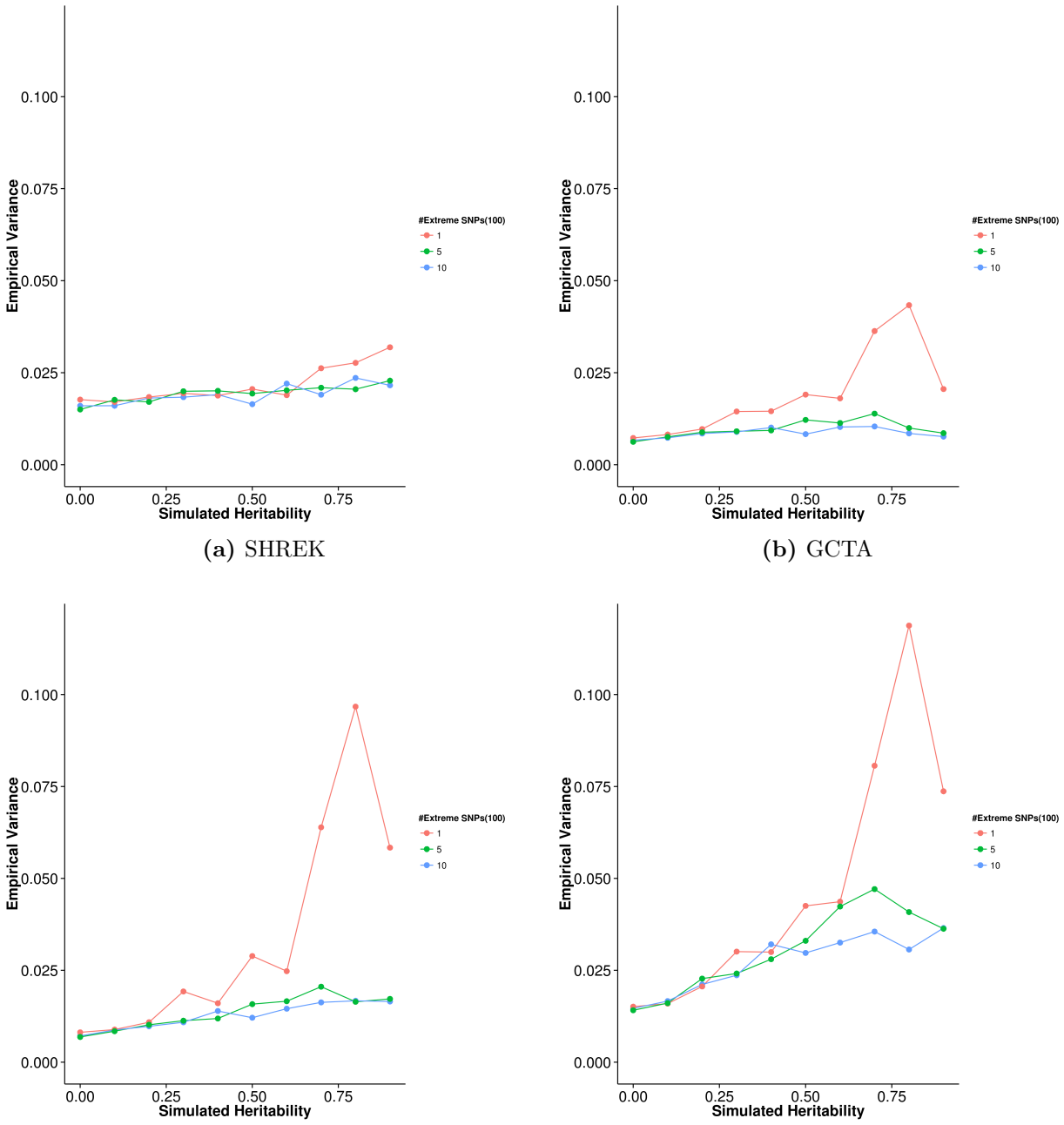
(a) SHREK

(b) GCTA

(c) LDSC with fix intercept

(d) LDSC with intercept estimation

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



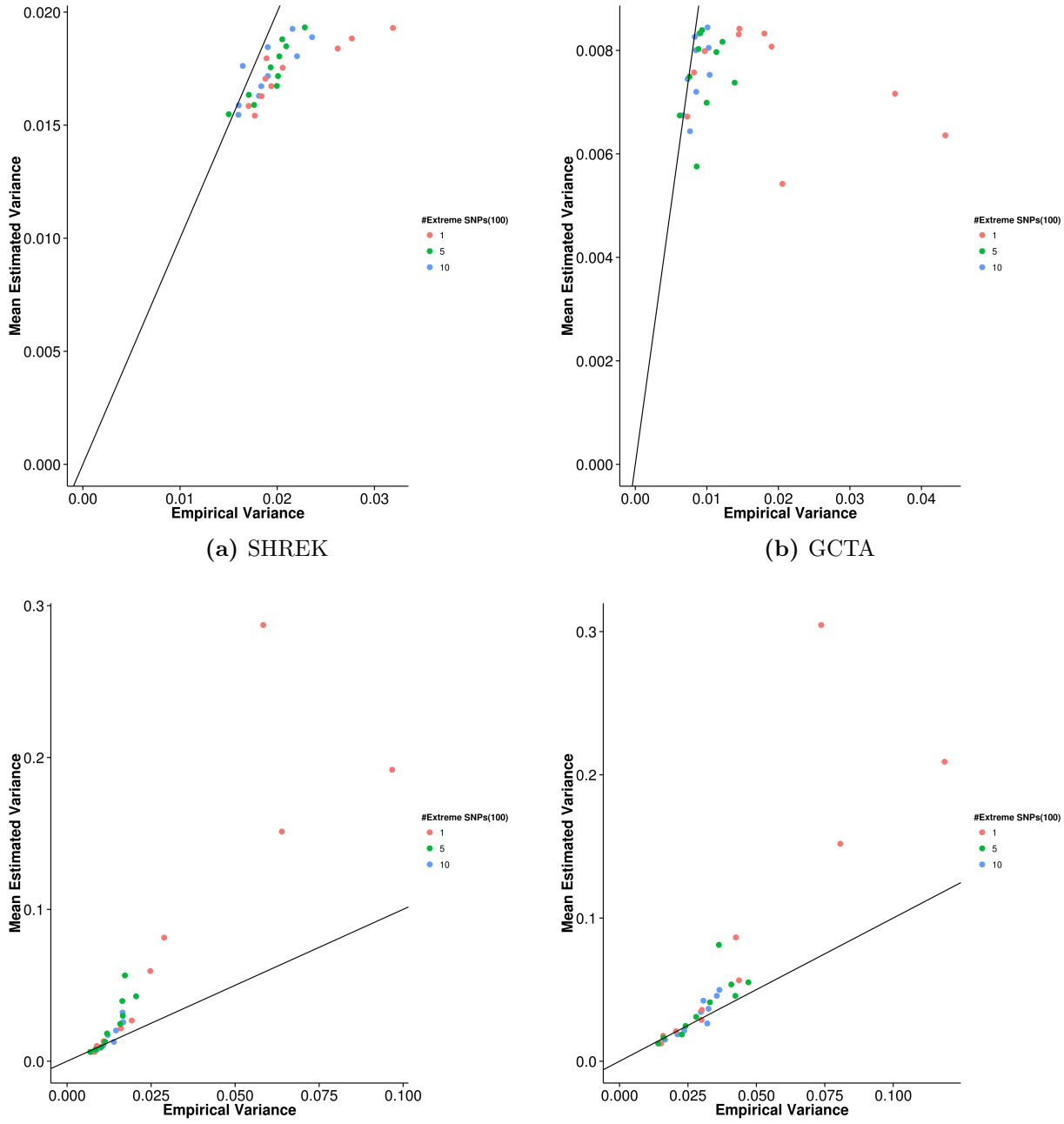
(a) SHREK

(b) GCTA

(c) LDSC with fix intercept

(d) LDSC with intercept estimation

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



(a) SHREK

(b) GCTA

(c) LDSC with fix intercept

(d) LDSC with intercept estimation

Figure 2.8: Estimated variance of results from quantitative trait simulation with extreme effect size simulation when compared to the empirical variance. 100 causal SNPs were simulated. SHREK and GCTA generally under-estimate the variance with the magnitude of bias being the highest when there is only 1 SNP with extreme effect. On the other hand, LDSC tends to over-estimate the variance and it can overestimate the variance by more than 3 folds when there is only 1 SNP with extreme effect.

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

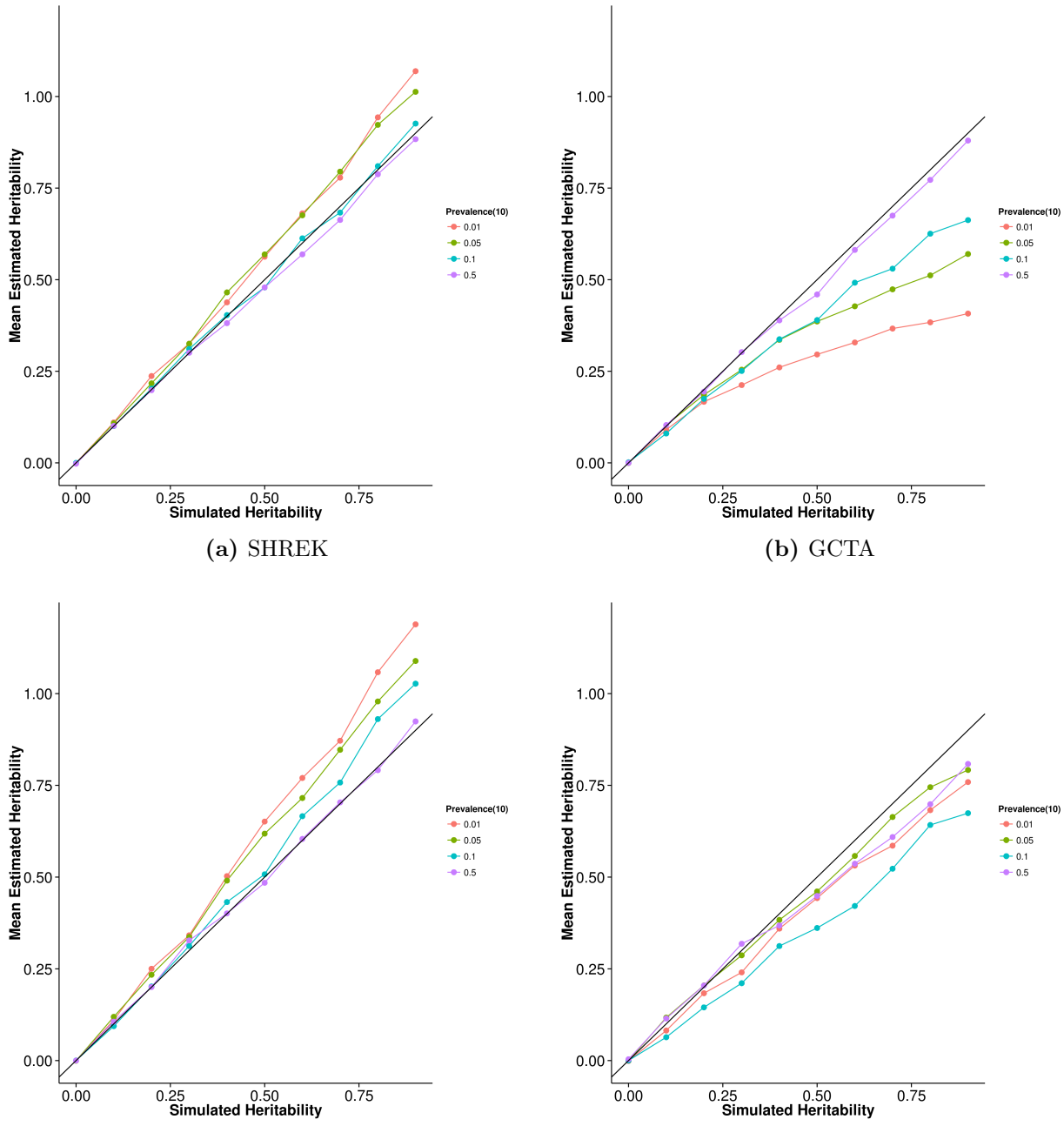
Table 2.2: MSE of quantitative trait simulation with extreme effect size. Of all the algorithms, GCTA has the lowest MSE. When comparing the performance of SHREK and LDSC, it is observed that LDSC performs better unless only 1 of the causal SNP has a large effect size. However, it is also observed that the performance of SHREK is robust to the change in number of SNPs with extreme effect size.

To conclude, GCTA has the best performance among the algorithms tested (table 2.2). However, for studies which only summary statistics are available, GCTA analysis cannot be performed. In such scenario, 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 Case Control Simulation

In order to estimate the heritability from case control studies, it is important to model the disease status under a liability threshold model, which require knowledge of the population prevalence of the disease. Given that both the population prevalence and trait heritability can have a large influence to the performance of SNP heritability estimation, simulations were performed where the population prevalence, the trait heritability and the number of causal SNPs were varied.

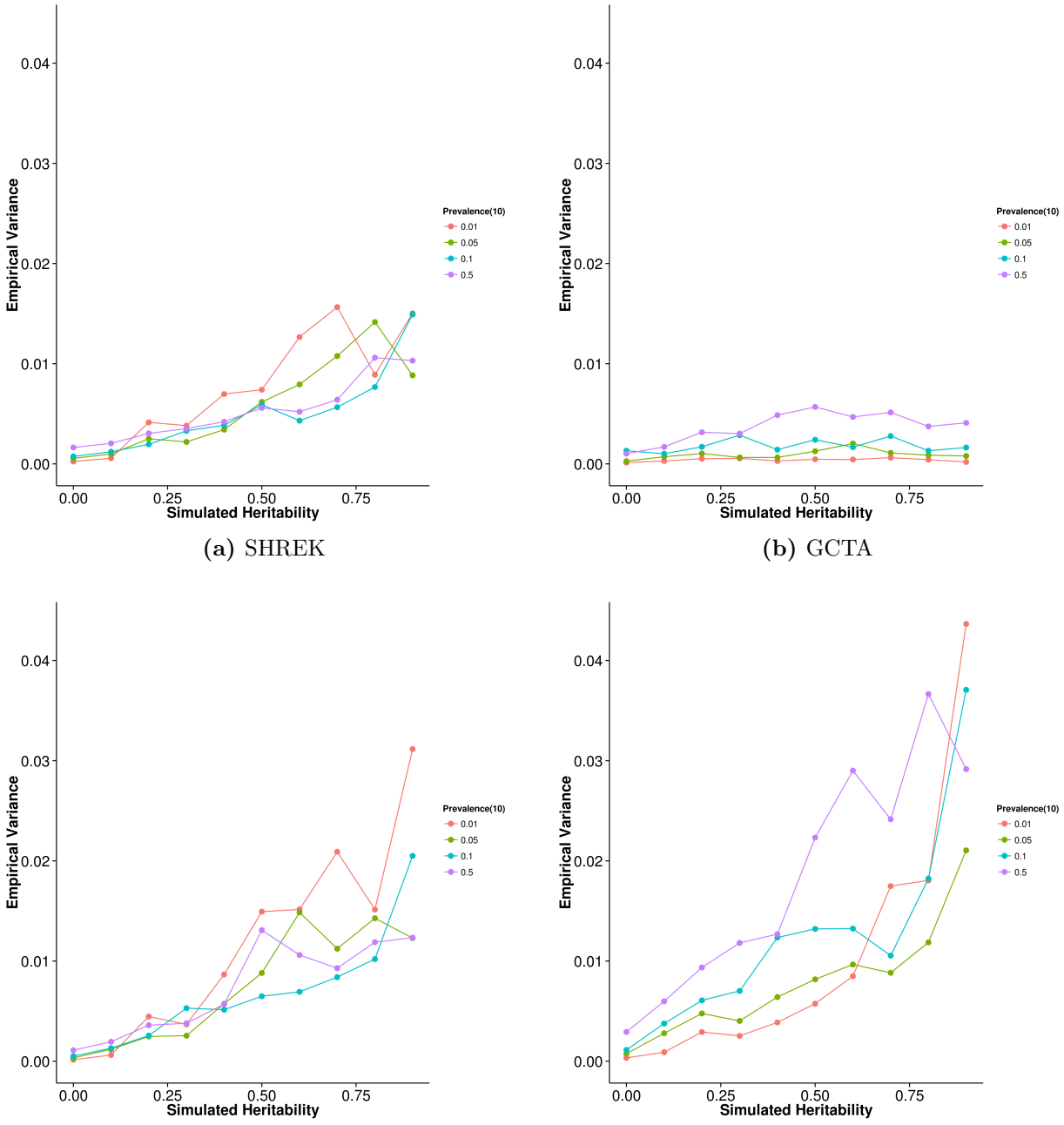
When only 10 causal SNPs were simulated, it is clear that the population prevalence has a significant impact to the performance of the algorithms (fig. 2.9). Of all the algorithms tested, GCTA is the most affected by the population prevalence (fig. 2.9b), where the estimates from GCTA are generally underestimated. As the population prevalence decreases, the magnitude of bias increases, as reported



