Understanding How Genetics and Environments Shape the Development of Schizophrenia

Choi Shing Wan

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

Department of Psychiatry
University of Hong Kong
Hong Kong
December 22, 2015

Abstract

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

Genome Wide Association Study (GWAS) is now a standard approach for investigating associations of common genetic variations (mainly the 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 (and other complex diseases) 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 estimated that prenatal infection may account for one-third of the cases of SCZ and 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

This thesis would not be possible without my supervisors, Professor Pak Sham, Dr Stacey Cherny and Dr Wanling Yeung and I would like to thank them for taking me in and giving my guidance whenever I need during my study. I would like to especially thanks Professor Pak Sham for his trust, guidance and support to me and provide me the valuable opportunity to work on different projects.

I am also blessed to have Dr Johnny Kwan, Dr Timothy Mak and Dr Desmond Campbell to have the patience to teach me all the statistical problems I have encountered during my studies and especially Dr Johnny Kwan for his constant guidances. Without my helpful and lovely colleagues, my life will be much different and I am in debt to them for giving me such a memorable and enjoyable time for the past 4 years. Thank you Beatrice Wu, Dr Li Qi, Tomy Hui, Vicki Lin, Nick Lin, John Wong, Dr Clara Tang, Dr Amy Butler, Dr Allen Gui, Dr Sylvia Lam, Yung Tse Choi, Oi Chi Chan Pui King Wong and Dr Miaoxin Li for their constant support and advice and for giving such a lovely atmosphere in the department.

Finally, I must thank you Beatrice Wu and my family for without their support and constant encouragement, I won't be able to complete my thesis.

THANK YOU!

Abbreviations

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

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

Contents

List of Figures

List of Tables

1 Introduction

1.1 Schizophrenia

Schizophrenia (SCZ) is a devastating psychiatric disorder affecting approximately $0.3 \sim 0.7\%$ of the population worldwide (American Psychiatric Association, 2013). According to one of the current standard classification manual Diagnostic and Statistical Manual of Mental Disorders (DSM)-V, a diagnosis of schizophrenia (F20.9) can only be reached if the patient suffered from 2 or more of the following symptoms for a significant portion of time during a 1-month period: 1) delusion; 2) hallucinations; 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). Signs of disturbance also need to persist for at least 6-month before the patient can be diagnosed with schizophrenia.

Because of the detrimental symptoms and the lack of effective treatments, schizophrenia imposes a long lasting health, social and financial burden to the patients and their families (Knapp, Mangalore, and Simon, 2004). Schizophrenia patient also have a higher tendency to suicide (Saha, Chant, and Mcgrath, 2007), leading to a higher mortality. Based on the World Health Organization (WHO) report, schizophrenia is 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), demonstrating the extent of impact from schizophrenia to patients. Due to the severity of schizophre-

Table 1.1: Top 20 leading cause of YLD calculated by WHO in year 2012. Schizophrenia was considered as one of the top 20 leading cause 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

nia, it has drawn much attention from the research community, hoping to delineate the disease mechanics and to identify risk factors associated with schizophrenia. Ultimately, the goal of schizophrenia research is to identify effective treatment(s) to help improving the quality of life of the patients.

1.2 Understanding the Disease Mechanism

An important first step in schizophrenia research is to understand whether if it is a genetic or environmental disorder. For example, if schizophrenia is a genetic disorder, then one should focus on collecting genetic data and identify genetic variants that might associate with schizophrenia. Yet if schizophrenia is an environmental disorder, one should instead focus on how the environmental factors affect the normal functioning of the patients. In order to study the relative contribution of genetic and environmental influence to individual differences in schizophrenia, one will need to calculate the *heritability* of schizophrenia. There are two definition of heritability: the broad sense heritability and the narrow sense heritability. The 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 is defined as the proportion of total variance of a trait in a population explained by the variation of *additive* genetic factors in the population.

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)

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

where the variance of the observed phenotype (σ_P^2) can be expressed as 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* measurement at a specific time point. As a result of that, the heritability estimation might differ from one population to another due to difference in minor allele frequency (maf) and one might obtain a different heritability estimate if the method or timepoint of measurement of the trait differs because of different environmental factors coming into play. A classic example was the study of intelligence quotient (IQ) where the heritability estimation increases with age (Bouchard, 2013). It was hypothesize that the shared environment has a larger effect on individuals when they were young, and as they become more independent, the effect of shared environment diminishes, leading to an *increased portion* of variance in IQ explained by the variance in genetic (Bouchard, 2013).

1.2.2 Narrow Sense Heritability

In reality, the problem of heritability was more complicated for there were different forms of genetic effects. For example, one can partition the genetic variance into variance of additive genetic effects (σ_A^2), variance of dominant genetic effects (σ_D^2) and other epistatic genetic effects (σ_I^2) such that

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

where additive genetic variance was the variance explained by the average effects of all loci involved in the determination of the trait, whereas dominant genetic effects and epistatic genetic effects were the interaction between alleles at the *same* locus or *different* loci respectively.

As individuals only transmit one copy of each allele to their offspring, relatives other than full siblings and identical twins will only share a maximum of one copy of the allele. Considering that dominance and non-additive genetic effects were the interactive effect, which usually involve more than one copy of the alleles, these effects are unlikely to contribute to the resemblance between relatives (Peter M Visscher, William G Hill, and Naomi R Wray, 2008). On the other hand, the additive genetic effects is usually transmitted from parent to offspring, thus it is more useful to consider the narrow sense heritability (h^2) which only consider the additive genetic effects:

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

To obtain the additive genetic effect, we can first consider the genetic effect of parents 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. On the other hand, as D' was specific to the child, it should be independent from A and D. Moreover, the covariance between the additive genetics and non-additive

genetics should be zero (Falconer and Mackay, 1996). Thus, eq. (1.2) becomes

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

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

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

If we consider the simple linear regression equation $Y = X\beta + \epsilon$, its 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 was that the relatives does not share anything other than the additive genetic factors. However, this was usually not the case as relatives does tends to be in the same cultural group and might have similar socio-economic status which might all contribute to the variance of the trait. This might therefore lead to bias in eq. (1.7) and we shall discuss the partitioning of variance in the later sections.

Nonetheless, eq. (1.7) was still useful for the understanding of the calcula-

tion of 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, we will need the concept of liability threshold model popularized by Falconer, 1965.

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 (Peter M Visscher, William G Hill, and Naomi R 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 such as schizophrenia where instead of having a continuous distribution of phenotype, only a dichotomous labeling of "affected" and "normal" were obtained. The variance of these phenotype were therefore more difficult to obtain.