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

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

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



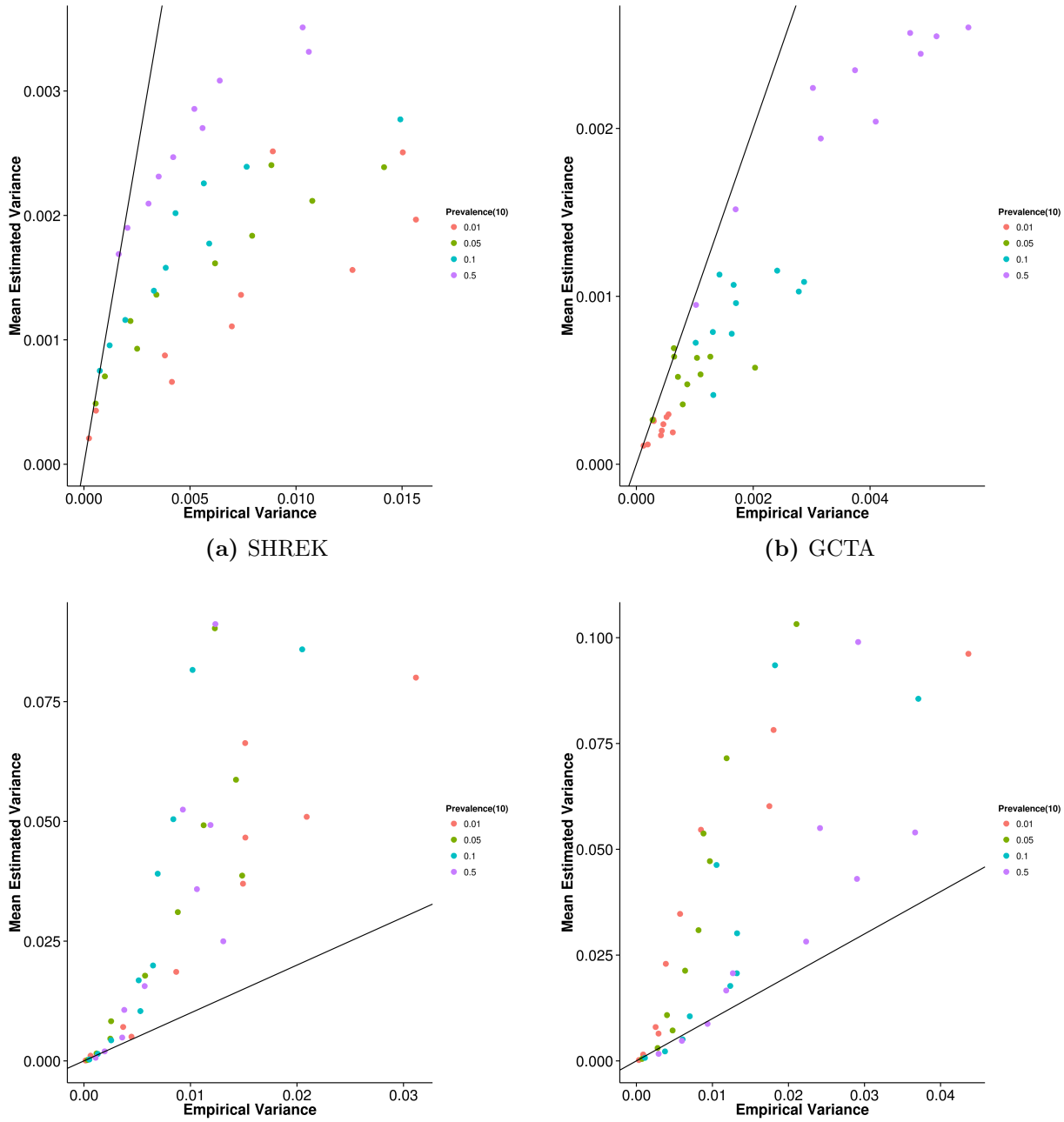
(a) SHREK

(b) GCTA

(c) LDSC with fix intercept

(d) LDSC with intercept estimation

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



(a) SHREK

(b) GCTA

(c) LDSC with fix intercept

(d) LDSC with intercept estimation

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

by Golan, Eric S Lander, and Rosset (2014). On the other hand, the estimates generated by LDSC with fixed intercept and SHREK are upwardly biased. The magnitude of bias also 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. 2.17, 2.20 and 2.23).

Of all the algorithms, GCTA has the smallest average empirical variance (fig. 2.10b) where LDSC with intercept estimation has the largest empirical variance. On the other hand, it is observed that the estimates from SHREK (fig. 2.10a) and LDSC (fig. 2.10c) with fixed intercept have similar empirical variance. As the number of causal SNPs increases, the empirical variance of all algorithms decreases (figs. 2.18, 2.21 and 2.24) 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. 2.11a). On the other hand, GCTA can provide a more accurate estimation for its empirical variance, only moderately underestimated the variance (fig. 2.11b). Again, it is observed that LDSC consistently overestimate its empirical variance (fig. 2.11). However, as the number of causal SNPs increases, the magnitude of bias observed in the estimation of variance decreases for LDSC (figs. 2.19, 2.22 and 2.25). When 500 causal SNPs were simulated, LDSC can provide a relatively accurate estimates of its empirical variance (fig. 2.25c).

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

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

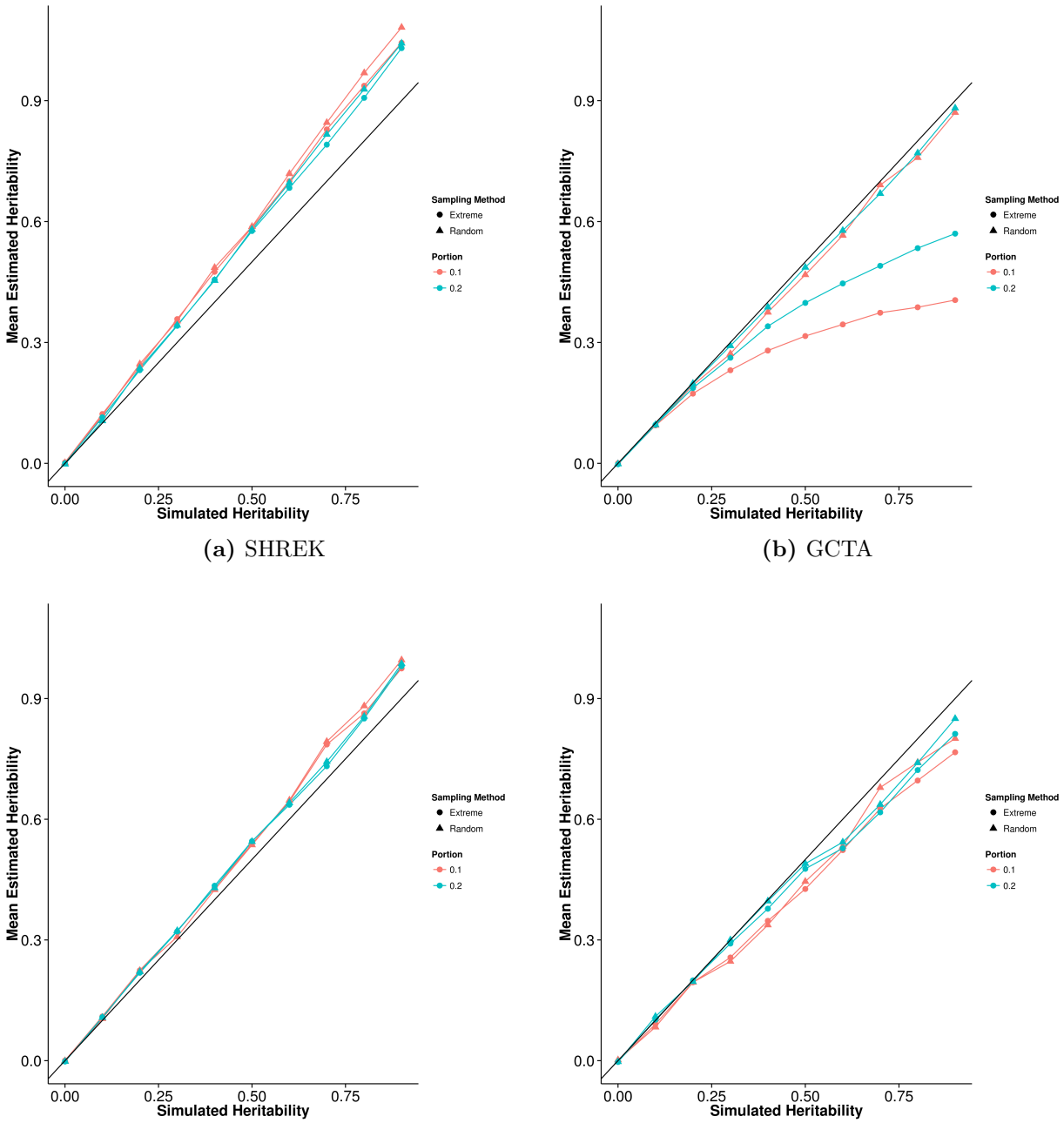
Table 2.3: MSE of Case Control simulation. Algorithm with the best performance under each condition were **bold-ed**. Of all the algorithms, SHREK has the best average performance. It is observed that as the number of causal SNPs increases, the MSE tends to decrease for all algorithms, similar to the results from quantitative trait simulation.

control sampling.

Finally, 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 case control 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 that of extreme phenotype sampling (Peloso et al., 2015), 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



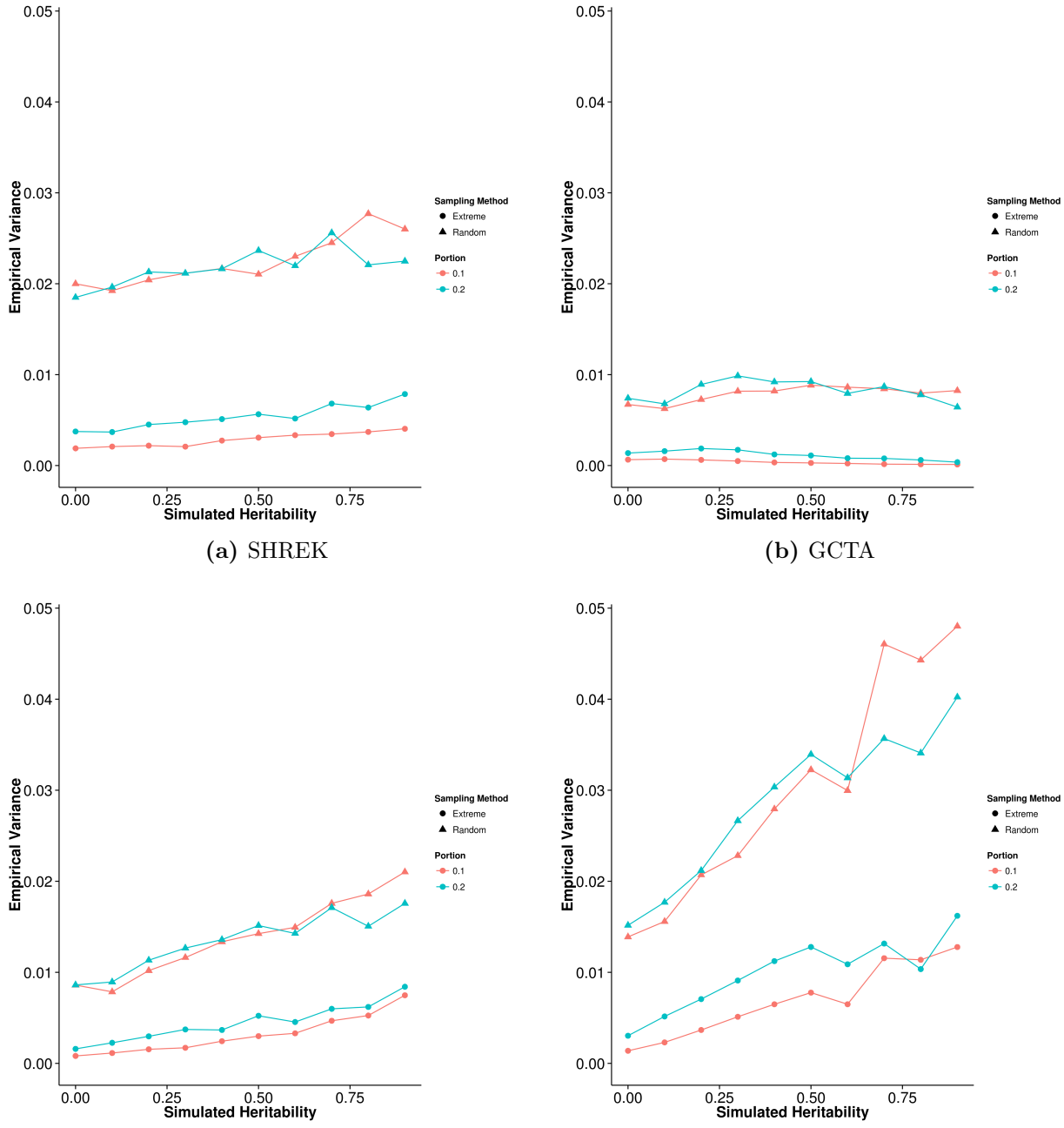
(a) SHREK

(b) GCTA

(c) LDSC with fix intercept

(d) LDSC with intercept estimation

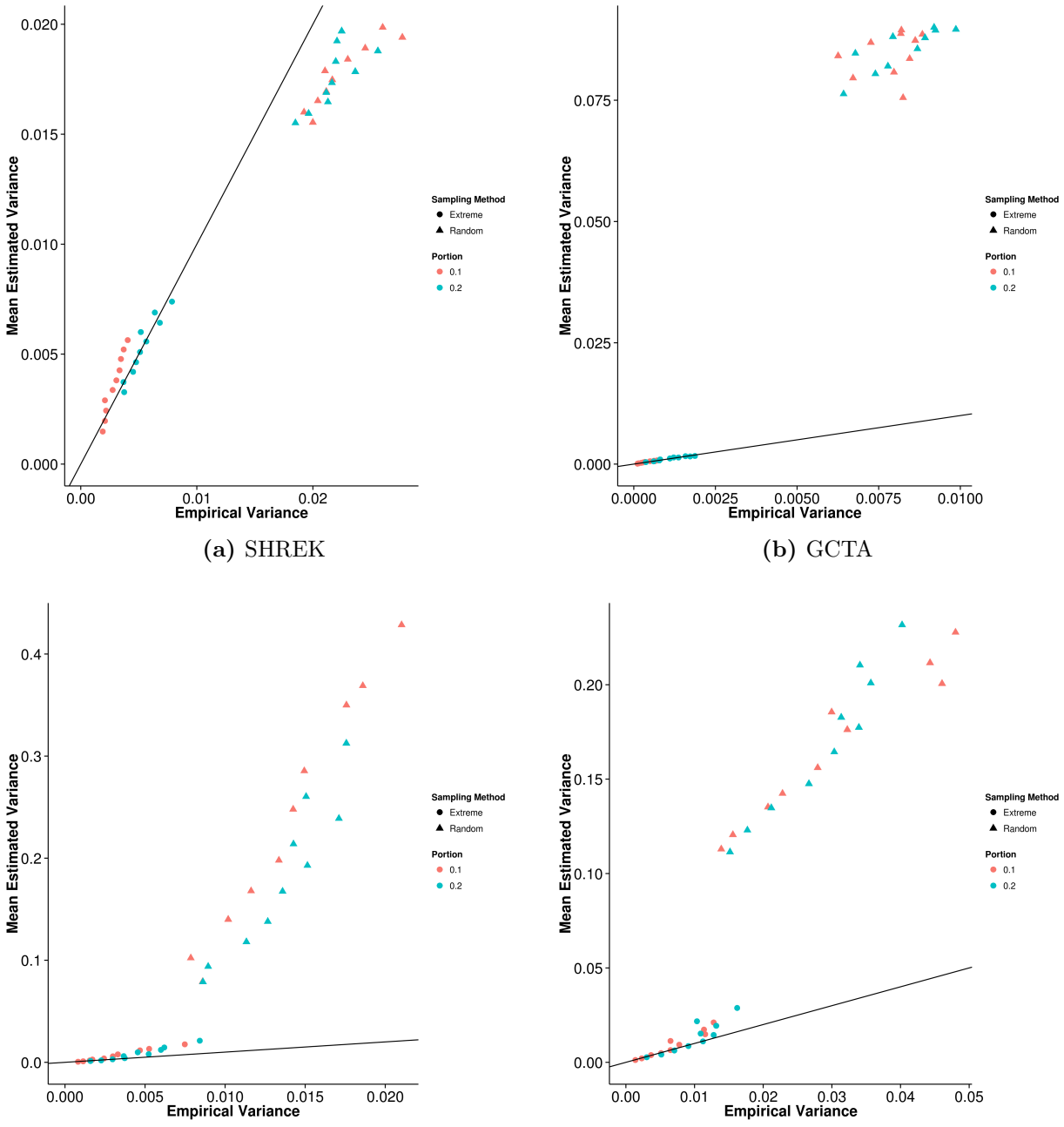
Figure 2.12: Mean of results from extreme phenotype simulation. The performance of the algorithms when random sampling was performed were similar to what was observed in the quantitative trait simulation. However, when extreme phenotype was performed, a larger under estimation was observed for GCTA and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



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

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

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



(a) SHREK

(b) GCTA

(c) LDSC with fix intercept

(d) LDSC with intercept estimation

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

the effect of 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 has been observed in the case control simulation (fig. 2.12b). 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. 2.12).

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. 2.14). On the other hand, the estimated variance of SHREK is unaffected. It is unclear what induces the inflation and further investigations are therefore required.

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

It is estimated that the heritability for major depression disorder is around 0.252 by SHREK and 0.154 by LDSC with intercept estimation (LDSC-In) whereas the heritability of bipolar is estimated to be around 0.308 by SHREK and 0.181 by

LDSC-In (table 2.5). For schizophrenia, the heritability is estimated to be around 0.185 by SHREK and 0.135 by LDSC-In. It is observed that the estimates from LDSC with fixed intercept and SHREK are generally larger than the estimates from LDSC with intercept estimation (LDSC-In) (table 2.5). Moreover, the SNP heritability estimated for schizophrenia is much smaller than previously reported by B. K. Bulik-Sullivan et al. (2015). Our results indicated that the common SNPs have relatively less contribution to the genetic predisposition of individuals to schizophrenia as measured by the heritability estimated. It is possible for genetic variations such as epigenetics and rare variants to account for the remaining 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)

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

The development of GWAS allow the systematic association of common genetic variants, such as SNPs, to traits on a genome-wide scale. In a recent GWAS conducted by the PGC, 108 genetic loci were found to be associated with schizophrenia (Stephan Ripke, B. M. Neale, et al., 2014). Despite the success of the PGC schizophrenia GWAS, it is uncertain whether if common genetic variants such as SNPs are the main contribution to individual differences in the liability to schizophre-

nia. Therefore, estimating the contribution of the common SNPs included in the GWAS, to schizophrenia has important implications for future research strategy.

The contribution of common SNPs to the heritability of a disease (SNP heritability) were usually performed using GCTA, which relies on the genetic relationship matrix. However, the calculation of genetic relationship matrix requires the sample genotypes. The sample genotypes are usually unavailable due to privacy concerns. Instead, summary statistics from the association analysis are usually available.

To our knowledge, LDSC and SHREK are the first algorithms that allows the estimation of SNP heritability using only the summary statistics from a GWAS. With LDSC and SHREK, it is now possible to estimates the SNP heritability from the PGC schizophrenia GWAS, which only the summary statistics of the association is available. The SNP heritability of schizophrenia should provide vital information to the future research direction of schizophrenia.

In this chapter, a series of extensive simulation analyses were performed to examine the effect of different sampling strategies and genetic architectures on the performance of LDSC and SHREK. The SNP heritability of schizophrenia were also estimated using the PGC schizophrenia GWAS summary statistics.

2.4.1 LD Correction

First, because SHREK heavily relies on the LD information for the estimation of SNP heritability, it is important to obtain an accurate estimation of LD. However, because the reference panels are a subsets of the population, sampling error exists in the estimated LD. When taking the square of LD, a positive bias is observed, which can affect the estimation of SNP heritability.

It is observed that of all the bias correction algorithms from section 2.2.6,

the equation from Weir and W G Hill (1980) eq. (2.40) has the best performance. Therefore, it was selected as the default bias correction algorithm.

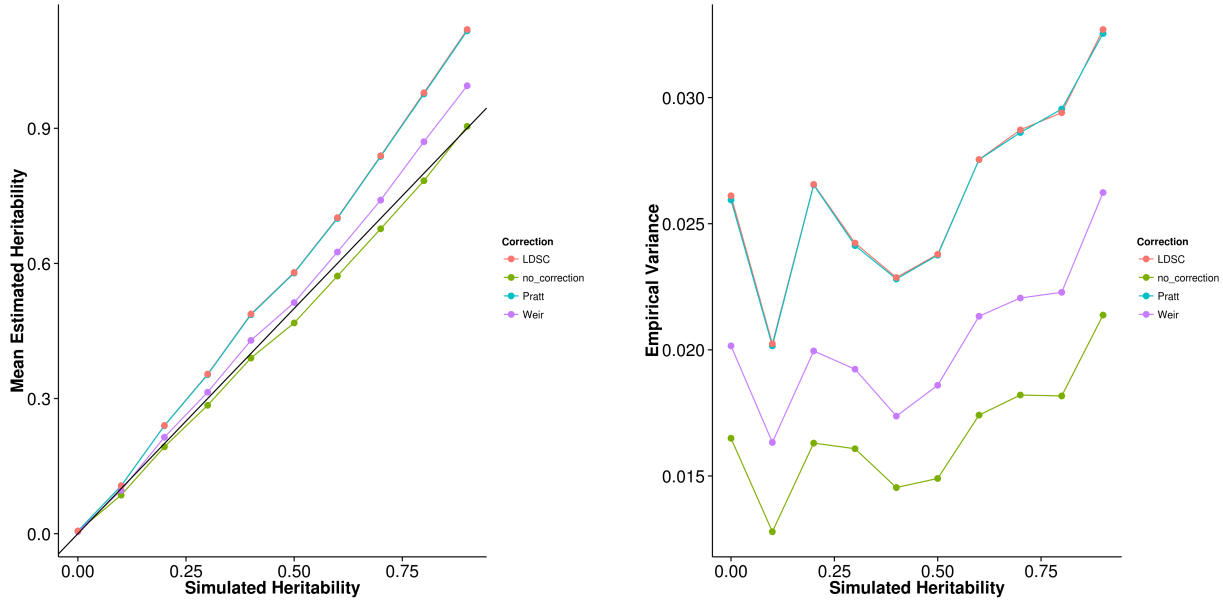
By applying the bias correction algorithm, a more accurate estimates were expected to be obtained. However, in the quantitative trait simulation, an upward bias is consistently observed in the estimates from SHREK. From the LD correction simulation, it is observed that the LD correction algorithms may cause an inflation in the estimates. As the number of simulated SNPs is much smaller in the LD correction simulation when compared to the quantitative trait simulation, it is possible for the increased number of SNPs to increase the magnitude of bias.

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

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

Given these information, it is important to investigate the performance of the LD correction algorithms when a larger number of SNPs are simulated (e.g. 50,000 SNPs on chromosome 1). The LD correction simulation was therefore repeated. Instead of simulating 5,000 SNPs from chromosome 22, 50,000 SNPs were simulated from chromosome 1. As most LD correction from section 2.2.6 produced similar results in previous simulation, only eq. (2.44), eq. (2.40) and eq. (2.38) were included in this scaled up simulation.

It is clear from fig. 2.15 that all LD correction algorithms inflate the estimates from SHREK, whereas a small downward bias is observed in the estimates when no LD correction was performed. Interestingly, when inspecting the MSE of the estimates, SHREK produced the best estimates when no LD correction was performed. This results suggest that as the number of SNPs increases, the LD correction algorithms might have a negative impact to the performance of SHREK.



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

Figure 2.15: 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.

One possible explanation is that small imprecisions might be introduced to the LD when the sampling error is being adjusted. As the number of SNPs increases, the imprecisions accumulates, therefore leading to bias in the estimates of SHREK.

Another interesting observation is that when the LD correction algorithm from LDSC was applied to SHREK, the estimates were upwardly biased. As the same inflation was not observed in the estimates from LDSC, this suggest that SHREK maybe more sensitive to errors within the LD matrix when compared to LDSC. It is noted that the algorithm of SHREK requires the inverse of the LD matrix. Because the LD matrix is ill-conditioned, SHREK is prone to large numerical errors. On the other hand, LDSC does not require to invert the LD matrix, therefore it is more numerically stable. Therefore, because of the fundamental difference in the two numerical methods, SHREK is more sensitive to errors when compared to LDSC.

2.4.2 Simulation Results

The main goal of the current study is to understand the impact of different sampling strategies and different genetic architectures on the performance of the SNP heritability estimation algorithms. A series of extensive simulations were therefore performed.

2.4.2.1 Quantitative Trait Simulation

From the quantitative trait simulation, it is observed that GCTA has the best overall performance. However, the main problem of GCTA is that it requires the sample genotypes for the calculation of the genetic relationship matrix. When the sample genotypes are unavailable, GCTA cannot be performed. On the other hand, LDSC and SHREK can still estimate the SNP heritability given only the summary statistics from a GWAS and a reference panel for the estimation of LD.

Similar to the results from B. K. Bulik-Sullivan et al. (2015), we observed that the variance of LDSC increases as the number of causal SNPs decrease. It is also observed that when the intercept estimation was performed, the variance of LDSC increases. On the other hand, although the variance of estimates from SHREK also increases as number of causal SNPs decrease, the magnitude of increase is much smaller, suggesting that SHREK is robust against change in number of causal SNPs. As a result, when the number of causal SNPs are small, SHREK will provide a more stable result.

Sometimes, it is possible for a small number of causal SNP(s) to have a larger effect size when compared to other causal SNPs. It is observed that only in the extreme scenario where 1 of the causal SNP was simulated with a larger effect size will the performance of LDSC be affected. However, upon re-examination of eq. (2.43) which was used for the simulation of SNPs with extreme effect size, it is

noted that an unnecessary upper bound of (h^2) was imposed to the effect sizes. A better alternative maybe to first simulate the effect sizes using eq. (2.41). Then, for the m "extreme" SNP(s), their effect sizes should be multiplied with a large constant (e.g 10). When compared to eq. (2.43), this should ensure the effect size(s) of the "extreme" SNP(s) to be larger than other causal SNPs. As a result, we are currently repeating the simulation from section 2.2.7.4 to investigate the effect of the presence of a small number of causal SNPs with large effect size to the performance of the algorithms.

Although the effect of number of causal SNPs, trait heritability and the impact of having small number of causal SNP(s) with large effect size were considered, the effect of confounding factors on the performance of the algorithms was not examined. Therefore, it is uncertain how the performance of SHREK and LDSC will be affected by the presence of confounding factors. Specifically, the core concept of the intercept estimation in LDSC is to delineate the contribution from confounding factors and common genetic variants. It is therefore expected that when confounding factors are presented, LDSC with the intercept estimation might outperform SHREK and LDSC with fixed intercept.

However, it is difficult to simulate confounding factors such as population stratification and cryptic relatedness. To our knowledge, there are no readily available softwares for the simulation of cryptically related samples. On the other hand, although it is possible to simulate stratified samples, a multitude of factors can confound the simulation with population stratification. The selection of reference panel is by far the most important factor to consider. For example, if half of the samples were simulated based on European samples and the remaining half of the samples were simulated based on African samples, then it is unclear whether if the European samples or the African samples should be used as the reference panel. Moreover, common practice in GWAS analysis is to adjust for the population stratification

either using principle component analysis (PCA) or EIGENSTRAT (Price et al., 2006). With the additional adjustment, the properties of the summary statistics will differ, and might therefore affect the SNP heritability estimation. As a result, to investigate the effect of population stratification, more factors have to be considered, e.g. different reference panel, different proportion of population included, different adjustment methods (e.g. PCA or EIGENSTRAT) e.t.c. Due to the complexity and scope of the problem, it is still a work in progress for us to investigate how the confounding factors affect the performance of LDSC and SHREK.

Overall, our results suggest that SHREK only outperform LDSC in extreme scenarios such as when the trait is oligogenic or when only 1 of the causal SNP has large effect size. In more general scenarios, the estimates of LDSC have a smaller variance and bias when compared to estimates from SHREK. However, it is noted that when no LD correction were performed, the bias observed in SHREK reduced (e.g. from 0.0217 to 0.0166 in the LD correction simulation with 50,000 SNPs simulated). Therefore, when large amount of SNPs were simulated, the performance of SHREK can be improved by not performing the LD correction. Without the LD correction, performance of SHREK may be comparable to LDSC under the polygenic scenario. Nonetheless, the sensitive to errors in the LD matrix remains to be one of the biggest weakness of SHREK.

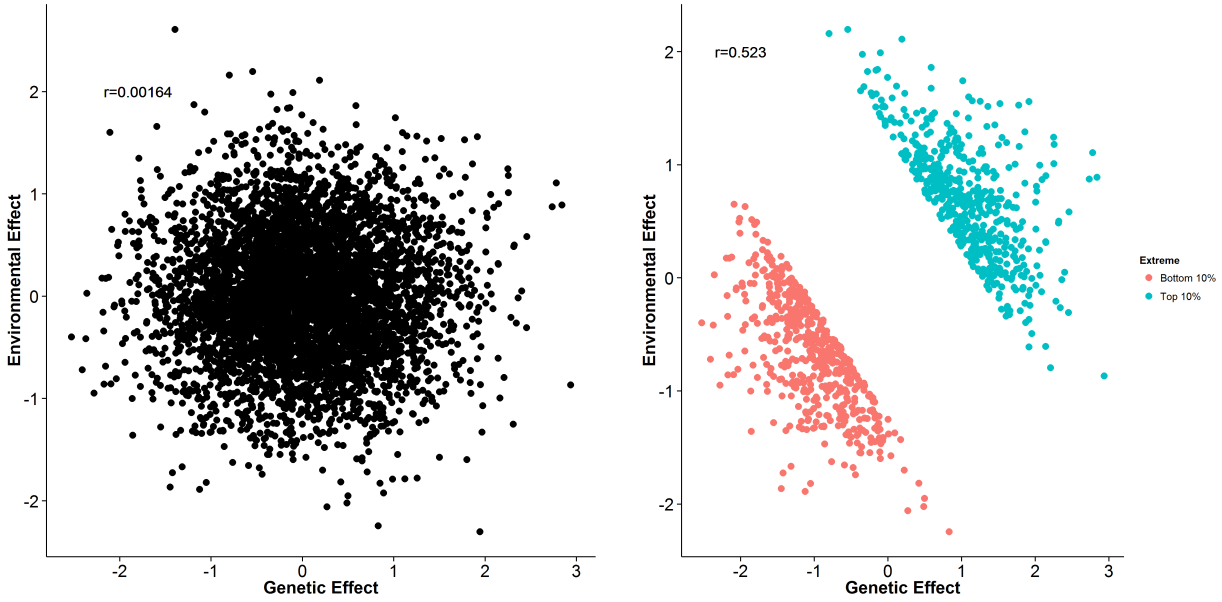
2.4.2.2 Case Control Simulation

In order to estimate heritability from case control samples, 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, Golan, Eric S Lander, and Rosset (2014) observed that GCTA underestimates the heritability explained by common variants for discontinuous traits. The magnitude of this bias is affected by the population prevalence of the trait, the observed prevalence, the true

underlying heritability and the number of genotyped SNPs (Golan, Eric S Lander, and Rosset, 2014). According to Golan, Eric S Lander, and Rosset (2014), there is an oversampling of the cases relative to their prevalence in the population in case control 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 (Golan, Eric S Lander, and Rosset, 2014). This leads to the estimates from GCTA to be strongly downward biased where the magnitude of bias increases as the population prevalence decreases, heritability increases and when the proportion of cases is closer to half.

It is therefore important to investigate whether if SHREK and LDSC are affected in a similar way. In our simulation, the same bias is observed for GCTA, whereas for SHREK and LDSC, an inflation is observed in the estimates when the population prevalence is small. As the population prevalence decreases, the magnitude of bias increases, suggesting that the population prevalence has a major influence to the estimates of LDSC and SHREK.

Surprisingly, by applying the intercept estimation, the estimates of LDSC becomes more robust to the change in population prevalence (figs. 2.9d, 2.17d, 2.20d and 2.23d). It was expected that as no confounding factors were simulated, the intercept estimation function will be redundant. However, results suggest that in the estimation fo SNP heritability from case control samples, the intercept estimation might be beneficial even when no confounding factors were presented. When population prevalence is small (e.g < 0.05), the performance of LDSC is better when the intercept estimation was performed even without the presence of any confounding factors table 2.3. Further investigation are required to understand how the intercept estimation can improve the performance under the case control scenario. This might provide insight for the development of a better algorithm in the estimation of SNP heritability from case control samples for LDSC and SHREK.



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

Figure 2.16: 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 (Golan, Eric S Lander, and Rosset, 2014).

2.4.2.3 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 distribution, one can achieve the same power as a study with random sampling design that has 4 times the sample size (Pak C Sham and Shaun M Purcell, 2014).

Interestingly, it is observed that when extreme phenotype sampling was performed, the estimates from GCTA are biased downward, similar to the bias observed in the case control simulation. The bias observed in GCTA might be a

result of artificial correlation between genetic and environmental effects introduced by extreme phenotype sampling (fig. 2.16), which is similar to the case control scenario. Therefore, as a smaller portion of samples were selected from the extreme ends of a population, the magnitude of bias observed in estimates of GCTA might increase.

On the other hand, an upward bias is observed in the estimates from SHREK and LDSC when extreme phenotype sampling was performed. The bias is slightly higher when a smaller portion of samples were selected from the extreme end. The pattern of the bias observed in the extreme phenotype sampling concurs with the bias observed in the case control scenario. This suggests that the ascertainment bias introduced by non-random sampling might affect the performance of the algorithms. Further investigation are therefore required to identify a better adjustment for the ascertainment bias in order to improve the performance of LDSC and SHREK when extreme phenotype sampling was performed.

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 for 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. However, in the simulation of random sampling, all parameters are the same as those in the quantitative trait simulation. It is therefore unclear why a different estimated variance are obtained. 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.

Finally, our simulation only considered limited dimension of parameters. For example, only traits with 100 causal SNPs were simulated. More simulations should therefore be performed in order to understand the effect of extreme phenotype sampling on the performance of the algorithms.

2.4.3 Application to Real Data

When applying SHREK and LDSC to estimate the SNP heritability in real data, it is observed that the estimates from LDSC is much smaller than the estimates from the supplementary materials of B. K. Bulik-Sullivan et al. (2015) (e.g. for schizophrenia, 0.555 compared to 0.135). After communicated with the corresponding author (B. Bulik-Sullivan, 2015), 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:

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

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

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

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

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

M was originally defined as the total number of SNPs in the reference panel used to estimated 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 which B. K. Bulik-Sullivan et al. (2015) deem 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." (B. K. Bulik-Sullivan et al., 2015). This explained the discrepancy between our estimates and the estimates observed in the supplementary table from B. K. Bulik-Sullivan et al. (2015).

It is observed that the estimates from LDSC with intercept estimation (LDSC-In) is smaller than the estimates from SHREK and LDSC with fixed intercept (table 2.5). From the case control simulation, it is observed that when the population prevalence is less than 0.5, LDSC-In underestimates the SNP heritability, whereas SHREK and LDSC with fixed intercept overestimates the SNP heritability. Therefore it is likely for the estimates from LDSC-In and SHREK to be the lower and upper bound of the true SNP heritability respectively.

However, in our simulation, confounding effects were not simulated, therefore it is unclear how the algorithms will perform in the presence of confounding effects. In the presence of population stratification or cryptic relatedness, spurious associations might be observed, leading to an inflated summary statistics (Zheng, Freidlin, and Gastwirth, 2006). It is therefore likely for SHREK and LDSC with fixed intercept to overestimate the true SNP heritability in the presence of population stratification or cryptic relatedness.

Moreover, from the LD correction simulations, it is noted that estimated from SHREK biased upward when LD correction was performed, especially when large number of SNPs were included in the study. As LD correction was performed when we apply SHREK to real data, it is likely for SHREK to provide an upward biased estimate. Together, it is very likely for SHREK to overestimate the true SNP heritability, therefore, its estimates should be treated as the upper bound of

the true SNP heritability. Further development of SHREK are required to reduce the bias and improve its performance.

Based on our estimation, the PGC schizophrenia GWAS can at most account for $\sim 20\%$ of the heritability of schizophrenia (SCZ). When compared to the heritability estimated from twin studies, there are around $40\% \sim 60\%$ of missing heritability unaccounted for. As the SNP heritability were only estimated based on the autosomal SNPs, it is possible for the sex chromosome to account for some of the "missing heritability". Different methods exists for the association analysis on the X chromosome where a different summary statistics can be obtained (Wong et al., 2014) (Supplementary materials). It is therefore difficult to perform the estimation of SNP heritability on the X chromosome with only the summary statistics. Further investigation are required before LDSC and SHREK can be applied to estimation the contribution of SNPs resides on the sex chromosomes.

On the other hand, as the common SNPs do not account for all the heritability of schizophrenia, other genetic variants such as the rare variants and epigenetic changes might be another possible source of heritability of schizophrenia.

As LD calculated from rare variants usually have a large SE, SHREK and LDSC cannot accurately estimate the contribution of rare variants to the heritability of a disease. B. K. Bulik-Sullivan et al. (2015) 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. Therefore, further developments are required in order to be able to estimate the true contribution of rare variants to the heritability of schizophrenia.