Falconer (1965) proposed the liability threshold model, which suggesting 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 were affected by combination of multitude of genetics and environmental factors, each with a small effects, as in the case of the continuous traits. The main difference was that the phenotype of an individual is determined by whether if the combined effects of these factors ("liability") were above a particular threshold ("liability threshold"). So for example, in the case of schizophrenia, only when an individual has a liability above the liability threshold will he/she be affected (fig. 1.1). One can then estimate the heritability of the

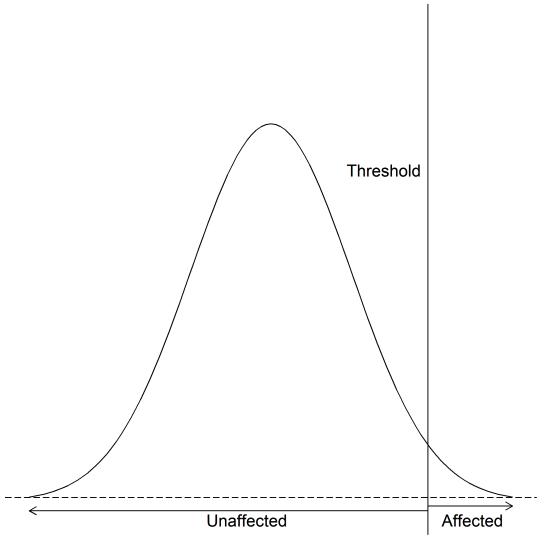


Figure 1.1: The liability threshold model. Only when an individual has a liability above the liability threshold will he/she be affected.

discontinuous 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 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 distribution 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

The key limitation of eq. (1.7) was its inability to discriminate the genetic factors from the shared environmental factors. Such problem arise as family not only shared some of their genes, but they also tends to share some of the environmental factors such as diet. In fact, this was the main reason for researchers to discord the argument that schizophrenia is a genetic disorder.

A classical adoption study carried out by Heston (1966) in 1966 set off to discriminate whether if the increased risk of schizophrenia in relatives of schizophrenia was caused by the shared environmental factors or the shared genetic factors. An advantages of adoption studies was that if the child was separated from their family early after birth, then the shared environmental factors should be minimized, thus any resemblance between the parent and child should be driven mainly by the shared genetic factors. Heston (1966) collected data of 47 individuals born from a schizophrenic mother during the period from 1915 to 1947. They were separated from their mother within three day of birth and were sent to a foster family. 50 matched control were also recruited to the study. It was observed that there was an increased risk of schizophrenia in individual born to schizophrenic mother when compared to the control group even-though they were brought up in a different environment as that of their mother. This result suggested that schizophrenia was likely driven by the shared genetic factors instead of the shared environmental factors.

1.2.5 Twin Studies

Despite the usefulness of adoption studies in delineating the effect of shared environment from the genetic factors, collection of adoption data were difficult. Moreover, any prenatal influence such as alcohol abuse during pregnancy might confound the results. Therefore, an alternative way would be the twin studies using the relationship between the monozygotic (MZ) and dizygotic (DZ) twins.

Theoretically, MZ twins should share all their genetic components (both additive (A) and non-additive (D) genetic factors) and also their common environmental factors (C) where the only difference between a twin pair would be the non-shared environmental factors (E). As for the DZ twins, they also share the same common environmental factors yet they only share $\frac{1}{2}$ of their additive genetic factors and $\frac{1}{4}$ of their non-additive genetic factors. The non-shared environmental was also by definition not shared among the twins (Rijsdijk and Pak C Sham, 2002). Based on these assumptions, Falconer and Mackay, 1996 derived the heritability as

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

where ρ_{MZ} and ρ_{DZ} were 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, strongly suggest schizophrenia as a genetic disorder. 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 that although there was a non-zero contribution of environmental influence on liability of schizophrenia (11%, confidence interval (CI)=$3\% - 19\%$), there was a much larger contribution from genetics (81%,

CI=$73\% - 90\%$), further supporting that schizophrenia was largely mediated by the genetic factors.

Such findings were not limited to twin-studies but were also reported in large scale population based studies. A recent large scale population based study in Sweden population (Lichtenstein et al., 2009) also found that there was a large genetic contribution in schizophrenia (64%). Although the estimated heritability (64% (Lichtenstein et al., 2009) vs 81% (Sullivan, Kendler, and M. C. Neale, 2003)) differs between the two studies, there is no doubt that schizophrenia is highly heritable, leading to the initiative of genetic research in schizophrenia.

1.3 Schizophrenia Genetics

The results from the twin studies strongly support schizophrenia as a genetic disorder. However, little was known about the mechanism of schizophrenia nor the genetic architecture of the disorder. All data from adoption studies, twin studies and family studies shown that schizophrenia does not follow the Mendelian framework (Gottesman and Shields, 1967a; Gottesman and James Shields, 1982). Specifically, shall schizophrenia be a Mendelian disorder, then we would expect all MZ siblings of the proband to also suffer from schizophrenia. However, the life time morbid risk of monozyogitc twins were only 48% (fig. 1.2) (Gottesman, 1991), making it unlikely for schizophrenia to follow a Mendelian pattern.

Based on these observations, Gottesman and Shields (1967b) proposed that schizophrenia follows a polygenic model where disease phenotype were determined by the additive effects from multiple genes. Thus, schizophrenia is likely to be a complex genetic disorder with complicated pattern of inheritance. Their hypothesis was supported by the calculation of Risch (1990a).

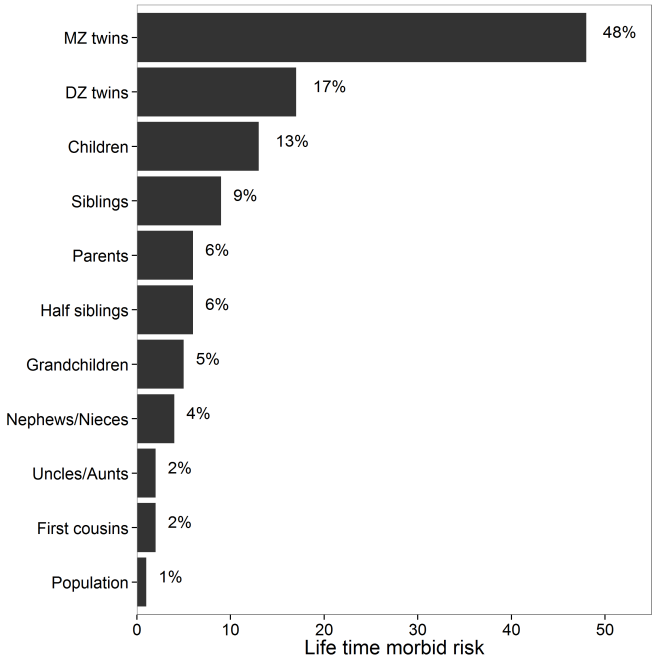


Figure 1.2: Lifetime morbid risks of schizophrenia in various classes of relatives of a proband. It was noted that the morbid risk of monozygotic (MZ) twins were only 48%, much lower than one would expect if schizophrenia follows a Mendelian pattern. Reproduced with permission from journal (Riley and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

Not only does Risch (1990a) supports the polygenic model for schizophrenia, Risch (1990a) also estimated the possible effect size of individual locus in schizophrenia. By comparing the observed life time morbid risk and the expected risk from different models, Risch (1990a) proposed that genetic models with a single locus with risk of 3.0 and with all other loci of small effect or models with two or three loci with risk of 2.0 were most consistent with the observed life time morbid risk of schizophrenia (Risch, 1990b).

Risch (1990a)'s calculation provided an explanation for the early inconsistent findings of linkage studies in schizophrenia (Harrison and Weinberger, 2005). As linkage studies were aimed to identify genetic variation of large effect size they failed to capture genetic loci with small effect size. It was therefore tempting to suggest that schizophrenia only follows the "common disease-common variant" model, which stated that schizophrenia is mediated by large amount of common variants

such as Single Nucleotide Polymorphism, each carries a small effect size.

However, another possible hypothesis was that the variation mediating schizophrenia were rare, therefore require a large sample size to detect and the inconsistent results of early linkage studies might be due to the inadequate sample size. This lead to some researchers suggesting the "common disease-rare variant" hypothesis, which propose that schizophrenia was mediated by a small amount of rare variants, each with a large effect size (McClellan, Susser, and King, 2007).

Nevertheless, success in genetic research of schizophrenia remains limited. Only until the initiation of Human Genome Project and technological advance resulted from that does genetic research of schizophrenia entered an era of success.

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) on the human genome, which is one of the major source of genetic variation.

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 maf ≥ 0.05. More importantly was that the HapMap Project also provided a detailed Linkage Disequilibrium (LD) map of the human genome.

LD was of particular importance to genetic research for it was the non-random correlation of genotypes between 2 genetic loci. SNPs in high LD were usually observed together in the human genome. When a large amount of SNPs were in high LD together, they form what was known as a LD block. By performing

association testing on SNPs representing a LD block ("tagging"), one can avoid the need of performing association on the whole genome, therefore reducing the cost of the experiment. This was the fundamental concept of Genome Wide Association Study (GWAS) which was now extensively used in the genetic research.

1.3.2 Genome Wide Association Study

In GWAS, genome-wide genotyping array were commonly used to systematically detect common genetic variants such as SNP and copy number variation (CNV). For quantitative traits, the association between the trait and frequency of the variants were calculated using methods such as linear regression. On the other hand, for dichotomous traits such as schizophrenia, the frequency of the variants were compared between the case and control samples using methods such as chi-square test or logistic regression. Because of the problem of multiple testing, only variants with a p-value passing a genome wide threshold (p-value $\leq 5 \times 10^{-8}$) were considered significant. Another possible method to decide the significant threshold was 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. When designing a GWAS, one need to take into account of the magnitude of effect, sample size, and required level of statistical significance (the false-positive, or type I, error rate) in order to have a powerful study (S Purcell, Cherny, and P C Sham, 2003).

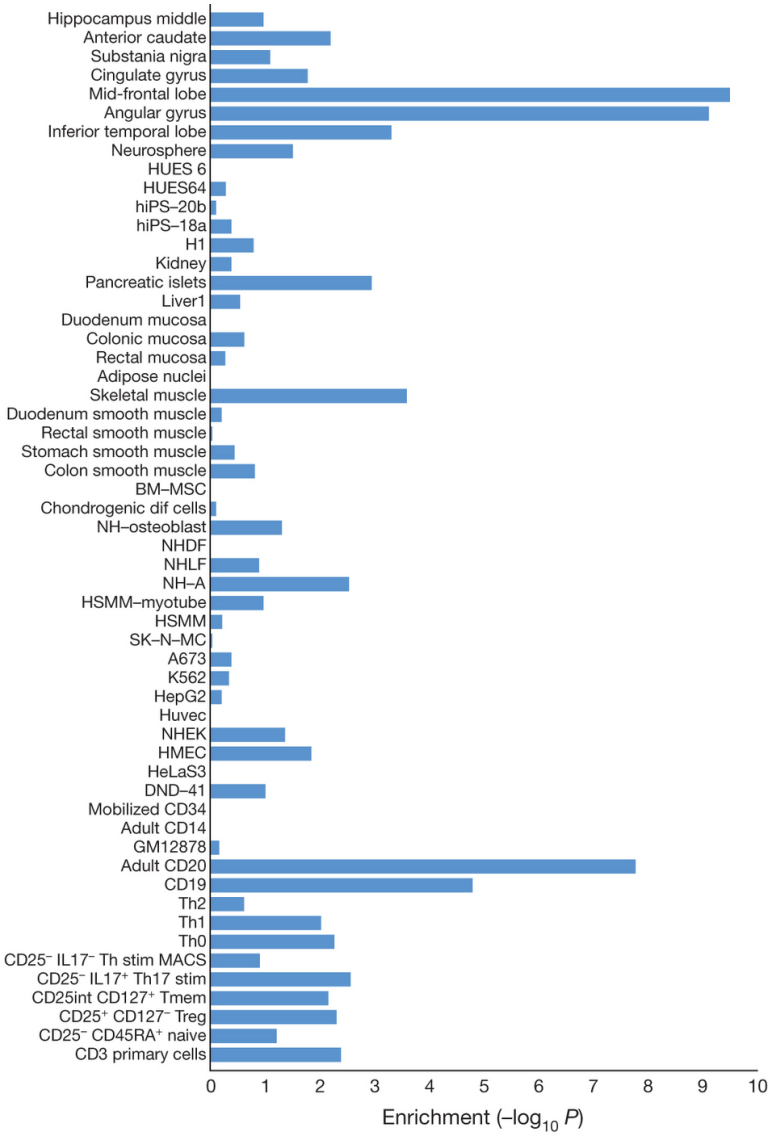
The Success of Psychiatric Genomic Consortium

Despite the great promise from GWAS, early GWAS in schizophrenia remain largely disappointing and were unable to identify any robust genetic markers associated with schizophrenia. The failure of early GWAS in schizophrenia were mainly due to the relative small sample size of the studies, which result in low detection power.

To overcome the problem of small sample size, large consortium were formed such that data from different research groups from different countries were combined, which provides a large sample size for the analysis. By 2014, the Schizophrenia Working group of the Psychiatric Genomics Consortium (PGC) has collected a total of 36,989 schizophrenia samples and 113,075 controls for the meta-analysis of schizophrenia (Stephan Ripke, B. M. Neale, et al., 2014). In their study (Stephan Ripke, B. M. Neale, et al., 2014), 128 linkage-disequilibrium-independent SNPs were found to exceeded the genome-wide significance (p-value $\leq 5 \times 10^{-8}$), corresponding to 108 genetic loci. 75% of these loci contain protein coding genes and a further 8% of these loci were within 20kilobase (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. Importantly, *DRD2*, the target of all effective anti-psychotic drug were also associated with schizophrenia. This result converges with existing knowledge of *DRD2* being involved in the pathology of schizophrenia, supported by multiple lines of research (Talkowski et al., 2007). It was further demonstrated that 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, provided further genetic support of the involvement of the immune system in the etiology of schizophrenia. Because of its role in neural development (B. Zhao and Schwartz, 1998; Deverman and Patterson, 2009), it is likely that the perturbation in the immune system might disrupt the brain development, therefore increasing the risk of schizophrenia.

Figure 1.3: Enrichment of enhancers of SNPs associated with schizophrenia. It was observed that the largest enrichment were in cell lines related to the brain and in tissues with important immune functions. Graphs reproduced with permission from the journal (Stephan Ripke, B. M. Neale, et al., 2014).