On the other hand, it was observed that individuals born to schizophrenic mother has a higher risk of schizophrenia when compared to individuals born to a schizophrenic father (Riley and Kendler, 2006). It is therefore possible for variations in methylation pattern or genetic mutation in the mitochondria, which are mainly

transmitted from the mother (Sutovsky et al., 1999), to contribute to the heritability of schizophrenia. Therefore epigenetic might also have an important role in the etiology of schizophrenia.

To conclude, the main advantage of SHREK is its robustness against different genetic architectures, however, further optimizations are required and will be done soon. Nonetheless, the development of LDSC and SHREK allows the estimation of SNP heritability using only summary statistics from GWAS. By understanding the relative contribution of common variants such as SNP to the heritability of a disease, a better research strategy can be developed, thus allow for a better use of research resources.

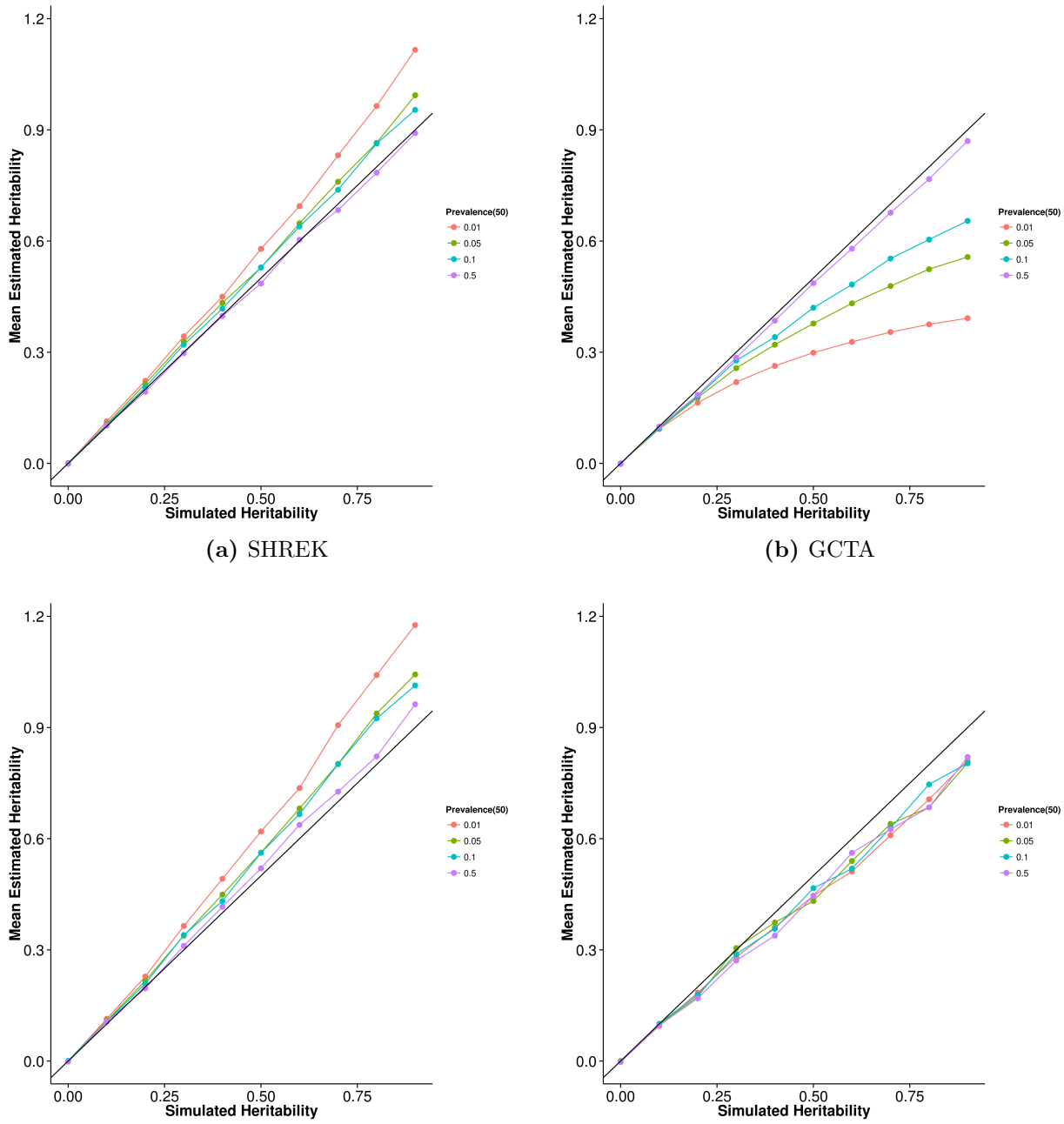
2.4.4 Limitations and Improvements

When compared to LDSC, SHREK is significantly slower because the inverse of the LD matrix are required. When the SNP density is high (e.g $> 2,000$ SNPs in a 1 mb region), SHREK requires a significant amount of time to estimate the SNP heritability and in the case of PGC schizophrenia GWAS dataset, the SNP density was too high, forcing us to use a smaller window size for the estimation. A possible improvement to SHREK will be to use a faster library such as the Armadillo library (Sanderson, 2010). Armadillo library can be 3 times faster than EIGEN C++ library, which is currently used for the implementation of SHREK, by utilizing the multi-threading high performance OpenBLAS library for the computation of SVD. Therefore, by using Armadillo and OpenBLAS in the implementation of SHREK, the speed of the analysis can be increased. However, because of the fundamental nature of the algorithm (requiring the inverse of LD), it is expected that SHREK will be slower in comparison to LDSC.

Finally, we acknowledge that more simulations can be performed. For

example, the effect of the observed prevalence in case control simulations were not investigated. Also, in the extreme phenotype sampling simulation, traits with different number of causal SNPs can also be simulated. However, the current simulations have provided a general concept to the effect of different sampling strategies and genetic architecture on the performance of LDSC and SHREK.

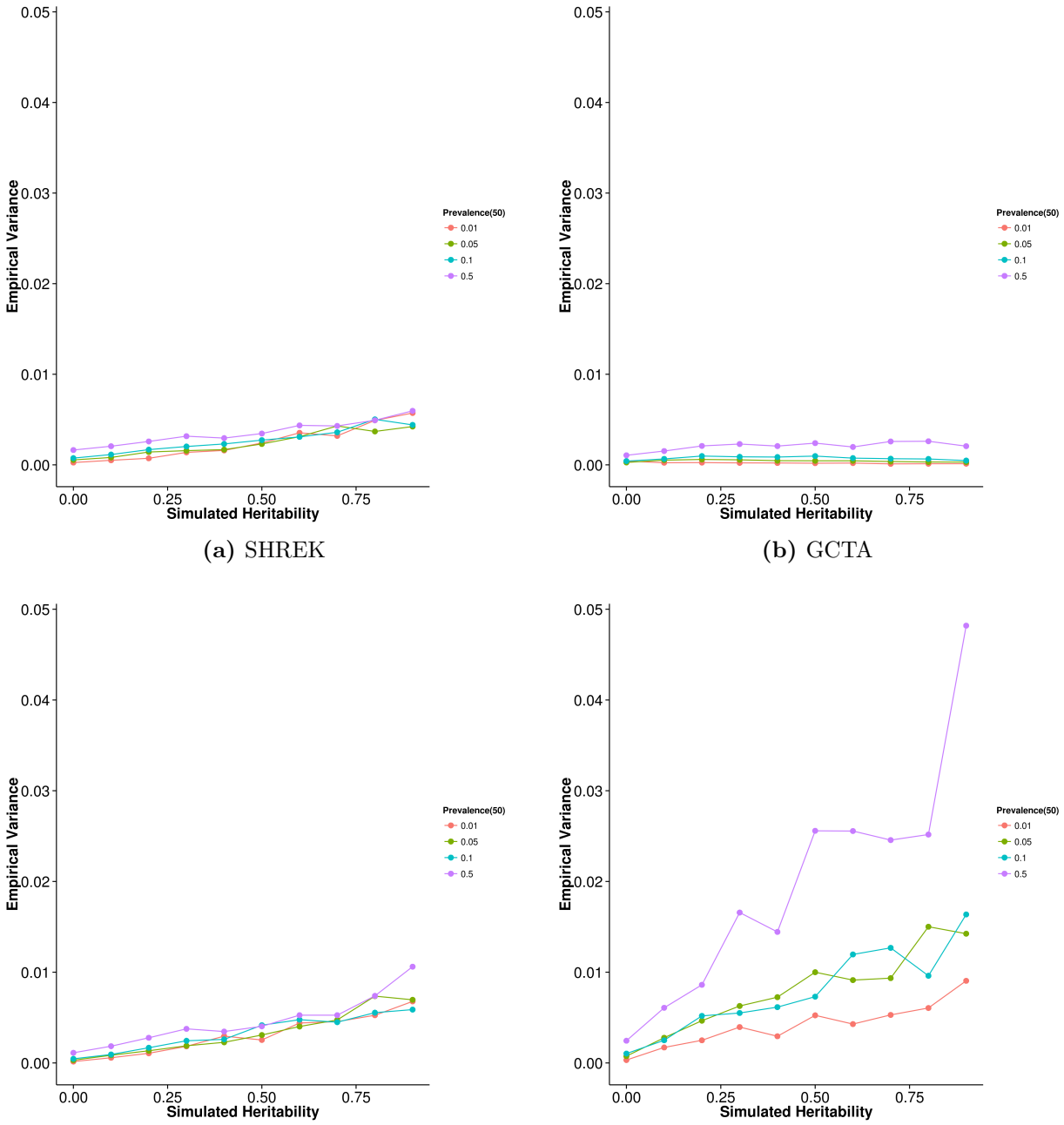
2.5 Supplementary



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

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

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



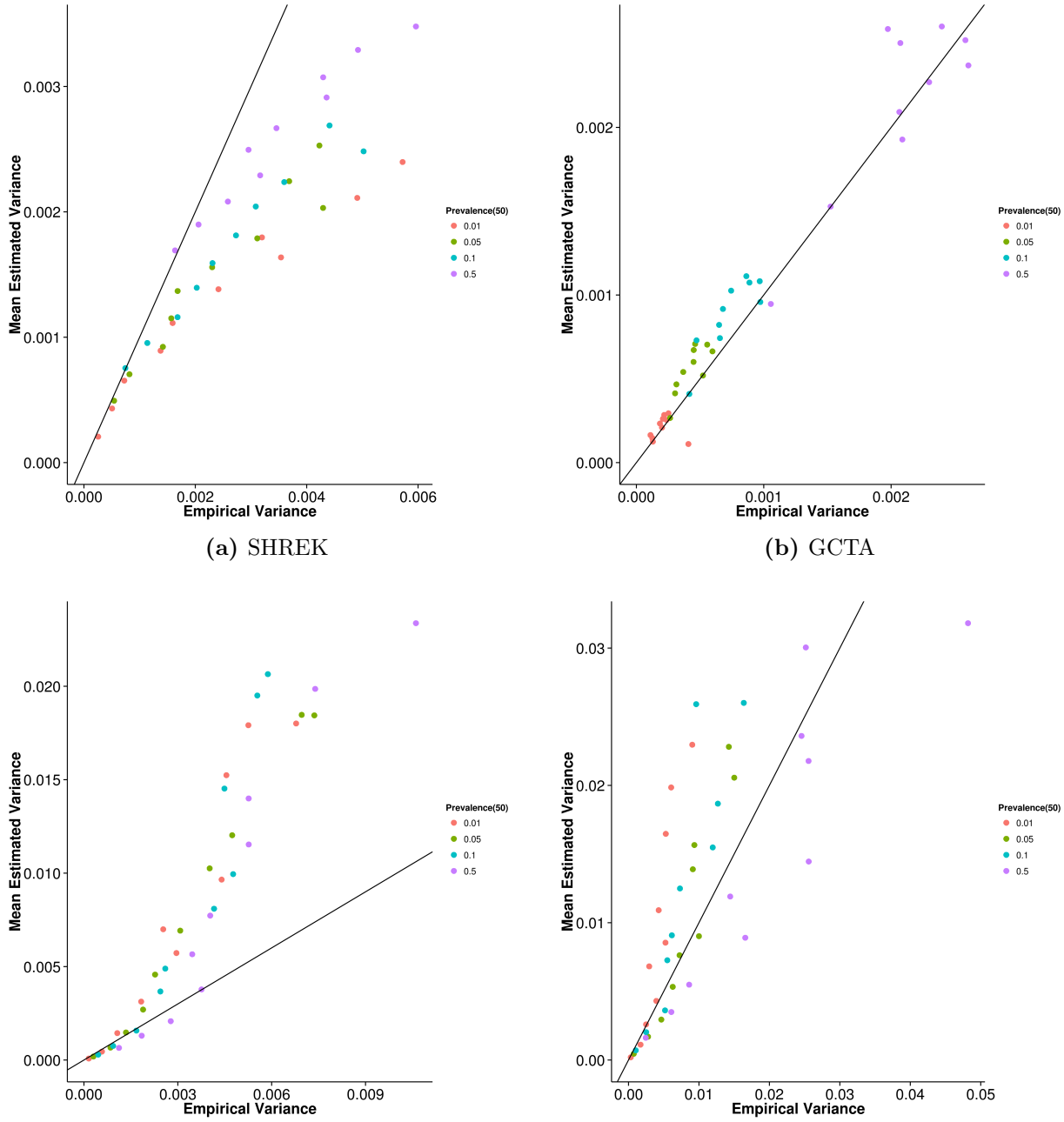
(a) SHREK

(b) GCTA

(c) LDSC with fix intercept

(d) LDSC with intercept estimation

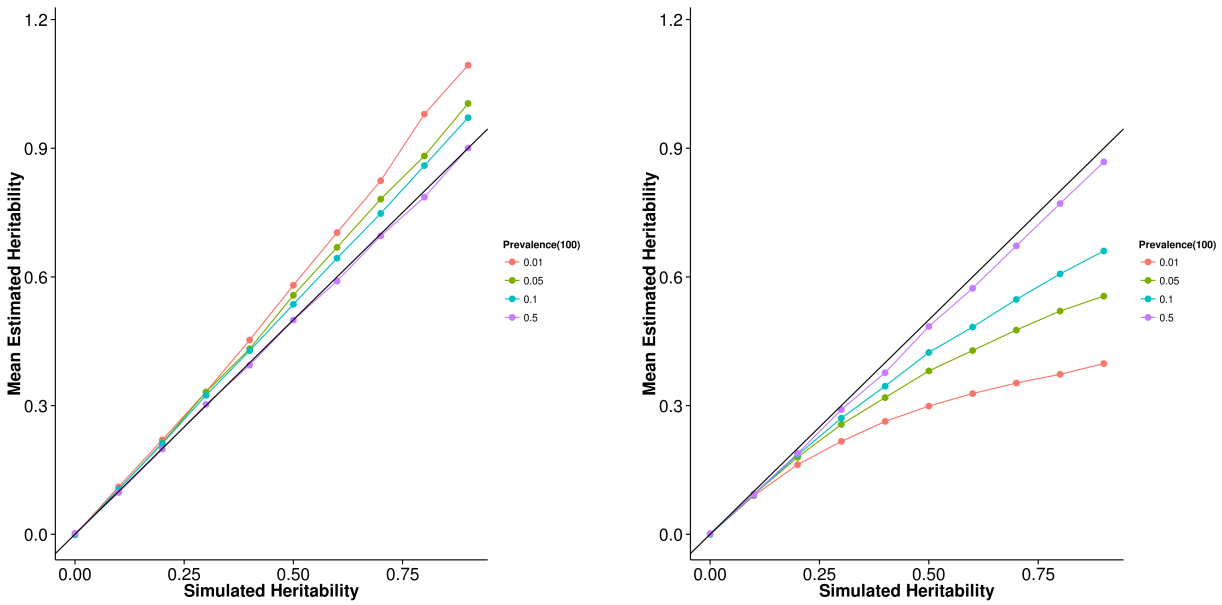
Figure 2.18: Variance of results from case control simulation with random effect size simulation with 50 causal SNPs. For most algorithm except that of LDSC with fixed intercept, the empirical variance of the estimates increases as the population prevalence of the trait increases, with the estimations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