Although the PGC schizophrenia GWAS is very successful, 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 given 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

In a typical GWAS, a stringent genome wide significant threshold were usually employed to avoid false positive findings. However, if individual SNPs have a small effect on the trait, the real association might be missed. Therefore, to estimate the true contribution of common SNPs to a disease (SNP-heritability), one should try to use all SNPs in the estimation.

Genome-wide Complex Trait Analysis

Currently, the most popular algorithm used for the estimation of SNP-heritability is Genome-wide Complex Trait Analysis (GCTA), which uses information from the Genetic Relationship Matrix (GRM) (J Yang et al., 2011). The GRM is 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. This is based on the fact that

genotypes were usually coded as 0, 1 or 2 (homozygous reference, heterozygous and homozygous alternative respectively) and should follow the binomial distribution. From the binomial distribution, the expected mean and variance of the genotype i will be $2p_i$ and $2p_i(1-p_i)$ respectively. Thus $A_{jk} = \frac{1}{N}\sum_{i=1}^{N} z_{ij}z_{ik}$ 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) then fit the effects of all the SNPs as random effects by a mixed linear model (MLM)

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

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

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

The main concept of GCTA is that instead of testing the associations for individual SNPs, one fit the effects of all SNPs as random effects in a MLM and estimate a single parameter, 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 estimates the σ_g^2 and $\sigma_\epsilon^2$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 $\frac{\sigma_g^2}{\sigma_g^2+\sigma_e^2}$.

Based on the above concept, 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%. The main difference in the estimates was because the MLM REML were able to consider all SNPs simultaneously without limited on significant SNPs. Although the estimates was still less

than 80% which was the expected heritability of height, Jian Yang, Benyamin, et al. (2010) was able to demonstrated that one possible source of "missing heritability" might be due to incomplete LD. By taking into consideration of incomplete LD, it was estimated that the proportion of variance explained by causal variants can be as high as 0.84 with standard error (SE) of 0.16 (Jian Yang, Benyamin, et al., 2010), close to the expected heritability. Together, J Yang et al. (2011) provide a possible method for the estimation of the variance explained by SNPs in GWAS data and the method is now implemented in GCTA which is wildly adopted.

The problem with GCTA was that genotype data are required to calculate the GRM. For complex disease like schizophrenia, the data were usually obtained from multiple data source where the raw genotypes were unavailable due to privacy concerns. Instead, summary statistics were usually provided. Therefore estimation of variance explained by SNPs in these GWAS can only rely on the summary statistics.

LD SCore regression

In large scale GWAS studies, a general inflation of summary statistics can sometimes be observed. It was usually considered to be contributed by the presence of confounding factors such as population stratification, under the assumption that most of the SNPs should have no association to the disease. It was therefore a common practice for one to perform the Genomic Control (GC) on the GWAS results (Zheng, Freidlin, and Gastwirth, 2006).

The problem of GC was that the basic assumption of a small number of causal SNPs might not be true, especially in complex disease like 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 (Jian Yang,

Weedon, et al., 2011). More importantly, they 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.

The observation of Jian Yang, Weedon, et al. (2011) provide important foundation for the estimation of SNP heritability based on summary statistics where a possible method will be to elucidate the heritability based on the magnitude of inflation of the summary statistics. However, when confounding factors such as population stratification and cryptic relatedness are presented, they can also inflate the summary statistics. Therefore, in order to estimate the SNP-heritability, one must be able to delineate the confounding factors from the polygenicity of the trait.

Based on the work of Jian Yang, Weedon, et al. (2011), B. K. Bulik-Sullivan et al. (2015) hypothesized that strength of "tagging" of a SNP should be correlated with the probability of it to "tag" the causal SNP yet should be independent to confounding factors such as population stratification and cryptic relatedness. B. K. Bulik-Sullivan et al. (2015) then define the strength of "tagging" of a SNP as the LD score, which is the sum of r^2 of k SNPs within a 1cM window of SNP_j:

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

Based on their hypothesis, the expected χ^2 of association of SNP_j with the trait can 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 were 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 and the intercept minus one will represent the mean contribution of the confounding bias such as those of population stratification. The LD SCore regression (LDSC) was implemented by B. K. Bulik-Sullivan et al. (2015), hoping to use eq. (1.16) to delineate the contribution from confounding factors and common genetic variants.

To test their hypothesis, B. K. Bulik-Sullivan et al. (2015) simulated multiple GWAS where the trait can have a polygenic architecture or where confounding factors can present. When the simulated trait is polygenic and no confounding factors were 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.

Moreover, when a polygenic trait was 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.

Given the success of the simulation, B. K. Bulik-Sullivan et al. (2015) estimated the SNP heritability of schizophrenia using the summary statistics from the PGC schizophrenia GWAS (Stephan Ripke, B. M. Neale, et al., 2014). By applying the liability threshold adjustment, B. K. Bulik-Sullivan et al. (2015) estimated the SNP-heritability of schizophrenia should be 0.555 with SE of 0.008. 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)) suggesting that it is possible for variants other than common SNPs to account for variations in schizophrenia.

Partitioning of Heritability

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

Traditionally, functional enrichment analysis in GWAS only take into account of SNPs that passed the genome wide significance threshold. However, for complex traits such as that of schizophrenia, much of the heritability might lies in SNPs that 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, if one only consider the significant loci, risk loci that have not reach genome wide significance threshold might be ignored from the analysis, decreasing the power of the functional enrichment analysis.

In order to estimate whether if a functional categories was associated with the trait, LDSC takes into consideration of the summary statistic of all the SNPs. 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) tried to identify functional categories that were most enriched in schizophrenia. In their study, it was found that brain cell types were most enriched in schizophrenia, especially those related to the central nervous system (CNS). Of all the functional categories, the most enriched category in schizophrenia was the H3K4me3 mark in the fetal brain(table 1.2). As H3K4me3 was mostly linked to active promoters, this suggest that genes that were activated in fetal brain (e.g. genes related to brain development) were associated with schizophrenia, supporting the idea of schizophrenia as a neuro-developmental disorder.

Moreover, it was also observed that the second most enriched cell types were those related to immunity. Undoubtedly, the CNS and the immune system have an important role in the disease etiology of schizophrenia.

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 estimated SNP-heritability using the common variants captured by the PGC schizophrenia GWAS suggest that variants other than common SNPs were accounting for the variation in schizophrenia. Based on the "common disease-rare variant" hypothesis, another interesting direction of schizophrenia research will be to identify rare variants associated with schizophrenia.

Copy Number Variation

A possible source of rare variants can be copy number variations (CNVs). CNV were classified as segment of DNA that is 1kb or larger and that 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 used the Swedish national sample (4,719 schizophrenia samples and 5,917 controls). In their study, they were able to association between schizophrenia and CNV such as 16p11.2 duplications, 22q11.2 deletions, 3q29 deletions and 17q12 duplications were identified. Through the gene set association

analysis, calcium channel signaling and binding partners of the fragile X mental retardation protein were found to be associated with these CNV (Szatkiewicz et al., 2014). Interestingly, the calcium channel signaling were also enriched in the PGC GWAS on SNP association, suggesting that the variants were converging on similar set of pathway or gene sets.

Similarly, Walsh et al. (2008) also found that genes disrupted by structure variants in their cases 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 (Walsh et al., 2008).

An important observation in these CNV studies was that the CNV were generally rare ($\leq$ 12 in 4,719 samples (Szatkiewicz et al., 2014)) and has a relative large effect (e.g. odd ratio $>$ 2 (Szatkiewicz et al., 2014; Walsh et al., 2008)), following the "common disease-rare variant" model.

Rare Single Nucleotide Mutation

Unlike CNV which affects a large region, it is difficult to capture rare SNP using current genotyping chips. Therefore, large scale association of rare SNPs was unavailable until the development of the next generation sequencing (NGS) technology. The NGS generates high-throughput sequencing data with per base resolution, allow one to investigate the whole human genome or the human exome without relying on "tagging".

Using exome sequencing, S M Purcell et al. (2014) sequenced the exome of 2,536 schizophrenia cases and 2,543 normal controls. They were able to identify a common missense allele in *CCHCR1* in the MHC that were associated with

schizophrenia. Although none of the genes showed a significant burden of rare mutation in cases, a significant increased burden of rare nonsense and disruptive variants was observed in cases in gene sets likely to be associated with schizophrenia such as voltage-gated calcium ion channel, genes affected by *de novo* mutations in schizophrenia (Fromer et al., 2014) and the postsynaptic density.

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

On top of rare variants, another possible source of "missing" heritability can comes from interaction between the genetic and environmental risk factors. Although previous studies (Gottesman and Shields, 1967a) suggested that the non-additive genetic factors were unlikely to contribute to schizophrenia, the possibility of 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

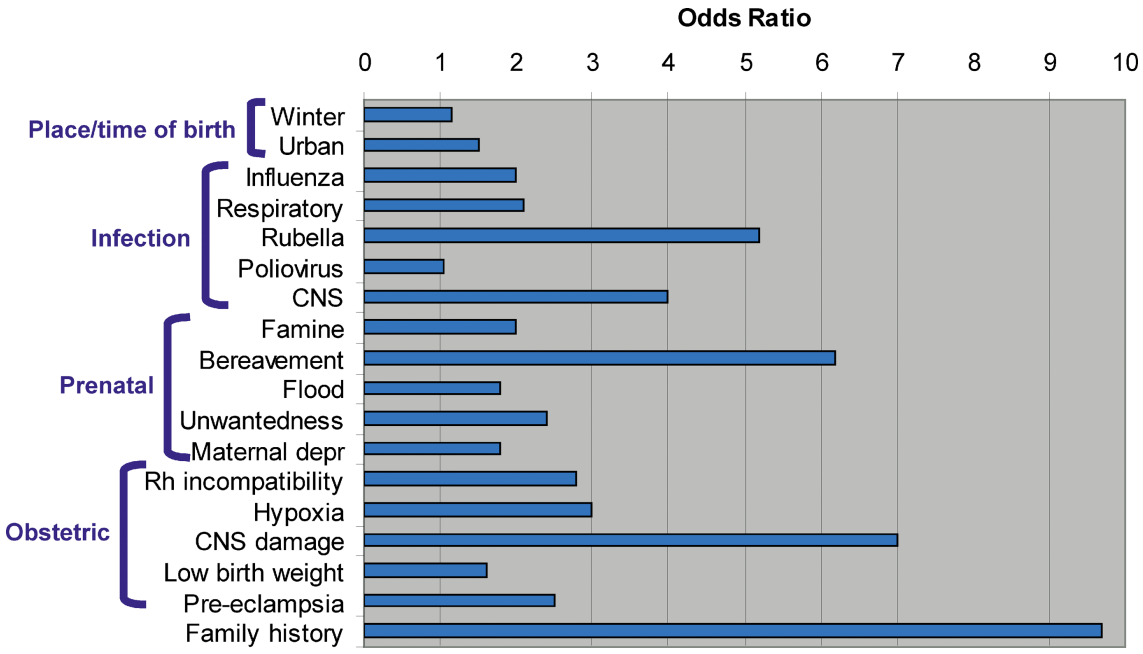


Figure 1.4: Risk factors of schizophrenia. It was observed that family history of schizophrenia was the largest risk factors. Risk of schizophrenia can be more than 9 times higher than the general population for individual with a family history of schizophrenia

status (McGrath et al., 2008). They 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 and existing evidence suggest that there are indeed an interaction between prenatal infection and genetic variations (Clarke et al., 2009). It is therefore interesting to investigate how prenatal infection trigger schizophrenia and how it interacts with genetic variations in the development of schizophrenia.

1.4.1 Prenatal Infection

Prenatal infection has always been an important risk factor of schizophrenia, being the single largest non-genetic risk factor of schizophrenia (fig. 1.4)(Sullivan, 2005). Initial clues indicated that births during the winter and spring months and 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 disease varies by season and infectious disease can spread more quickly in urban regions due to higher population density, these evidence suggest 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. This leads to difficulties in replication of the findings. Subsequently, researchers uses birth cohorts where infection was documented using different biomarkers during pregnancies to provide a better labeling of the exposure status (A S Brown and Derkits, 2010). Through these rigorous studies it was found that the risk of schizophrenia increases as long as an individual's mother was infected by different form of infectious agents such as influenza, HSV-2 and *T.gondii* during gestation (A S Brown and Derkits, 2010). As different infectious agents all increase the risk of schizophrenia, it leads to the hypothesis of maternal immune activation (MIA) (A S Brown and Derkits, 2010) where it was suggested that instead of a particular infectious agents, it was the maternal immune response that disrupt the brain development in the offspring, thus leading to an elevated risk of schizophrenia.

To really understand how MIA increase the risk of schizophrenia, it is important to understand the molecular mechanism. A great challenge in the study of MIA was that one cannot carry out empirical experiment in human samples due to ethical concerns. Thus a popular alternative is to employ rodent models. However, unlike physiological traits, psychiatric disorder such as that of schizophrenia often contain symptoms related to higher level functioning such as hallucinations,

delusion, disorganized speech etc (American Psychiatric Association, 2013) that are not readily detectable in rodents. This raises challenge in diagnosing whether if the rodent has demonstrated the symptoms of schizophrenia for not only it was difficult to check whether if the high level functioning of the rodent is disrupted, there were no available biomarkers for schizophrenia. Therefore instead of labeling whether if the rodent is "schizophrenic" or "normal", one would rather consider whether if the rodent demonstrate any "schizophrenia-like" behaviours such as impaired prepulse inhibition, impaired working memory and reduced social interaction (U Meyer, Yee, and J Feldon, 2007). An important point to note here is that as autism and schizophrenia shares most of these behavioral abnormality, and that risk of autism is also increased by MIA (Alan S Brown, 2012), studies using these rodent models were usually non-specific to schizophrenia or autism. Rather, autism and schizophrenia were usually considered together in these models. However, the discussion of the etiology of autism and the similarity and difference between autism and is beyond the scope of the current thesis. Therefore, for the simplicity and focus of the current thesis, we would 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 in rodents. It was found that offspring exposed to PolyI:C displays phenotypes mirrors that 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. What was most interesting was 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, suggesting that IL-6 is central to the process by which MIA causes long-term behavioral changes.