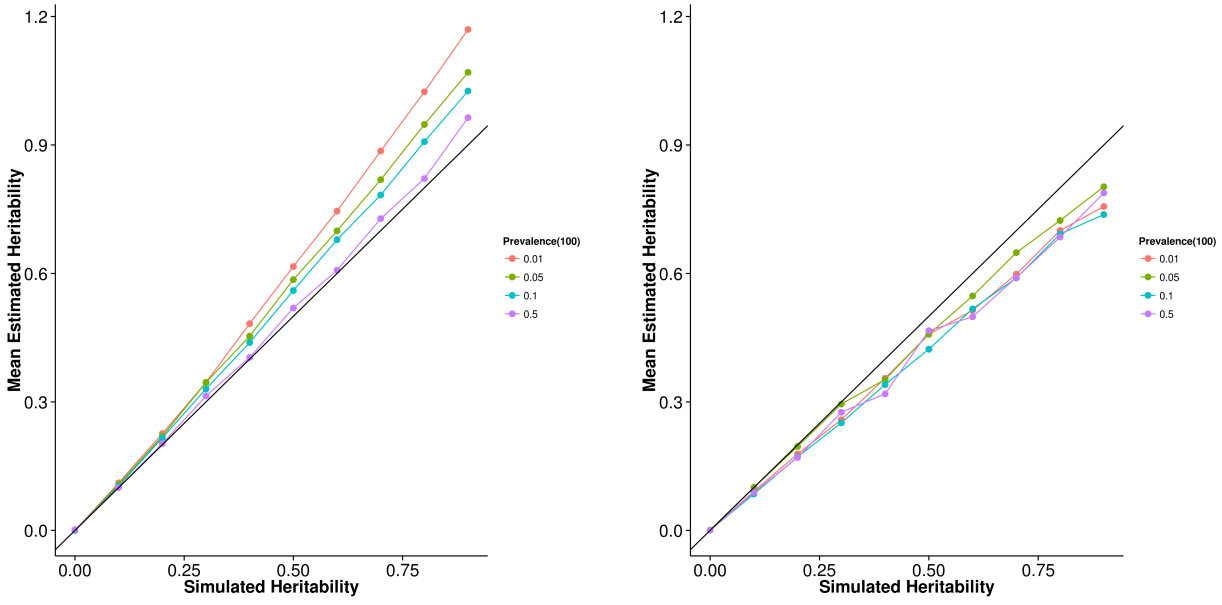
(a) SHREK **(b)** GCTA

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

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

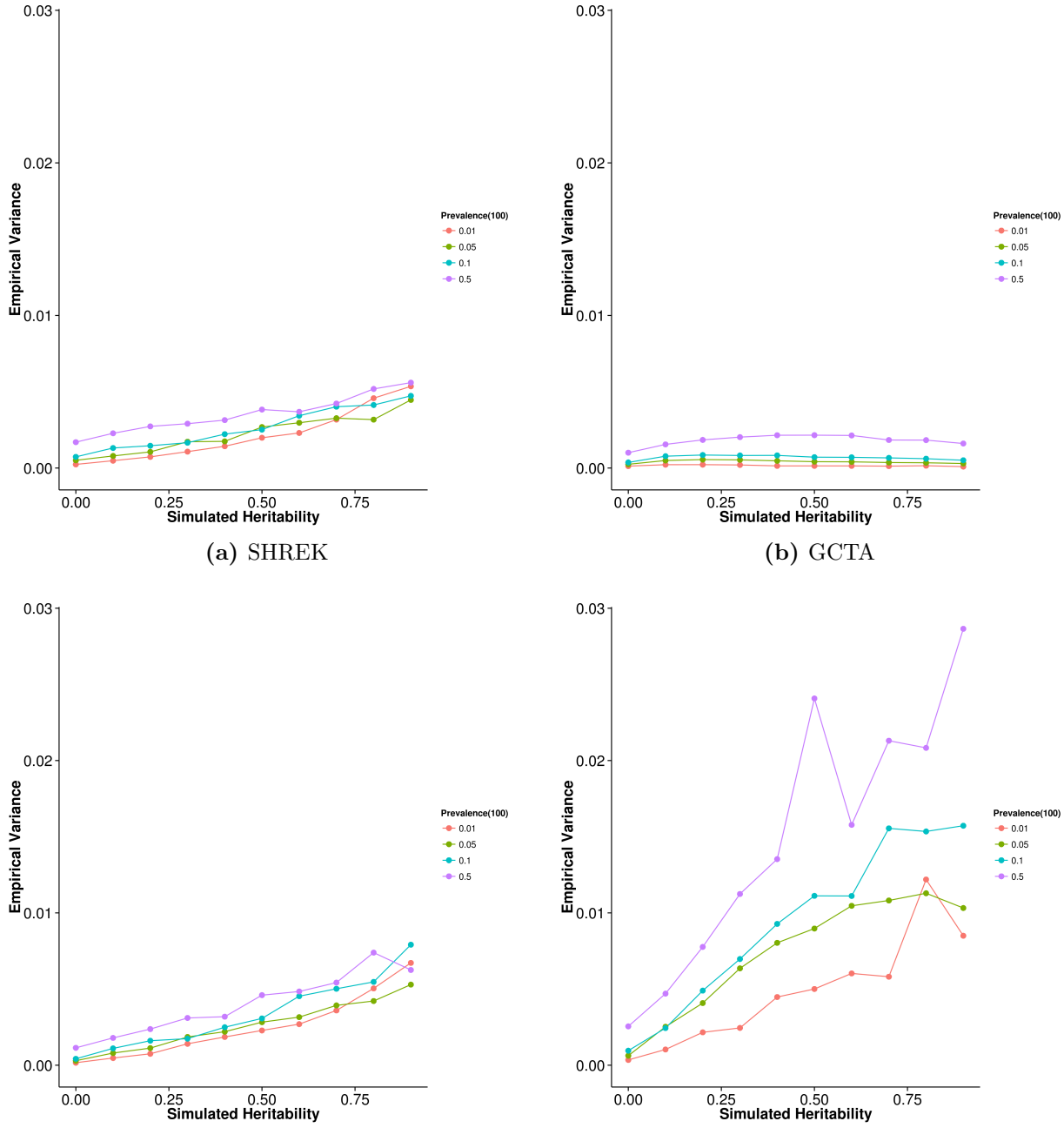


(a) SHREK **(b)** GCTA



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

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



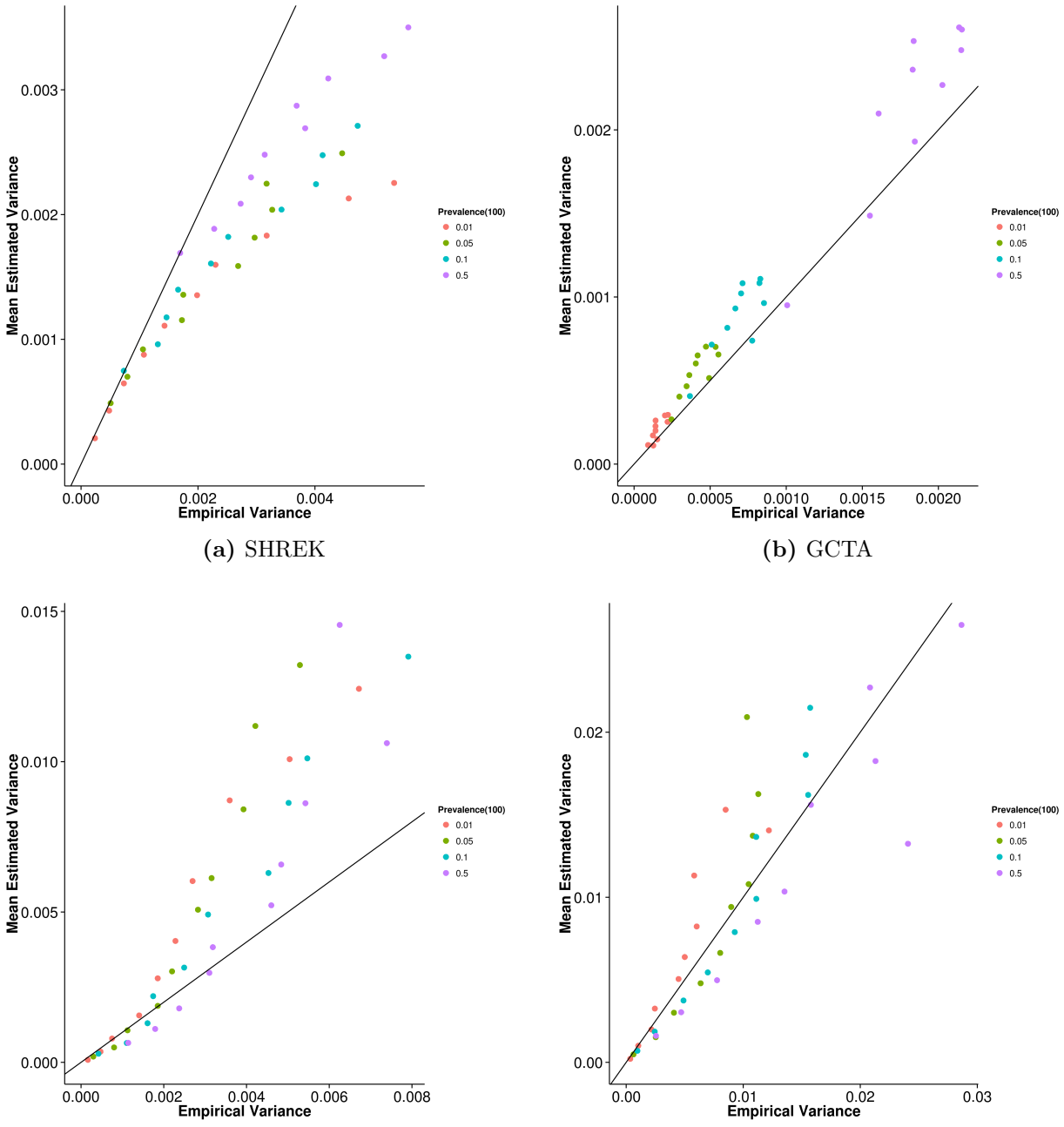
(a) SHREK

(b) GCTA

(c) LDSC with fix intercept

(d) LDSC with intercept estimation

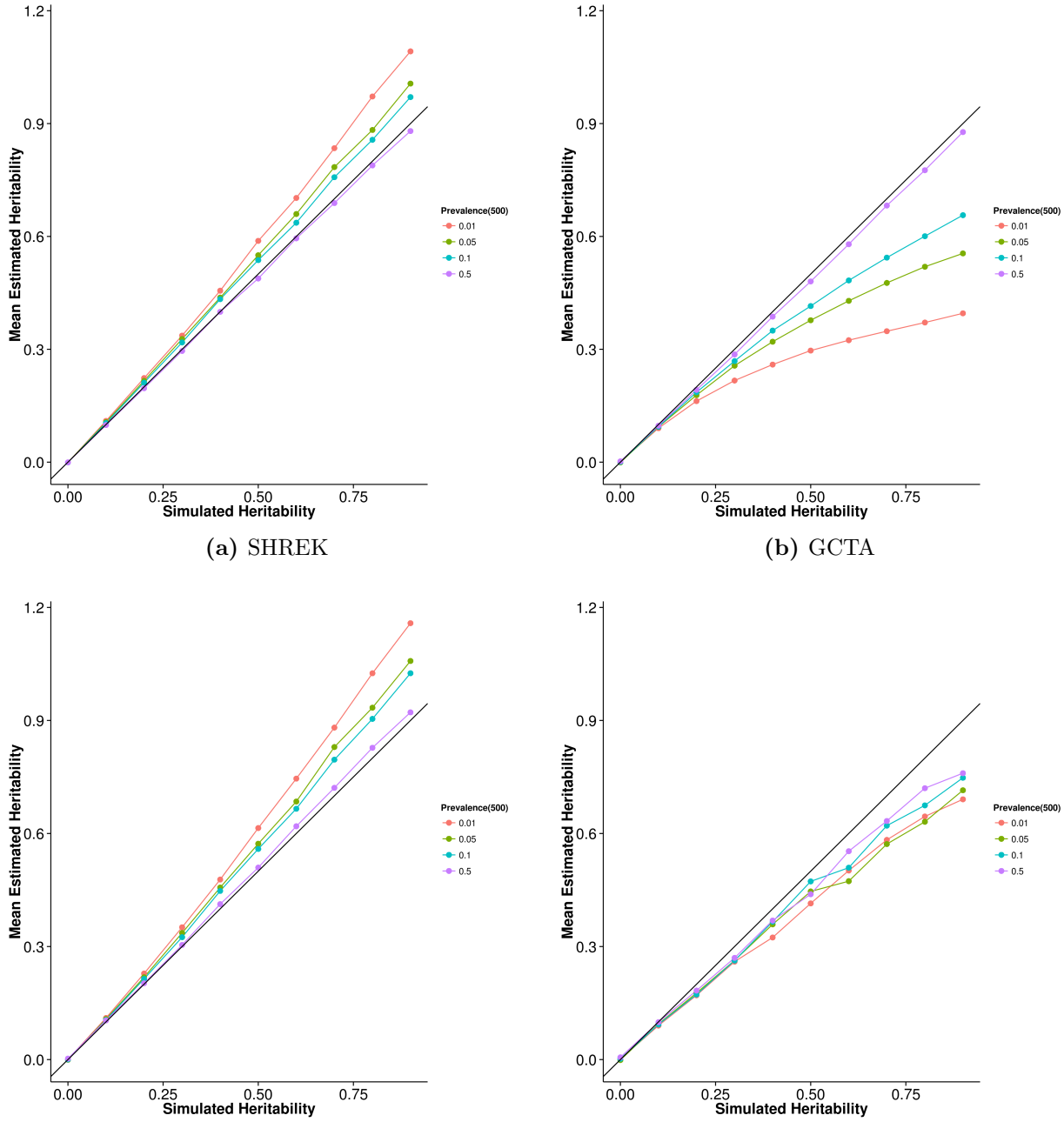
Figure 2.21: Variance of results from case control simulation with random effect size simulation with 100 causal SNPs. As the number of causal SNPs increased to 100, the relationship between the population prevalence and the empirical variance of the algorithms become clear where as the population prevalence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



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

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

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



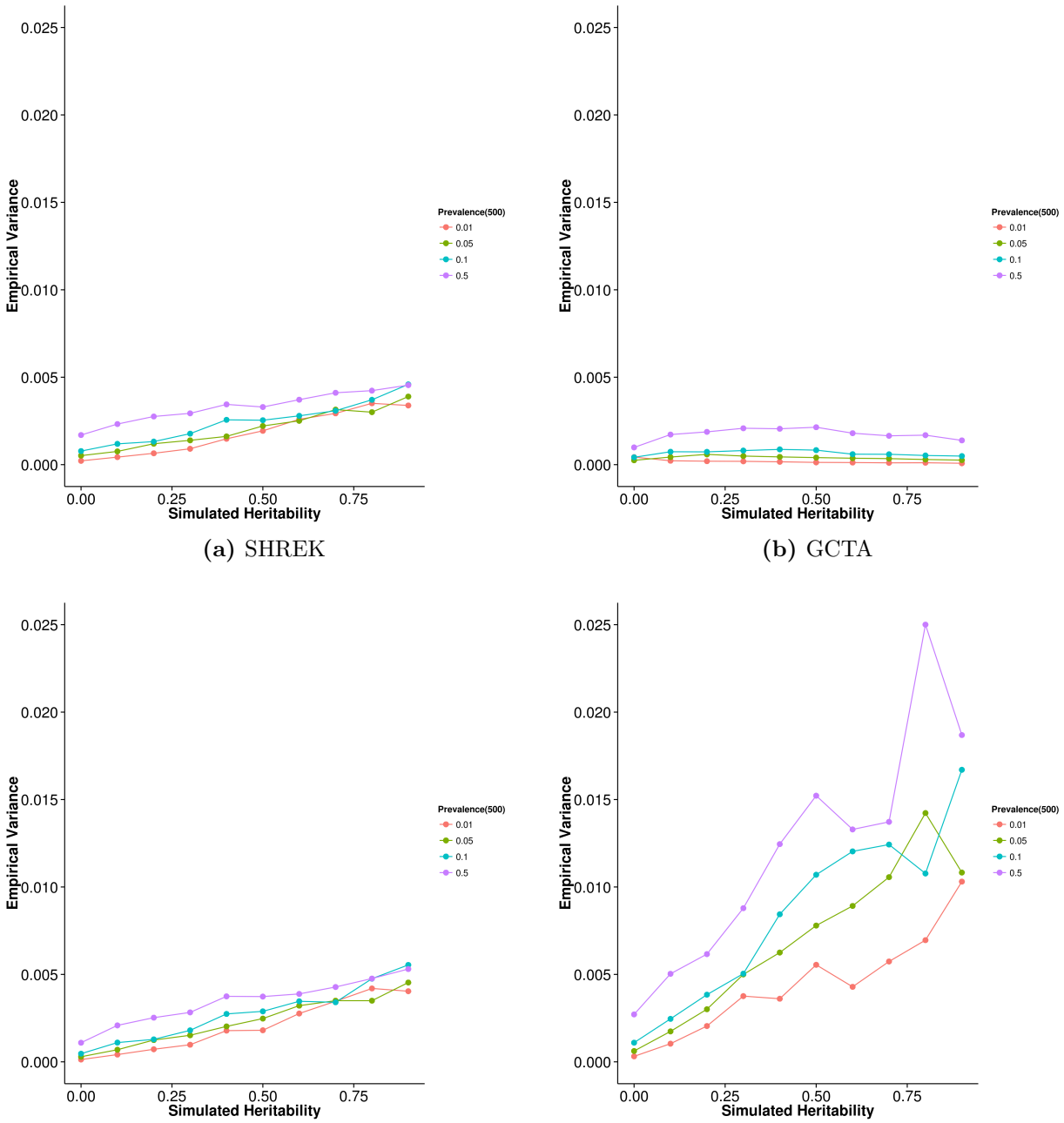
(a) SHREK

(b) GCTA

(c) LDSC with fix intercept

(d) LDSC with intercept estimation

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



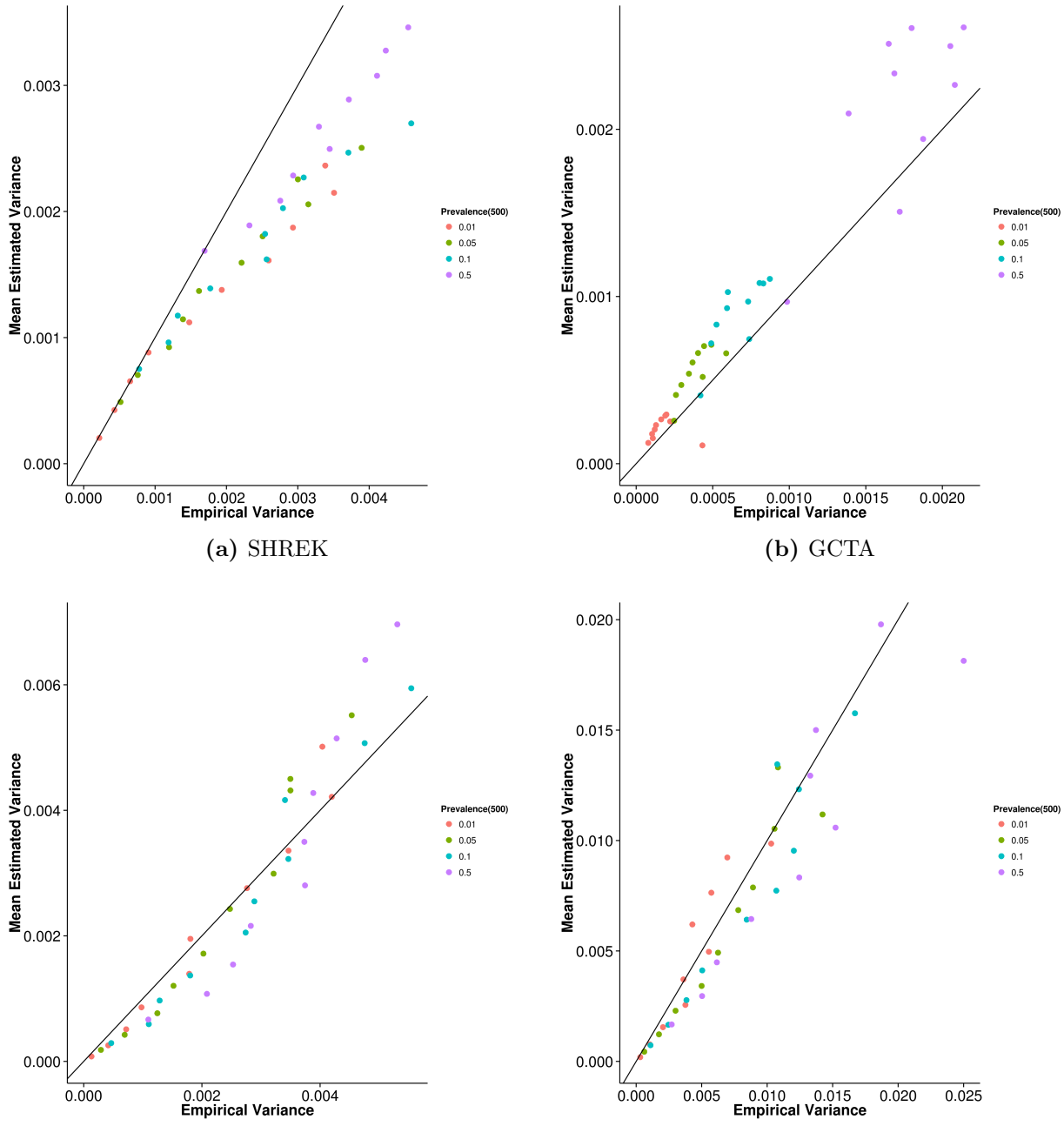
(a) SHREK

(b) GCTA

(c) LDSC with fix intercept

(d) LDSC with intercept estimation

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



(a) SHREK

(b) GCTA

(c) LDSC with fix intercept

(d) LDSC with intercept estimation

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

3 n-3 Polyunsaturated Fatty Acid Rich Diet in Schizophrenia

3.1 Introduction

In the previous chapter, we have found that the SNP heritability of schizophrenia is estimated to be at most 20% which is much lower than expected in schizophrenia. This suggests that other factors such as rare variants such, epigenetic factors and gene-environmental interaction ($G \times E$) might contribute to the "missing" heritability of schizophrenia.