Further 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-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) hypothesize 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

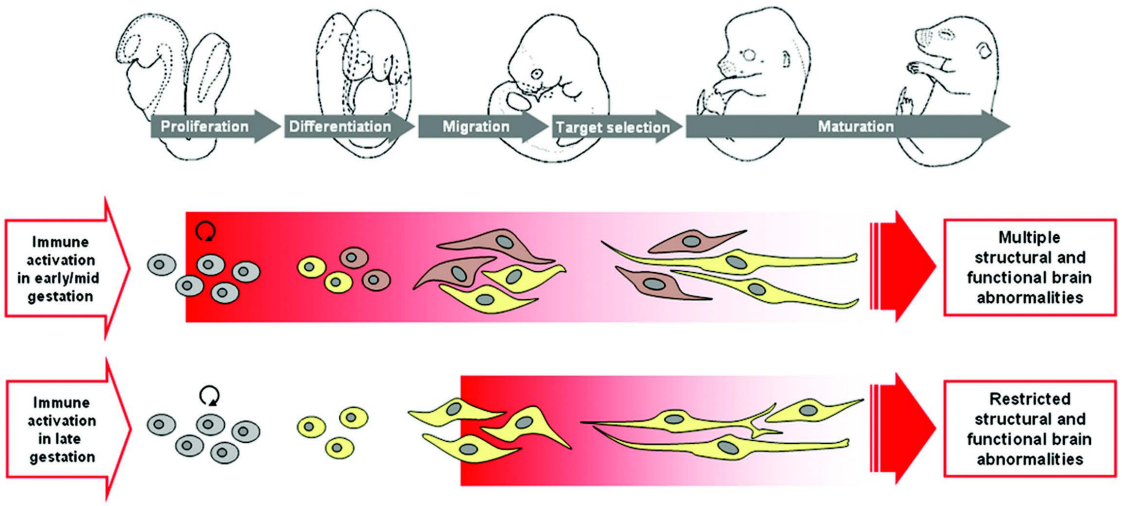


Figure 1.5: Hypothesized model of the impact of prenatal immune challenge on fetal brain development. Maternal infection in early/mid pregnancy may affect early neurodevelopmental events in the fetal brain, thereby influencing the differentiation of neural precursor cells (grey) into particular neuronal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

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. 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 to whether if they were exposed to postnatal stress whereas 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. It was even estimated that one third of all schizophrenia cases could have been prevented shall all infection were prevented from the entire pregnant population (A S Brown and Derkits, 2010).

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 have been the vulnerability period. However, in the birth cohorts such as the Prenatal Determinants of SCZ, it was found that the time window with maximal risk for infection-mediated disturbance in brain development is earlier than the second trimester of human pregnancy and can be as early as the first trimester (U Meyer, Yee, and J Feldon, 2007). Through the review of existing MIA studies on rodent models, U Meyer, Yee, and J Feldon (2007) suggests that effect of MIA during late pregnancy can be 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 might 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 and the understanding of the full molecular implication of early MIA events in adult brain were lacking. As technology advances, we can now employ the RNA Sequencing technique to examine the global 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, one can only inspect the global expression changes using the microarray using probe hybridization. As NGS developed, one can now use poly-T probes to "extract" the mRNA fragments and sequence them. The depth of coverage of each gene then provide a general representation of the concentration fo the mRNA in the cell. When compared to microarray, the RNA Sequencing has a number of advantages, most notably, because RNA Sequencing does not rely on specific probe hybridization, it doesn't suffer from bias introduced by probe performances such as signal saturation, cross-hybridization, background noises and non-specific hybridization (S. Zhao et al., 2014). Moreover, RNA Sequencing has the additional advantage that one can perform not only the differential expression analysis, but also detect alternative splicing events and de novo transcripts.

However, the analysis of RNA Sequencing is more complicated when compared to microarray. The first hurdle in the analysis of RNA Sequencing data is the sequence alignment. RNA sequencing will typically generate sequence reads from the mRNA transcripts and only need to align these reads to either the genome or the transcriptome in order to be able to calculate the depth of coverage for each genes, thus allowing the differential expression analysis. The different alignment strategies have their own pros and cons.

Alignment to transcriptomes were most straightforward as the reads were originated from the transcripts and should have sequence composition similar to the transcriptome. The problem of transcriptome alignment is that multiple isoform can share the same exon, leading to read mapping uncertainties (B. Li and Dewey, 2011). Without taking into consideration of the uncertainties, the downstream analysis might be biased and inaccurate. When one is only interested in analyzing the gene level expression difference, this complication might be unnecessary.

On the other hand, alignment to the genome should help to reduce the problem of multiple mapping yet it will require a splice aware aligner such as TopHat2 (Kim et al., 2013), STAR (Dobin et al., 2013) and MapSplice (K. Wang et al., 2010). The reason behind was that as the reads were originated from the mRNA where alternative splicing might have occurred, the reads might span multiple exons which were separated by intronic regions. The splicing algorithm will be able to "split" the reads and correctly align them onto the exons. with the accurate alignment, one can then quantify the "expression" of each individual genes.

The expression of a gene is usually represented in terms of number of reads aligned to the gene. Given this information, statistic analysis can then be performed on the count data. 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 were more complicated. Early RNA Sequencing experiment assumes the gene expression counts follows the Poisson distribution (Marioni et al., 2008) where the variance is assumed to be equal to the mean of the expression. However, it was found that the assumption of Poisson distribution is too restrictive where an over-dispersion was typically observed in RNA Sequencing data (S Anders and W Huber, 2010). Therefore, to overcome the problem of over-dispersion, modern RNA Sequencing statistical package usually models the RNA Sequencing counts 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.

Nonetheless, as our knowledge with RNA Sequencing advances, we are getting better in utilizing the information provided by RNA Sequencing and it should serves as an important tool for the analysis of gene expression changes induced by MIA event.

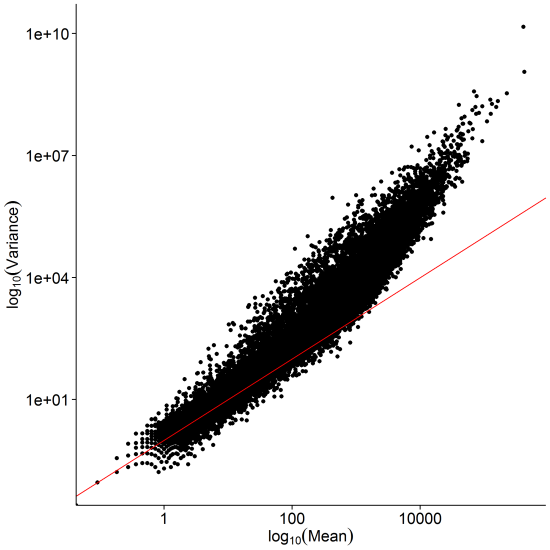


Figure 1.6: Over-dispersion observed in RNA Sequencing Count Data. If the RNA Sequencing count data follows the Poisson distribution, then the mean and variance of the data should be equal (follow the diagonal). However, it was observed that as the mean increases, the variance increases even more, suggesting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

1.5 Summary

In this thesis, we would like to first perform a series of empirical simulations to 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 suggested by B. K. Bulik-Sullivan et al. (2015), under certain conditions such as when the trait is oligogenic, the performance of LDSC might be subpar. Thus we would also like to develop an alternative algorithm for the estimation of SNP heritability that is robust to different genetic architecture. Ultimately, we would like to repeat the analysis by B. K. Bulik-Sullivan et al. (2015) to estimate the true contribution of common SNPs to the variance in schizophrenia.