Previous studies have reported the possibility of interaction between prenatal infection and genetic variation in risk of developing schizophrenia (Tienari et al., 2004; Clarke et al., 2009). It has been suggested that the effect of prenatal infection was mainly mediated by maternal immune response, instead of the specific type of infection (A S Brown and Derkits, 2010). Therefore it is likely that the perturbation induced by maternal immune activation (MIA) interacts with genetic variations in the development of schizophrenia.

Slowly but steadily, progress has been made in the research of schizophrenia. Converging evidence from GWAS, CNV and sequencing studies suggest that rare and common variants in genes related to postsynaptic density (PSD) (S M Pur-

cell et al., 2014; T. N. Consortium and Psychiatric Genomics, 2015) and calcium ion channels (S M Purcell et al., 2014; Stephan Ripke, B. M. Neale, et al., 2014; Szatkiewicz et al., 2014) contribute to the etiology of schizophrenia. Given these, it is possible for the effect of prenatal infection to also act upon the same functional gene sets during the development of schizophrenia.

Additionally, with the development of LDSC, partitioning of SNP heritability can now be performed using summary statistics from GWAS. It is therefore possible to not only identify gene sets or pathways that are associated with schizophrenia, but also estimate their relative contribution to the heritability of schizophrenia.

Furthermore, a number of studies have reported the potential of n-3 polyunsaturated fatty acid (PUFA) in the treatment of schizophrenia (Q. Li, Leung, et al., 2015; Trebble et al., 2003). In mouse, it was found that n-3 PUFA can inhibits the production of IL-6 (Trebble et al., 2003) - a major mediator in MIA model (Smith et al., 2007). 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 (Clandinin, 1999; Kitajka et al., 2002). 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 (Q. Li, Leung, et al., 2015).

Herein, we conduct a hypothesis-driven 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 RNA Sequencing. Another goal of the current study is to investigate whether the effects of MIA or diet act upon the same functional gene sets as the genetic variants associated with schizophrenia in the development of the disease. Finally, the relative contribution of the candidate gene sets to the heritability of schizophrenia was also estimated using LDSC.

Although hippocampus (Velakoulis et al., 2006; Nugent et al., 2007) and prefrontal cortex (Knable and Weinberger, 1997; Perlstein et al., 2001) are the two brain regions that have been extensively studied in schizophrenia, cerebellum dysfunction has also been reported in schizophrenia (Yeganeh-Doost et al., 2011; Andreasen and Pierson, 2008). 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-Doost et al., 2011). Taken together, cerebellum might plays an important part in the etiology of schizophrenia and are therefore selected for the current study.

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 (Q. Li, C. Cheung, Wei, Hui, et al., 2009). A dose of 5mg kg^{-1} polyriboinosinic-polyribocytidilic acid (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. 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 (Reeves, Nielsen, and Fahey, 1993), 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 (Olivo and Hilakivi-Clarke, 2005). 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.

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.

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 (Andrews, n.d.), 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 remained in the final sequence. By using trim_glore, a wrapper for cutadapt (version 1.9.1) (Martin, 2011), 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 Engstrom et al. (2013), it was demonstrated that STAR (Dobin et al., 2013) 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) (Dobin et al., 2013). And the quantification of the gene expression levels were conducted using featureCounts (version 1.5.0) (Liao, Gordon K Smyth, and Shi, 2014).

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. 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 Seyednasrollah, Laiho, and Elo (2015), 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) (Love, Wolfgang Huber, and Simon Anders, 2014) 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 Yue et al. (2014), where most of the findings reported were found to be confounded by lane and batch effect Gilad and Mizrahi-Man (2015). 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.

To further investigate whether batch effect may can lead to false positives, we performed the likelihood ratio test (LRT) to investigate the effect of batch on our result. The LRT examines two models for the counts, a full model with a certain number of terms and a reduced model, in which some of the terms of the full model are removed. The test determines if the increased likelihood of the data using the extra terms in the full model is more than expected if those extra terms are truly zero. Thus we compared the full model $\sim Batch + Condition + Diet + Condition : Diet$ with $\sim Condition + Diet + Condition : Diet$ to understand the effect of batch on our data.

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 PSD (S M Purcell et al., 2014; T. N. Consortium and Psychiatric Genomics, 2015) and calcium ion channel (S M Purcell et al., 2014; Stephan Ripke, B. M. Neale, et al., 2014; Szatkiewicz et al., 2014) 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 S M Purcell et al. (2014) and T. N. Consortium and Psychiatric Genomics (2015) were retrieved. Gene sets and pathway related to PSD and calcium ion channel were also

retrieved from Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000) and Gene Ontology (GO) (T. G. O. Consortium, 2015), both have been widely used for systematic genetic analysis.

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.

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 T. N. Consortium and Psychiatric Genomics (2015)

Then, the partitioning of heritability was performed using LDSC (B. K. Bulik-Sullivan et al., 2015) --annot and --overlap-annot options, with window size of 1000kb window size and the LD score generated in section 2.2.8. The MHC region (chr6:25,000,000-35,000,000) was removed from the analysis due to its unusual LD and genetic architecture (Finucane et al., 2015).

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 (Busby et al., 2013). 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) (Dobin et al., 2013).

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.

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)

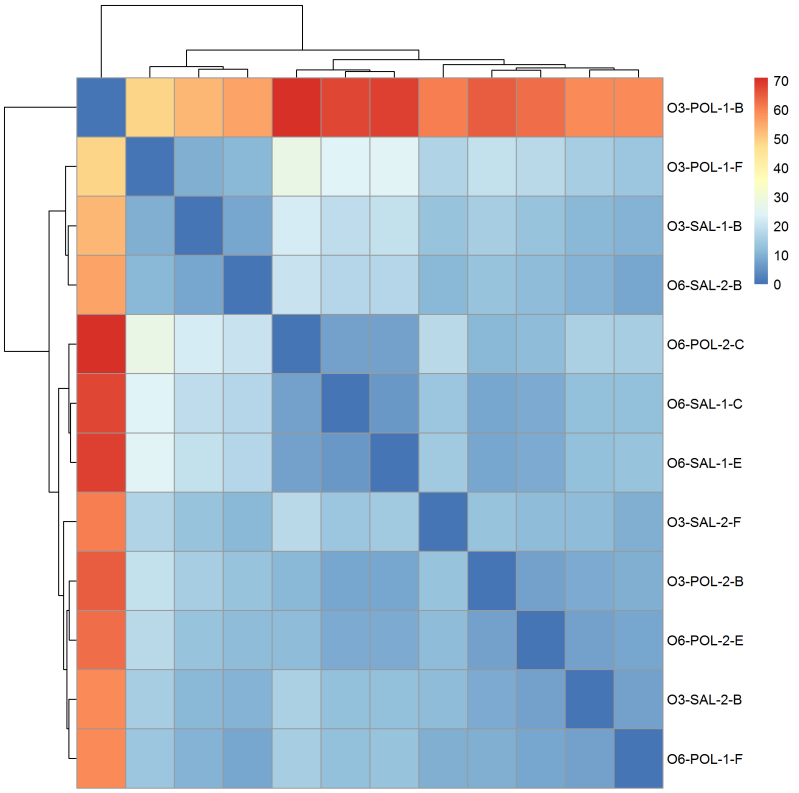
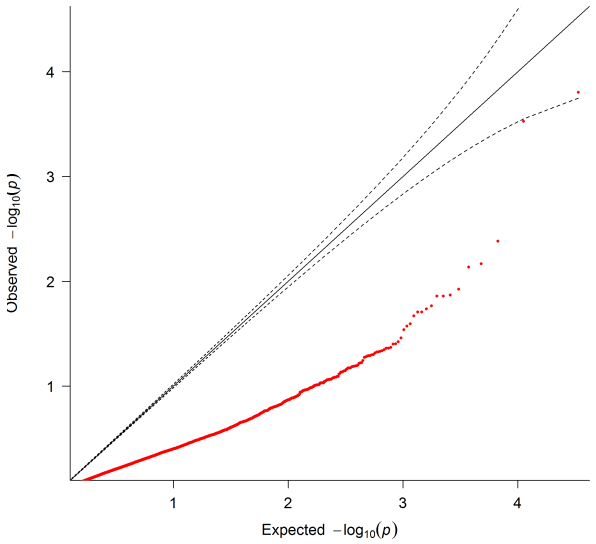
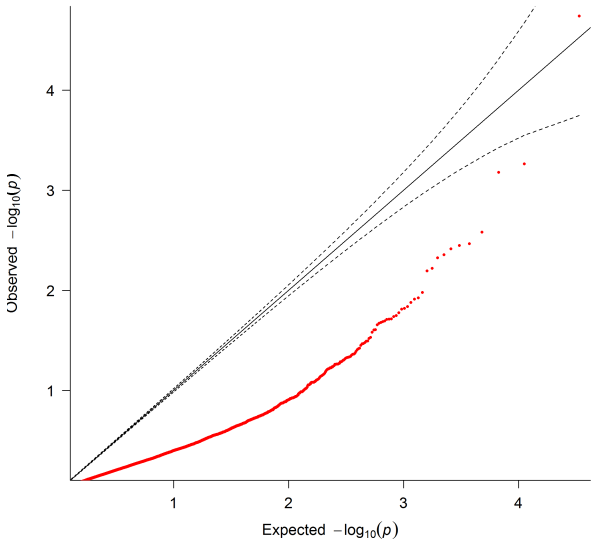


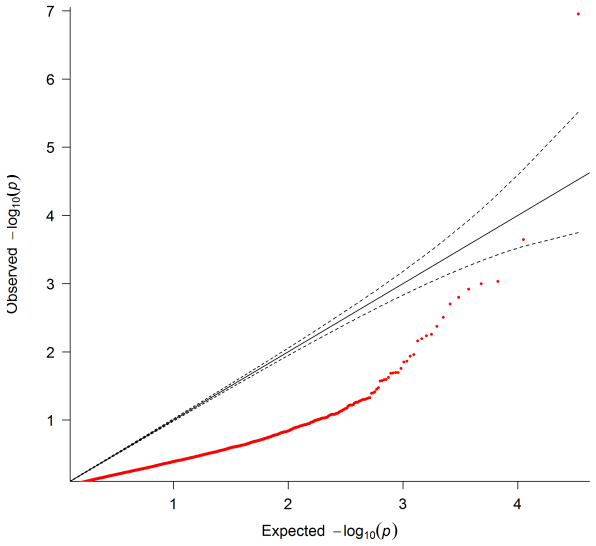
Figure 3.1: Sample Clustering results. Samples were labeled as <Diet>-<Condition>-<Lane>-<Batch> where O3 = n-3 PUFA rich diet; O6 = n-6 PUFA rich diet; POL = PolyI:C; SAL = Saline.
No clear clustering for lane or batch effects are observed. However, one sample from the n3-PUFA-PolyI:C group is found 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



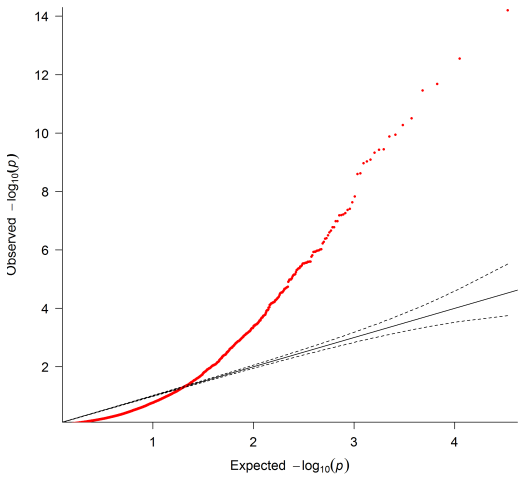
(a) O6-Saline mice vs O6-PolyI:C mice



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



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



(d) Batch Effect

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.
Meanwhile, the results from LRT suggested that the full model, which adjusted for batch effect, might provide a better fit to our data. Therefore it is important to adjust for the batch effect.

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).

LRT was performed to test the goodness of fit of model with and without the batch effect include. It is observed that when "Batch" was not included in the model, 178 genes are found to be significant (fig. 3.2d). This indicates that by including "Batch" in the statistic model, a significant better fit can be obtained.

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.

For all the gene sets related to PSD, the PSD gene set from S M Purcell et al. (2014) is the only one that is not found to be significant in all conditions. Upon further investigation, the PSD gene set from S M Purcell et al. (2014) is found to be based on the work of Kirov et al. (2012) 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.

Table 3.2: Results of gene set analysis. In total, 7 gene sets were retrieved from S M Purcell et al. (2014), KEGG and GO. Firstly, Wilcoxon Rank sum test was performed. Except for the PSD gene set obtained from S M Purcell et al. (2014), 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 S M Purcell et al. (2014) was found to be based on the work of Kirov et al. (2012) 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 S M Purcell et al. (2014) might explain the difference observed between the PSD set from S M Purcell et al. (2014) and PSD set from GO.

Gene Set	Source	ID	Category	Diet in PolyIC Mice	MIA Effect	Diet in Saline Mice	Proportion of h^2 explained	Enrichment P-value
Calcium Ion Signaling Pathway	KEGG	hsa04020	Calcium Ion	0.0402	4.40×10^{-7}	0.231	0.0135	0.421
Glutamatergic synapse	KEGG	hsa04724	PSD	0.118	0.00490	0.123	0.0134	0.382
Voltage-Gated Calcium Channel Activity	GO	GO:05245	Calcium Ion	0.0262	3.45×10^{-6}	0.137	0.00771	0.313
Calcium Channel Activity	GO	GO:05262	Calcium Ion	0.0942	0.00209	0.0880	0.0119	0.593
PSD	GO	GO:14069	PSD	4.86×10^{-3}	6.31×10^{-9}	0.0383	0.0352	0.00624
PSD	Purcell		PSD	0.113	0.328	0.977	0.0486	0.131
GWAS	Purcell		GWAS	0.3048	6.91×10^{-3}	0.551	0.0998	7.42×10^{-8}

3.3.4 Designing the Replication Study

Using Scotty (Busby et al., 2013), 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 (Tsai et al., 2002; Lee et al., 2003). For example, Tsai et al. (2002) 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 (Tsai et al., 2002).

On the other hand, it was found that *Sgk1* can regulates the AMPA and kainate glutamate receptors, especially GluR6, which is encoded by *Grik2* (Lang, Böhmer, et al., 2006; Lang, Strutz-Seebohm, et al., 2010). The kainate receptors contribute to the excitatory postsynaptic current and are important to the synaptic transmission and plasticity in the hippocampus (Lang, Böhmer, et al., 2006). The upregulation of AMPA and kainate receptors are therefore expected to enhance the excitatory effects of glutamate (Lang, Strutz-Seebohm, et al., 2010).

Furthermore, *Sgk1* can up-regulates the glutamate transporters such as EAAT4 (Bohmer et al., 2004), which are vital for the clearance of glutamate from

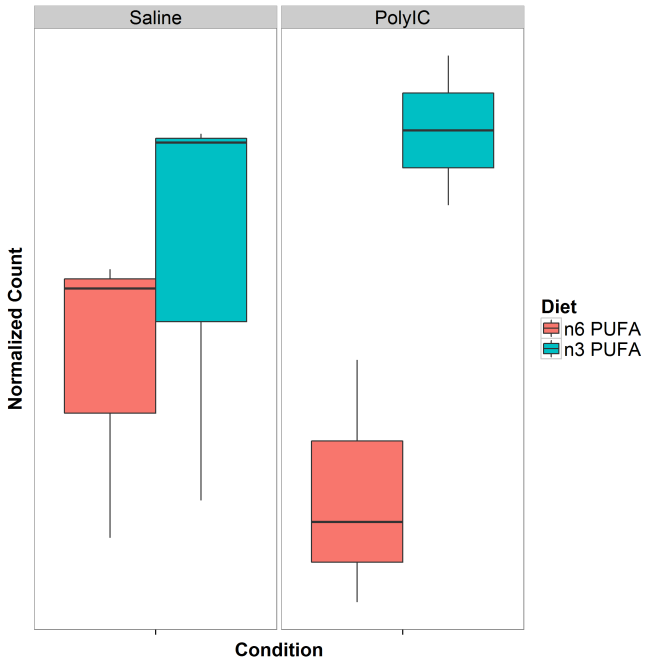


Figure 3.3: Normalized Expression of *Sgk1*. It was observed that the expression level of *Sgk1* increases after the mice was given a n3-PUFA rich diet where a significant increase was observed in mice exposed to PolyI:C.

the synaptic cleft. This prevents excessive glutamate accumulation, thus help to prevent the neurotoxic effects of glutamate (Lang, Strutz-Seebohm, et al., 2010). In addition, Schoenebeck et al. (2005) 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 exposed 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*, Zhang et al.