Currently, there are evidences suggesting that there might be interaction between prenatal infection and genetic variations in the development of schizophrenia (Tienari et al., 2004; Clarke et al., 2009). We therefore hypothesize that the differential gene expression induced by MIA and genetic mutation might have act

upon the same functional pathway. To test this hypothesis, we performed a hypothesis generation RNA Sequencing study to capture gene expression changes induced by early MIA events (Gestation Day (GD)9) in the cerebellum of mouse using the PolyI:C mouse model. Based on the gene expression changes, we hope to identify pathways perturbed by early MIA events. Most importantly, we would like to test whether if these pathways contribute disproportionately to the heritability of schizophrenia. As a result of that, we would also perform the partitioning of heritability using LDSC on the pathways affected by MIA.

Moreover, recent study from our lab suggested that n-3 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 would also like to take this opportunity to assess the effect of n-3 PUFA rich diet on the gene expression pattern in the brain of the adult offspring.

This thesis will be divided into three parts. First, in Chapter 2, we performed a series of empirical simulations to assess the performance of LDSC in the estimation of SNP heritability. We also proposed an alternative approach for the estimation of SNP-heritability from GWAS summary statistics that is robust to different genetic architectures.

In Chapter 3, a hypothesis generation study was performed to study the effect of MIA on the gene expression pattern of mouse cerebellum. On top of that, as recent study suggested that n-3 PUFA rich diet can help to reduce the schizophrenia-like behaviour observed mouse exposed to early MIA (Q. Li, Leung, et al., 2015), we also investigated the effect of n-3 PUFA rich diet on the gene expression pattern of mouse cerebellum.

Lastly, we summarize and conclude all findings in Chapter 4 and give future perspectives on the research.

2 Heritability Estimation

2.1 Introduction

The development of LDSC (B. K. Bulik-Sullivan et al., 2015) has allow researchers to estimate the true contribution of common SNPs to the variance in different diseases. Its ability in delineating the contributions from confounding factors such as population stratification and common SNPs were vital to its success. However, limited simulations were performed by B. K. Bulik-Sullivan et al. (2015) and it is unclear how different sampling strategies (e.g. extreme phenotype sampling) or genetic architectures (e.g. different population prevalence) affect the performance of LDSC.

Moreover, as schizophrenia is usually defined as "affected" or "normal", the estimation of SNP heritability might have to adjusted for the ascertainment bias introduced by the case control sampling. The ascertainment bias correction was usually performed based on the liability threshold model. However, this adjustment might not be as straightforward as it seems. For example, it has been suggested that GCTA, the most popular SNP heritability estimation tools for GWAS, will provide highly biased estimates for case control studies (Golan, Eric S Lander, and Rosset, 2014). As B. K. Bulik-Sullivan et al. (2015) did not perform any empirical simulation of the effect of the case control sampling to their estimates, it is therefore important for us to perform empirical simulations to investigate whether if the case

control sampling has any impact to the estimates of LDSC.

Finally, as noted by B. K. Bulik-Sullivan et al. (2015), the performance of LDSC can be subpar under certain condition, for example when the trait is oligogenic. We are therefore interested to see if there is a robust algorithm for the estimation of SNP heritability without being affected by the genetic architecture of the trait.

In this chapter, we will first introduce SNP HeRitability Estimation Kit (SHREK), an alternative algorithm to LDSC for the robust estimation of SNP heritability based on GWAS summary statistics. We then perform a series of empirical simulation to test the performance of LDSC and SHREK when the trait has different genetic architectures. Most importantly, we would like to repeat the analysis of B. K. Bulik-Sullivan et al. (2015) to estimate the true contribution of SNP to schizophrenia using the PGC schizophrenia GWAS summary statistics. This would provide us insights into possible direction of future researches in schizophrenia.

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

The overall aims of this study is to develop a robust algorithm for the estimation of the narrow sense heritability using only the summary statistic from a GWAS and to study the performance of the heritability estimation algorithms for different type of traits.

It was noted that in GWAS, the test statistic of a particular SNP should increases with its own effect size and the effect size from all the other SNPs in LD with it. Based on this property, we may use the information from the LD matrix and the test statistic of the GWAS SNP the estimate the narrow sense heritability.

2.2.1 Heritability Estimation

Remember that the narrow-sense heritability is defined as

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

where $\mathrm{Var}(\boldsymbol{x})$ is the variance of the genotype and $\mathrm{Var}(\boldsymbol{y})$ is the variance of the phenotype. In a GWAS, regression were performed between the SNPs and the phenotypes, giving

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

where $\boldsymbol{y}$ and $\boldsymbol{x}$ are the standardized phenotype and genotype respectively. $\boldsymbol{\epsilon}$ is then the error term, accounting for the non-genetic elements contributing to the phenotype (e.g. environment factors). Based on eq. (2.1), and by assuming that $\beta \boldsymbol{x}$

independent of $\boldsymbol{\epsilon}$, one can then have

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

β^2 is then considered as the portion of phenotype variance explained by the variance of genotype, which can also be considered as the narrow-sense heritability of the phenotype.

A challenge in calculating the heritability from GWAS data is that usually only the summary statistic or p-value were provided and one will not be able to directly calculate the heritability based on eq. (2.2). In order to estimate the heritability of a trait from the GWAS summary statistic, we exploit the fact that when both $\boldsymbol{x}$ and $\boldsymbol{y}$ are standardized, β^2 will be equal to the coefficient of determination (r^2). Thus, based on properties of the Pearson product-moment correlation coefficient:

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

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

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

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

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

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

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

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

When there are LD between each individual SNPs, the situation will become more complicated as each SNPs' observed effect will be influenced by other SNPs in LD with it:

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

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

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

Where $f(m, \boldsymbol{V})$ denote a general distribution with mean m and variance $\boldsymbol{V}$ and $\boldsymbol{R}$ is the LD matrix between SNPs.

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

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

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

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

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

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

where $\boldsymbol{H}$ is the variance of the "true" effect. It is shown that heritability can be expressed as

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

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

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

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

$$\rho_i = \sum_j R_{ij}\boldsymbol{\beta}_j \qquad (2.12)$$

In reality, the *observed* correlation usually contains error. 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

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

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

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

If we define

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

we got

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

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

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

And the heritability is then defined as

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

where the $\mathbf{1}^t$ were multiplied to $\boldsymbol{R_{sq}^{-1}} \boldsymbol{f}$ 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

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

Therefore, to obtain the variance of $\hat{Heritability}$, we first need to calculate the variance covariance matrix of $\boldsymbol{f}$.