(2015) 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* (Lang, Strutz-Seebohm, et al., 2010).

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.

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 (Lidow, 2003; S M Purcell et al., 2014; Stephan Ripke, B. M. Neale, et al., 2014). For example, in exome sequencing study of schizophrenia conducted by S M Purcell et al. (2014), 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 (Stephan Ripke, B. M. Neale, et al., 2014).

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 (Berridge, 2014). Although it is unclear the exact role of the calcium signaling pathway in the etiology of , 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 ot 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 (Cline, 2005).

Most importantly, it is observed that the gene set containing genes within the associated GWAS LD-intervals from S M Purcell et al. (2014) 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, where their expression might change. Thus, genetic variants associated with and differential expression induced by early MIA might be affecting similar genes. The converging evidences suggested these genes might serve as an important candidates for future functional study in order to understand the etiology of schizophrenia.

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 (Bazinet and Laye, 2014; Calon et al., 2005). Subsequent research shown that the expression of PSD proteins are significantly down-regulated in n-3 PUFA depleted mouse brains (Sidhu, Huang, and H.-Y. Kim, 2011). Sidhu, Huang, and H.-Y. Kim (2011) 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 B. K. Bulik-Sullivan et al. (2015). It is observed that only the PSD gene set from GO and the gene set containing genes within the associated GWAS LD-intervals from S M Purcell et al. (2014) are found to have a significant disproportionate contribution to the heritability of schizophrenia. Our results are consistent with previous findings in T. N. Consortium and Psychiatric Genomics (2015), where only the GO PSD gene set is found to be significantly enriched by *common* variants.

Not surprisingly, the GWAS gene set is significantly enriched by common variants included in the PGC schizophrenia GWAS and is accounting for around 10% of the SNP heritability. On the other hand, as all included gene sets have been

reported to enriched by genetic variation observed in schizophrenia, the insignificant enrichment only suggest that they are not enriched by *common* genetic variants. Instead, it is likely that these other gene sets are more affected by the rare genetic variants such as CNV. Therefore, further genetic studies on the involvement of these gene sets in the etiology of schizophrenia might want to focus on identify rare variants instead of common variants.

To conclude, 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 schizophrenia in the etiology of schizophrenia. Considering 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 may therefore 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 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 (Bergeron et al., 2013; Lein et al., 2007). 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.

Forthly, a high RNA expression level does not guarantee a high protein concentration (Vogel and Marcotte, 2012). 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 (Vogel and Marcotte, 2012). 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 (Jiang et al., 2011).

4 Conclusion

With the rapid advancement of technologies, it is now possible to perform association studies on genome-wide scale using Genome Wide Association Study (GWAS). After years of failure in research of schizophrenia genetics, the Schizophrenia Working group of Psychiatric Genomics Consortium (PGC) has finally identified 108 genetic loci associated with schizophrenia (Stephan Ripke, B. M. Neale, et al., 2014).

In this thesis, we presented SNP HeRitability Estimation Kit (SHREK), an robust algorithm for the estimation of Single Nucleotide Polymorphism (SNP) heritability using summary statistics from GWAS, an alternative to LD SCore regression (LDSC). Through simulations, it was suggested that when compared to LDSC, SHREK can provide a more robust estimate for oligogentic traits and in case-control designs where no confounding variables was present. Using the latest GWAS summary statistics released by the PGC, we estimated that schizophrenia has a SNP-heritability of 0.185 (SD=0.00450), which is similar to the estimate of 0.198 (SD=0.0057) by LDSC.

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

On the other hand, we also performed an RNA sequencing on the polyriboinosinic-

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

4.1 Schizophrenia: Future Perspectives

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

On the other hand, when estimating the SNP-heritability of schizophrenia, it was found that no more than 20% of the heritability has been accounted for by the current GWAS which is lower than the 81% estimated based on twin studies

(Sullivan, Kendler, and M. C. Neale, 2003). This suggested that factors other than common SNP were contributing to the heritability of schizophrenia.

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

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

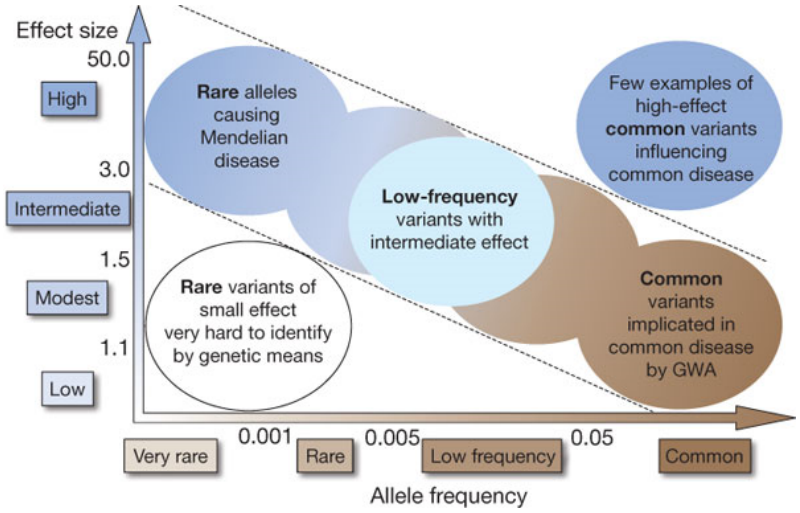


Figure 4.1: Relationship between effect size and allele frequency. It is expected that rare variants with large effect size were actively selected against in the population and therefore should be rare.

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

environmental interaction that contributes to the disease etiology of schizophrenia.

With the sophistication of technologies, we can now perform whole genome sequencing with the HiSeq X Ten system costing less than $1,000. Therefore, the largest challenge now resides in how to make sense of the data instead of data generation. For example, the alignment of sequence read to low complexity sequence or low-degeneracy repeats remains challenging and might be error prone, thus have a negative impact to the quality of the results(Sims et al., 2014). New sequencing technology such as Oxford Nanopore which can provide extra long-reads, might help to make alignment easier due to the extra information for each individual reads. However, the Oxford Nanopore is still under development and has a relatively high error rate (Mikheyev and Tin, 2014). Only until the error rate is dramatically decreased can the use of Oxford Nanopore system become feasible.

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

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

mechanism of schizophrenia be obtained.

Bibliography

Altshuler, David M et al. (2010). "Integrating common and rare genetic variation in diverse human populations." In: *Nature* 467.7311, pp. 52–58 (cit. on pp. 50, 51).

American Psychiatric Association (2013). *Diagnostic and Statistical Manual of Mental Disorders*. American Psychiatric Publishing, p. 991 (cit. on pp. 1, 28).

Anders, S and W Huber (2010). "Differential expression analysis for sequence count data". eng. In: *Genome Biol* 11.10, R106 (cit. on p. 34).

Andreasen, Nancy C and Ronald Pierson (2008). "The role of the cerebellum in schizophrenia." eng. In: *Biological psychiatry* 64.2, pp. 81–88 (cit. on p. 111).

Andrews, S. *FastQC A Quality Control tool for High Throughput Sequence Data* (cit. on p. 114).

Bazinet, Richard P and Sophie Laye (2014). "Polyunsaturated fatty acids and their metabolites in brain function and disease". In: *Nat Rev Neurosci* 15.12, pp. 771–785 (cit. on p. 127).

Bergeron, J D et al. (2013). "White matter injury and autistic-like behavior predominantly affecting male rat offspring exposed to group B streptococcal maternal inflammation". eng. In: *Dev Neurosci* 35.6, pp. 504–515 (cit. on p. 128).

Bernstein, Bradley E et al. (2010). "The NIH Roadmap Epigenomics Mapping Consortium." eng. In: *Nature biotechnology* 28.10, pp. 1045–1048 (cit. on p. 21).

Berridge, Michael J (2014). "Calcium signalling and psychiatric disease: bipolar disorder and schizophrenia." eng. In: *Cell and tissue research* 357.2, pp. 477–492 (cit. on p. 126).

Bohmer, Christoph et al. (2004). "Stimulation of the EAAT4 glutamate transporter by SGK protein kinase isoforms and PKB." eng. In: *Biochemical and biophysical research communications* 324.4, pp. 1242–1248 (cit. on p. 123).

Bouchard, Thomas J (2013). "The Wilson Effect: the increase in heritability of IQ with age." In: *Twin research and human genetics : the official journal of the International Society for Twin Studies* 16.5, pp. 923–30 (cit. on p. 4).

Brown, A S and E J Derkits (2010). "Prenatal infection and schizophrenia: a review of epidemiologic and translational studies". eng. In: *Am J Psychiatry* 167.3, pp. 261–280 (cit. on pp. 26–28, 31, 109, 136).

Brown, Alan S (2012). "Epidemiologic studies of exposure to prenatal infection and risk of schizophrenia and autism." eng. In: *Developmental neurobiology* 72.10, pp. 1272–1276 (cit. on p. 29).

Bulik-Sullivan, Brendan (2015). *Replicating MDD heritability Estimation* (cit. on p. 94).

Bulik-Sullivan, Brendan K et al. (2015). "LD Score regression distinguishes confounding from polygenicity in genome-wide association studies". In: *Nature Genetics* 47.3, pp. 291–295 (cit. on pp. 19, 20, 23, 35, 37, 38, 55, 63, 65, 69, 84, 88, 94–96, 117, 127).

Busby, Michele A et al. (2013). "Scotty: a web tool for designing RNA-Seq experiments to measure differential gene expression". In: *Bioinformatics* 29.5, pp. 656–657 (cit. on pp. 118, 123).

Cadenhead, K S et al. (2000). "Modulation of the startle response and startle laterality in relatives of schizophrenic patients and in subjects with schizotypal personality disorder: evidence of inhibitory deficits." eng. In: *The American journal of psychiatry* 157.10, pp. 1660–1668 (cit. on p. 29).

Calon, Frederic et al. (2005). "Dietary n-3 polyunsaturated fatty acid depletion activates caspases and decreases NMDA receptors in the brain of a transgenic mouse model of Alzheimer's disease." eng. In: *The European journal of neuroscience* 22.3, pp. 617–626 (cit. on p. 127).

Clandinin, M T (1999). "Brain development and assessing the supply of polyunsaturated fatty acid." eng. In: *Lipids* 34.2, pp. 131–137 (cit. on p. 110).

Clarke, Mary C et al. (2009). "Evidence for an interaction between familial liability and prenatal exposure to infection in the causation of schizophrenia." eng. In: *The American journal of psychiatry* 166.9, pp. 1025–1030 (cit. on pp. 26, 35, 109, 136).

Cline, H (2005). "Synaptogenesis: a balancing act between excitation and inhibition". eng. In: *Curr Biol* 15.6, R203–5 (cit. on p. 126).

Consortium, The Gene Ontology (2015). "Gene Ontology Consortium: going forward". In: *Nucleic Acids Research* 43.D1, pp. D1049–D1056 (cit. on p. 117).

Consortium, The GTEx (2015). "The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans". In: *Science* 348.6235, pp. 648–660 (cit. on p. 137).

Consortium, The International HapMap (2005). "A haplotype map of the human genome". In: *Nature* 437, pp. 1299–1320 (cit. on p. 12).

Consortium, The Network and Pathway Analysis Subgroup of the Psychiatric Genomics (2015). "Psychiatric genome-wide association study analyses implicate neuronal, immune and histone pathways". In: *Nat Neurosci* 18.2, pp. 199–209 (cit. on pp. 110, 116, 117, 127).

Deverman, B E and P H Patterson (2009). "Cytokines and CNS development". eng. In: *Neuron* 64.1, pp. 61–78 (cit. on p. 14).

Dobin, A et al. (2013). "STAR: ultrafast universal RNA-seq aligner". eng. In: *Bioinformatics* 29.1, pp. 15–21 (cit. on pp. 33, 114, 118).

ENCODE Project Consortium (2012). "An integrated encyclopedia of DNA elements in the human genome". In: *Nature* 489.7414, pp. 57–74 (cit. on pp. 21, 137).

Engstrom, Par G et al. (2013). "Systematic evaluation of spliced alignment programs for RNA-seq data". In: *Nat Meth* 10.12, pp. 1185–1191 (cit. on p. 114).

Falconer, Douglas S (1965). "The inheritance of liability to certain diseases, estimated from the incidence among relatives". In: *Annals of Human Genetics* 29.1, pp. 51–76 (cit. on p. 7).

Falconer, Douglas S and Trudy F C Mackay (1996). *Introduction to Quantitative Genetics (4th Edition)*. Vol. 12, p. 464 (cit. on pp. 3, 5, 10).

Feuk, Lars, Andrew R Carson, and Stephen W Scherer (2006). "Structural variation in the human genome". In: *Nat Rev Genet* 7.2, pp. 85–97 (cit. on p. 24).

Finucane, Hilary K et al. (2015). "Partitioning heritability by functional annotation using genome-wide association summary statistics". In: *Nat Genet* advance online publication (cit. on pp. 21, 23, 117).

Fromer, M et al. (2014). "De novo mutations in schizophrenia implicate synaptic networks". eng. In: *Nature* 506.7487, pp. 179–184 (cit. on p. 25).

Garbett, K a et al. (2012). "Effects of maternal immune activation on gene expression patterns in the fetal brain". In: *Translational Psychiatry* 2.4, e98 (cit. on pp. 28–30).

Gilad, Yoav and Orna Mizrahi-Man (2015). "A reanalysis of mouse ENCODE comparative gene expression data." eng. In: *F1000Research* 4, p. 121 (cit. on p. 115).

Giles, Peter J and David Kipling (2003). "Normality of oligonucleotide microarray data and implications for parametric statistical analyses." eng. In: *Bioinformatics (Oxford, England)* 19.17, pp. 2254–2262 (cit. on p. 34).

Giovanoli, S. et al. (2013). "Stress in puberty unmasks latent neuropathological consequences of prenatal immune activation in mice". eng. In: *Science* 339.6123, pp. 1095–1099 (cit. on p. 31).

Golan, David, Eric S Lander, and Saharon Rosset (2014). "Measuring missing heritability: Inferring the contribution of common variants". In: *Proceedings of the National Academy of Sciences* 111.49, E5272–E5281 (cit. on pp. 37, 75, 78, 90–92).

Gottesman, Irving I (1991). *Schizophrenia genesis: The origins of madness.* WH Freeman/Times Books/Henry Holt & Co (cit. on p. 11).

Gottesman, Irving I and James Shields (1982). *Schizophrenia: The Epigenetic Puzzle.* Cambridge University Press (cit. on p. 11).

Gottesman, Irving I and J Shields (1967a). "A polygenic theory of schizophrenia". In: *Proceedings of the National Academy of Sciences* 58.1, pp. 199–205 (cit. on pp. 10, 11).

— (1967b). "A polygenic theory of schizophrenia". In: *Proceedings of the National Academy of Sciences* 58.1, pp. 199–205 (cit. on p. 26).

Guey, Lin T. et al. (2011). "Power in the phenotypic extremes: A simulation study of power in discovery and replication of rare variants". In: *Genetic Epidemiology* 35.4, pp. 236–246 (cit. on p. 47).

Gui, Hongsheng et al. (2013). "RET and NRG1 interplay in Hirschsprung disease." eng. In: *Human genetics* 132.5, pp. 591–600 (cit. on p. 57).

Hansen, Per Christian (1987). "The truncated SVD as a method for regularization". In: *Bit* 27.4, pp. 534–553 (cit. on p. 50).

Harrison, P J and D R Weinberger (2005). "Schizophrenia genes, gene expression, and neuropathology: on the matter of their convergence." In: *Molecular psychiatry* 10.1, 40–68, image 5 (cit. on p. 12).

Heston, Leonard L (1966). "Psychiatric Disorders in Foster Home Reared Children of Schizophrenic Mothers". In: *The British Journal of Psychiatry* 112.489, pp. 819–825 (cit. on p. 9).

Hinrichs, A S et al. (2006). "The UCSC Genome Browser Database: update 2006." eng. In: *Nucleic acids research* 34.Database issue, pp. D590–8 (cit. on p. 62).

Hoyle, David C et al. (2002). "Making sense of microarray data distributions." eng. In: *Bioinformatics (Oxford, England)* 18.4, pp. 576–584 (cit. on p. 34).

Jiang, Lichun et al. (2011). "Synthetic spike-in standards for RNA-seq experiments". In: *Genome Research* 21.9, pp. 1543–1551 (cit. on p. 132).

Kanehisa, M and S Goto (2000). "KEGG: kyoto encyclopedia of genes and genomes". eng. In: *Nucleic Acids Res* 28.1, pp. 27–30 (cit. on p. 117).

Kelly, C and R G McCreadie (1999). "Smoking habits, current symptoms, and premorbid characteristics of schizophrenic patients in Nithsdale, Scotland." eng. In: *The American journal of psychiatry* 156.11, pp. 1751–1757 (cit. on p. 26).

Kim, Daehwan et al. (2013). "TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions". In: *Genome Biology* 14.4, R36 (cit. on p. 33).

Kirov, G et al. (2012). "De novo CNV analysis implicates specific abnormalities of postsynaptic signalling complexes in the pathogenesis of schizophrenia." eng. In: *Molecular psychiatry* 17.2, pp. 142–153 (cit. on pp. 121, 122).

Kitajka, Klára et al. (2002). "The role of n-3 polyunsaturated fatty acids in brain: Modulation of rat brain gene expression by dietary n-3 fatty acids". In: *Proceedings of the National Academy of Sciences of the United States of America* 99.5, pp. 2619–2624 (cit. on p. 110).

Knable, M B and D R Weinberger (1997). "Dopamine, the prefrontal cortex and schizophrenia." eng. In: *Journal of psychopharmacology (Oxford, England)* 11.2, pp. 123–131 (cit. on p. 111).

Knapp, Martin, Roshni Mangalore, and Judit Simon (2004). "The global costs of schizophrenia." In: *Schizophrenia bulletin* 30.2, pp. 279–293 (cit. on p. 1).

Lander, E S et al. (2001). "Initial sequencing and analysis of the human genome." eng. In: *Nature* 409.6822, pp. 860–921 (cit. on p. 12).

Lang, Florian, Christoph Böhmer, et al. (2006). "(Patho)physiological Significance of the Serum- and Glucocorticoid-Inducible Kinase Isoforms". In: *Physiological Reviews* 86.4, pp. 1151–1178 (cit. on p. 123).

Lang, Florian, Nathalie Strutz-Seebohm, et al. (2010). "Significance of SGK1 in the regulation of neuronal function". In: *The Journal of Physiology* 588.18, pp. 3349–3354 (cit. on pp. 123–125).

Lee, Eminy H Y et al. (2003). "Enrichment enhances the expression of sgk, a glucocorticoid-induced gene, and facilitates spatial learning through glutamate AMPA receptor mediation." eng. In: *The European journal of neuroscience* 18.10, pp. 2842–2852 (cit. on p. 123).

Lein, E S et al. (2007). "Genome-wide atlas of gene expression in the adult mouse brain". eng. In: *Nature* 445.7124, pp. 168–176 (cit. on p. 128).

Li, Bo and Colin N Dewey (2011). "RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome." eng. In: *BMC bioinformatics* 12, p. 323 (cit. on p. 33).

Li, Miao-Xin Xin et al. (2011). "Evaluating the effective numbers of independent tests and significant p-value thresholds in commercial genotyping arrays and public imputation reference datasets". In: *Human Genetics* 131.5, pp. 747–756 (cit. on pp. 13, 45).

Li, Na and Matthew Stephens (2003). "Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data." eng. In: *Genetics* 165.4, pp. 2213–2233 (cit. on p. 53).

Li, Q, C Cheung, R Wei, V Cheung, et al. (2010). "Voxel-based analysis of postnatal white matter microstructure in mice exposed to immune challenge in early or late pregnancy". eng. In: *Neuroimage* 52.1, pp. 1–8 (cit. on pp. 29, 32).

Li, Q, C Cheung, R Wei, E S Hui, et al. (2009). "Prenatal immune challenge is an environmental risk factor for brain and behavior change relevant to schizophre-

nia: evidence from MRI in a mouse model". eng. In: *PLoS One* 4.7, e6354 (cit. on pp. 29, 32, 112).

Li, Q, Y O Leung, et al. (2015). "Dietary supplementation with n-3 fatty acids from weaning limits brain biochemistry and behavioural changes elicited by prenatal exposure to maternal inflammation in the mouse model." eng. In: *Translational psychiatry* 5, e641 (cit. on pp. 36, 110, 134).

Liao, Yang, Gordon K Smyth, and Wei Shi (2014). "featureCounts: an efficient general purpose program for assigning sequence reads to genomic features." eng. In: *Bioinformatics (Oxford, England)* 30.7, pp. 923–930 (cit. on p. 114).

Lichtenstein, Paul et al. (2009). "Common genetic determinants of schizophrenia and bipolar disorder in Swedish families: a population-based study". In: *The Lancet* 373.9659, pp. 234–239 (cit. on pp. 10, 20, 133).

Lidow, Michael S (2003). "Calcium signaling dysfunction in schizophrenia: a unifying approach." eng. In: *Brain research. Brain research reviews* 43.1, pp. 70–84 (cit. on p. 125).

Love, Michael I, Wolfgang Huber, and Simon Anders (2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." eng. In: *Genome biology* 15.12, p. 550 (cit. on p. 115).

Marioni, J C et al. (2008). "RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays". eng. In: *Genome Res* 18.9, pp. 1509–1517 (cit. on p. 34).

Martin, Marcel (2011). "Cutadapt removes adapter sequences from high-throughput sequencing reads". In: *EMBnet.journal; Vol 17, No 1: Next Generation Sequencing Data Analysis* (cit. on p. 114).

McGrath, John et al. (2008). "Schizophrenia: A Concise Overview of Incidence, Prevalence, and Mortality". In: *Epidemiologic Reviews* 30.1, pp. 67–76 (cit. on p. 26).

Mednick (1988). "Schizophrenia Following Prenatal Exposure to an Influenza Epidemic". In: *Arch Gen Psychiatry* 45.1 (cit. on p. 27).

Meyer, U, B K Yee, and J Feldon (2007). "The neurodevelopmental impact of prenatal infections at different times of pregnancy: the earlier the worse?" eng. In: *Neuroscientist* 13.3, pp. 241–256 (cit. on pp. 28, 30–32, 136).

Meyer, Urs, Joram Feldon, and S Hossein Fatemi (2009). "In-vivo rodent models for the experimental investigation of prenatal immune activation effects in neurodevelopmental brain disorders". In: *Neuroscience & Biobehavioral Reviews* 33.7, pp. 1061–1079 (cit. on p. 29).

Mikheyev, Alexander S and Mandy M Y Tin (2014). "A first look at the Oxford Nanopore MinION sequencer." eng. In: *Molecular ecology resources* 14.6, pp. 1097–1102 (cit. on p. 137).

Neumaier, Arnold (1998). "Solving Ill-Conditioned and Singular Linear Systems: A Tutorial on Regularization". In: *SIAM Review* 40.3, pp. 636–666 (cit. on p. 48).

Nishioka, Masaki et al. (2012). "DNA methylation in schizophrenia: progress and challenges of epigenetic studies." eng. In: *Genome medicine* 4.12, p. 96 (cit. on p. 135).

Nugent, Tom F. et al. (2007). "Dynamic mapping of hippocampal development in childhood onset schizophrenia". In: *Schizophrenia Research* 90.1-3, pp. 62–70 (cit. on p. 111).

O'Callaghan, E et al. (1991). "Season of birth in schizophrenia. Evidence for confinement of an excess of winter births to patients without a family history of mental disorder." eng. In: *The British journal of psychiatry : the journal of mental science* 158, pp. 764–769 (cit. on p. 26).

Olivo, Susan E and Leena Hilakivi-Clarke (2005). "Opposing effects of prepubertal low- and high-fat n-3 polyunsaturated fatty acid diets on rat mammary tumorigenesis." eng. In: *Carcinogenesis* 26.9, pp. 1563–1572 (cit. on p. 112).

Orr, H Allen (1998). "The Population Genetics of Adaptation: The Distribution of Factors Fixed during Adaptive Evolution". In: *Evolution* 52.4, pp. 935–949 (cit. on p. 53).

Oskviga, Devon B. et al. (2012). "Maternal immune activation by LPS selectively alters specific gene expression profiles of interneuron migration and oxidative stress in the fetus without triggering a fetal immune response". In: *Brain, Behavior, and Immunity* 26.4, pp. 623–634 (cit. on p. 29).

Peloso, Gina M et al. (2015). "Phenotypic extremes in rare variant study designs." ENG. In: *European journal of human genetics : EJHG* (cit. on p. 79).

Perlstein, W M et al. (2001). "Relation of prefrontal cortex dysfunction to working memory and symptoms in schizophrenia." eng. In: *The American journal of psychiatry* 158.7, pp. 1105–1113 (cit. on p. 111).

Peters, Bas J M et al. (2010). "Methodological and statistical issues in pharmacogenomics." eng. In: *The Journal of pharmacy and pharmacology* 62.2, pp. 161–166 (cit. on p. 13).

Price, Alkes L et al. (2006). "Principal components analysis corrects for stratification in genome-wide association studies". In: *Nature Genetics* 38, pp. 904–909 (cit. on p. 90).

Project, Genomes et al. (2012). "An integrated map of genetic variation from 1,092 human genomes". In: *Nature* 491.7422, pp. 56–65 (cit. on pp. 51, 62).

Psychiatric GWAS Consortium Bipolar Disorder Working Group (2011). "Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4." eng. In: *Nature genetics* 43.10, pp. 977–983 (cit. on p. 62).

Purcell, S M et al. (2014). "A polygenic burden of rare disruptive mutations in schizophrenia". eng. In: *Nature* 506.7487, pp. 185–190 (cit. on pp. 25, 109, 110, 116, 121, 122, 125–127, 135).

Purcell, S, S S Cherny, and P C Sham (2003). "Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits". en. In: *Bioinformatics* 19, pp. 149–150 (cit. on p. 13).

Purcell, Shaun et al. (2007). "PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses". In: *The American Journal of Human Genetics* 81.3, pp. 559–575 (cit. on pp. 53, 56).

Reeves, P G, F H Nielsen, and G C Jr Fahey (1993). *AIN-93 purified diets for laboratory rodents: final report of the American Institute of Nutrition ad hoc writing committee on the reformulation of the AIN-76A rodent diet.* eng (cit. on p. 112).

Rijsdijk, Fruhling V and Pak C Sham (2002). "Analytic approaches to twin data using structural equation models." eng. In: *Briefings in bioinformatics* 3.2, pp. 119–133 (cit. on p. 10).

Riley, Brien and Kenneth S Kendler (2006). "Molecular genetic studies of schizophrenia." In: *European journal of human genetics : EJHG* 14.6, pp. 669–680 (cit. on pp. 11, 96).

Ripke, Stephan, Benjamin M. Neale, et al. (2014). "Biological insights from 108 schizophrenia-associated genetic loci". In: *Nature* 511, pp. 421–427 (cit. on pp. 14, 15, 20, 21, 25, 62, 84, 110, 116, 125, 133, 134).

Ripke, Stephan, Naomi R Wray, et al. (2013). "A mega-analysis of genome-wide association studies for major depressive disorder." eng. In: *Molecular psychiatry* 18.4, pp. 497–511 (cit. on p. 62).

Ripke, S et al. (2013). "Genome-wide association analysis identifies 13 new risk loci for schizophrenia". eng. In: *Nat Genet* 45.10, pp. 1150–1159 (cit. on p. 21).

Risch, N (1990). "Linkage strategies for genetically complex traits. II. The power of affected relative pairs." In: *American Journal of Human Genetics* 46.2, pp. 229–241 (cit. on pp. 11, 12).

Robinson, M D, D J McCarthy, and G K Smyth (2010). "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data". eng. In: *Bioinformatics* 26.1, pp. 139–140 (cit. on p. 34).

Saha, Sukanta, David Chant, and John Mcgrath (2007). "A Systematic Review of Mortality in Schizophrenia". In: *Archives of general psychiatry* 64.10, pp. 1123–1131 (cit. on pp. 1, 135).

Sanderson, Conrad (2010). *Armadillo: An Open Source C++ Linear Algebra Library for Fast Prototyping and Computationally Intensive Experiments*. Tech. rep. (cit. on p. 97).

Schoenebeck, Bodo et al. (2005). "Sgk1, a cell survival response in neurodegenerative diseases." eng. In: *Molecular and cellular neurosciences* 30.2, pp. 249–264 (cit. on p. 124).

Seyednasrollah, Fatemeh, Asta Laiho, and Laura L Elo (2015). "Comparison of software packages for detecting differential expression in RNA-seq studies". In: *Briefings in Bioinformatics* 16.1, pp. 59–70 (cit. on p. 115).

Sham, Pak C and Shaun M Purcell (2014). "Statistical power and significance testing in large-scale genetic studies." In: *Nature reviews. Genetics* 15.5, pp. 335–46 (cit. on pp. 47, 61, 62, 92).

Sidhu, Vishaldeep K, Bill X Huang, and Hee-Yong Kim (2011). "Effects of docosahexaenoic acid on mouse brain synaptic plasma membrane proteome analyzed by mass spectrometry and (16)O/(18)O labeling". In: *Journal of proteome research* 10.12, pp. 5472–5480 (cit. on p. 127).

Sims, David et al. (2014). "Sequencing depth and coverage: key considerations in genomic analyses". In: *Nat Rev Genet* 15.2, pp. 121–132 (cit. on p. 137).

Smith, S E et al. (2007). "Maternal immune activation alters fetal brain development through interleukin-6". eng. In: *J Neurosci* 27.40, pp. 10695–10702 (cit. on pp. 29, 110, 136).

Su, Zhan, Jonathan Marchini, and Peter Donnelly (2011). "HAPGEN2: Simulation of multiple disease SNPs". In: *Bioinformatics* 27.16, pp. 2304–2305 (cit. on pp. 50, 52, 56).

Sullivan, Patrick F (2005). "The Genetics of Schizophrenia". In: *PLoS Med* 2.7, e212 (cit. on p. 27).

Sullivan, Patrick F, Kenneth S Kendler, and Michael C Neale (2003). "Schizophrenia as a Complex Trait". In: *Archives of general psychiatry* 60, pp. 1187–1192 (cit. on pp. 10, 20, 133, 135).

Sutovsky, Peter et al. (1999). "Development: Ubiquitin tag for sperm mitochondria". In: *Nature* 402.6760, pp. 371–372 (cit. on p. 97).

Szatkiewicz, J P et al. (2014). "Copy number variation in schizophrenia in Sweden". In: *Mol Psychiatry* 19.7, pp. 762–773 (cit. on pp. 24, 25, 110, 116).

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

Tienari, Pekka et al. (2004). "Genotype-environment interaction in schizophrenia-spectrum disorder". In: *The British Journal of Psychiatry* 184.3, pp. 216–222 (cit. on pp. 26, 35, 109).

Trapnell, Cole et al. (2012). "Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks". In: *Nat. Protocols* 7.3, pp. 562–578 (cit. on p. 34).

Trebble, Timothy et al. (2003). "Inhibition of tumour necrosis factor-alpha and interleukin 6 production by mononuclear cells following dietary fish-oil supplementation in healthy men and response to antioxidant co-supplementation." eng. In: *The British journal of nutrition* 90.2, pp. 405–412 (cit. on p. 110).

Tsai, Kuen J et al. (2002). "sgk, a primary glucocorticoid-induced gene, facilitates memory consolidation of spatial learning in rats." eng. In: *Proceedings of the National Academy of Sciences of the United States of America* 99.6, pp. 3990–3995 (cit. on p. 123).

Velakoulis, Dennis et al. (2006). "Hippocampal and amygdala volumes according to psychosis stage and diagnosis". In: *Archives of general psychiatry* 63, pp. 139–149 (cit. on p. 111).

Visscher, Peter M, William G Hill, and Naomi R Wray (2008). "Heritability in the genomics era [mdash] concepts and misconceptions". In: *Nat Rev Genet* 9.4, pp. 255–266 (cit. on pp. 5, 7).

Vogel, Christine and Edward M Marcotte (2012). "Insights into the regulation of protein abundance from proteomic and transcriptomic analyses." eng. In: *Nature reviews. Genetics* 13.4, pp. 227–232 (cit. on p. 129).

Vuillermot, Stéphanie et al. (2010). "A longitudinal examination of the neurodevelopmental impact of prenatal immune activation in mice reveals primary defects in dopaminergic development relevant to schizophrenia". eng. In: *J Neurosci* 30.4, pp. 1270–1287 (cit. on p. 30).

Walsh, Tom et al. (2008). "Rare Structural Variants Disrupt Multiple Genes in Neurodevelopmental Pathways in Schizophrenia". In: *Science* 320.5875, pp. 539–543 (cit. on pp. 24, 25).

Wang, K et al. (2010). "MapSplice: accurate mapping of RNA-seq reads for splice junction discovery". eng. In: *Nucleic Acids Res* 38.18, e178 (cit. on p. 33).

Wang, Zhongmiao and Bruce Thompson (2007). "Is the Pearson r 2 Biased, and if So, What Is the Best Correction Formula?" In: *The Journal of Experimental Education* 75.2, pp. 109–125 (cit. on p. 52).

Weir, B S and W G Hill (1980). "EFFECT OF MATING STRUCTURE ON VARIATION IN LINKAGE DISEQUILIBRIUM". In: *Genetics* 95.2, pp. 477–488 (cit. on pp. 52, 64, 86).

Welter, Danielle et al. (2014). "The NHGRI GWAS Catalog, a curated resource of SNP-trait associations". In: *Nucleic Acids Research* 42.D1, pp. 1001–1006 (cit. on p. 55).

Wockner, L F et al. (2014). "Genome-wide DNA methylation analysis of human brain tissue from schizophrenia patients". In: *Transl Psychiatry* 4, e339 (cit. on p. 135).

Wong, Emily H M et al. (2014). "Common variants on Xq28 conferring risk of schizophrenia in Han Chinese." eng. In: *Schizophrenia bulletin* 40.4, pp. 777–786 (cit. on p. 96).

World Health Organization (2013). *WHO methods and data sources for global burden of disease estimates*. Tech. rep. Geneva (cit. on p. 2).

Yang, Jian, Beben Benyamin, et al. (2010). "Common SNPs explain a large proportion of the heritability for human height." eng. In: *Nature genetics* 42.7, pp. 565–569 (cit. on p. 18).

Yang, Jian, Michael N Weedon, et al. (2011). "Genomic inflation factors under polygenic inheritance". In: *Eur J Hum Genet* 19.7, pp. 807–812 (cit. on pp. 18, 19).

Yang, J et al. (2011). "GCTA: a tool for genome-wide complex trait analysis". eng. In: *Am J Hum Genet* 88.1, pp. 76–82 (cit. on pp. 16, 17, 55).

Yeganeh-Doost, Peyman et al. (2011). "The role of the cerebellum in schizophrenia: from cognition to molecular pathways". In: *Clinics* 66.Suppl 1, pp. 71–77 (cit. on p. 111).

Yue, Feng et al. (2014). "A comparative encyclopedia of DNA elements in the mouse genome." eng. In: *Nature* 515.7527, pp. 355–364 (cit. on p. 115).

Zhang, Wenting et al. (2015). "n-3 Polyunsaturated Fatty Acids Reduce Neonatal Hypoxic/Ischemic Brain Injury by Promoting Phosphatidylserine Formation and Akt Signaling." eng. In: *Stroke; a journal of cerebral circulation* 46.10, pp. 2943–2950 (cit. on p. 124).

Zhao, B and J P Schwartz (1998). "Involvement of cytokines in normal CNS development and neurological diseases: recent progress and perspectives". eng. In: *J Neurosci Res* 52.1, pp. 7–16 (cit. on p. 14).

Zhao, Shanrong et al. (2014). "Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells". In: *PLoS ONE* 9.1. Ed. by Shu-Dong Zhang, e78644 (cit. on p. 33).

Zheng, Gang, Boris Freidlin, and Joseph L Gastwirth (2006). "Robust genomic control for association studies." eng. In: *American journal of human genetics* 78.2, pp. 350–356 (cit. on pp. 18, 95).