We first consider the standardized genotype x_i with standard normal mean

z_i and non-centrality parameter μ_i, we have

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

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

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

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

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

Remember that $\mathrm{E}[z_i z_j] = R_{ij}$, we then have

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

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 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 $\boldsymbol{z} = \sqrt{\boldsymbol{\chi^2}}$ from eq. (2.14), with the direction of effect as its sign and $\circ$ is the element-wise product (Hadamard product).

The problem with eq. (2.20) is that it requires the direction of effect. Without the direction of effect, the estimation of SE will be inaccurate. If we consider that $n \times \boldsymbol{f} + 1$ 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

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

eq. (2.21) should in theory gives us an heuristic estimation of the SE. Moreover,

the direction of effect was not required for eq. (2.21), reducing the number of input required from the user.

2.2.3 Case Control Studies

When dealing with case control data, we cannot directly use eq. (2.17) to estimate the heritability. Instead, we will need to employ the concept of liability threshold model from section 1.2.3.

Based on the derivation of Jian Yang, Naomi R. Wray, and Peter M. Visscher (2010), the approximate ratio between the non-centrality parameter (NCP) obtained from case control studies (NPC_{CC}) and quantitative trait studies(NCP_{QT}) were

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

where

K = Population Prevalence

v = Proportion of Cases

N = Total Number of Samples

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

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

Using this approximation, we can directly transform the NCP between the case control studies and quantitative trait studies. As we are not interested in transforming the NCP between two different studies, the sample size of the case control study (N_{CC}) and sample size of the quantitative trait study (N_{QT}) will be

the same in eq. (2.22), therefore eq. (2.22) becomes

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

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

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

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

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

2.2.4 Extreme Phenotype Sampling

The development of GWAS now provide unprecedented power to perform hypothesis free association throughout the whole genome. However, a challenge for the studies of complex traits is to obtain sufficient sample size with a limited budgets. It is therefore important to design the experiment in a way where sample size can be reduced without affecting the power of the study. A common technique is to perform extreme phenotype sampling in the detection stage of the study. The extreme phenotype sampling will inflate the frequency distortion between samples from the two extreme end of phenotype, thus increase the statistical power (Guey et al., 2011). It was estimated that for a 0.5% variant with a fivefold effect in the general population, a discovery studies using extreme phenotype sampling requires four times less samples in the replication to achieve 80% power when compared to studies using random samples (Guey et al., 2011). This allows studies to be conducted using a smaller amount of samples with the same degree of power, therefore reducing the cost of the study.

A problem of extreme phenotype sampling was that the variance of the selected phenotype will not be representative of that in the population. The effect size are generally overestimated (Guey et al., 2011). Thus, to adjust for this bias, one can multiple the effect size by the ratio between the variance before V_P and after V_P' the selection process (Pak C Sham and Shaun M Purcell, 2014), which is equivalent to the multiplication of $\frac{V_P'}{V_P}$ to f in 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

In order to obtain the heritability estimation, we will require to solve eq. (2.17). If $\boldsymbol{R_{sq}}$ is of full rank and positive definite, it will be straight-forward to solve the matrix equation. However, more often than not, the LD matrix are rank-deficient and suffer from multicollinearity, making it ill-conditioned, therefore highly sensitive to changes or errors in the input. To be exact, we can view eq. (2.17) as calculating the sum of $\hat{\boldsymbol{h^2}}$ from eq. (2.16). This will involve solving for

$$\hat{\boldsymbol{h^2}} = \boldsymbol{R}_{\boldsymbol{sq}}^{-1}\boldsymbol{f} \tag{2.27}$$

which requires the inverse of $\boldsymbol{R_{sq}}$.

In normal circumstances (e.g. when $\boldsymbol{R_{sq}}$ is full rank and positive semi-definite), one can easily solve eq. (2.27) using the QR decomposition or LU decomposition. However, when $\boldsymbol{R_{sq}}$ is ill-conditioned, the traditional decomposition method will fail. Even if the decomposition can be performed, the result tends to be a meaningless approximation to the true $\hat{\boldsymbol{h^2}}$.

Therefore, to obtain an unique solution, regularization techniques such as the Tikhonov Regularization (also known as Ridge Regression) and Truncated

Singular Value Decomposition (tSVD) has to be performed(Neumaier, 1998). There are a large variety of regularization techniques, yet the discussion of which is beyond the scope of this study. In this study, we will focus on the use of tSVD in the regularization of the LD matrix. This is because the Singular Value Decomposition (SVD) routine has been implemented in the EIGEN C++ library (Guennebaud and Jacob, 2010), allowing us to implement the tSVD method without much concern with regard to the detail of the algorithm.

To understand the problem of the ill-conditioned matrix and regularization method, we consider the matrix equation $\boldsymbol{Ax} = \boldsymbol{B}$ where $\boldsymbol{A}$ is ill-conditioned or singular with $n \times n$ dimension. The SVD of $\boldsymbol{A}$ can be expressed as

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

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

$$\boldsymbol{A}^{-1} = \boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^t \tag{2.29}$$

Where $\boldsymbol{\Sigma}^{-1} = \text{diag}(\frac{1}{\sigma_1}, \frac{1}{\sigma_2}, \ldots, \frac{1}{\sigma_n})$. we can then represent $\boldsymbol{Ax} = \boldsymbol{B}$ as

$$\begin{aligned} \boldsymbol{Ax} &= \boldsymbol{B} \\ \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^t\boldsymbol{x} &= \boldsymbol{B} \\ \boldsymbol{x} &= \boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^t\boldsymbol{B} \end{aligned} \tag{2.30}$$

A matrix $\boldsymbol{A}$ is considered as ill-condition when its condition number $\kappa(\boldsymbol{A})$ is large or singular when its condition number is infinite where the condition number can be calculated as $\kappa(\boldsymbol{A}) = \frac{\sigma_1}{\sigma_n}$. Therefore it can be observed that when σ_n is tiny, $\boldsymbol{A}$ is likely to be ill-conditioned and when $\sigma_n = 0$, $\boldsymbol{A}$ will be singular.

One can also observe from eq. (2.30) that when the singular value σ_i is

small, the error ϵ_i in $\boldsymbol{B}_i$ will be drastically magnified by a factor of $\frac{1}{\sigma_i}$. Making the system of equation highly sensitive to errors in the input.

To obtain a meaningful solution from this ill-conditioned/singular matrix $\boldsymbol{A}$, we may perform the tSVD method to obtain a pseudo inverse of $\boldsymbol{A}$. Similar to eq. (2.28), the tSVD of $\boldsymbol{A}$ can be represented as

$$\boldsymbol{A}^{+} = \boldsymbol{U}\boldsymbol{\Sigma}_k\boldsymbol{V}^t \qquad \text{and} \qquad \boldsymbol{\Sigma}_k = \text{diag}(\sigma_1, \ldots, \sigma_k, 0, \ldots, 0) \tag{2.31}$$

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

$$\sigma_i = \begin{cases} \sigma_i & \text{for} \quad \sigma_i \geq t \\ 0 & \text{for} \quad \sigma_i < t \end{cases} \tag{2.32}$$

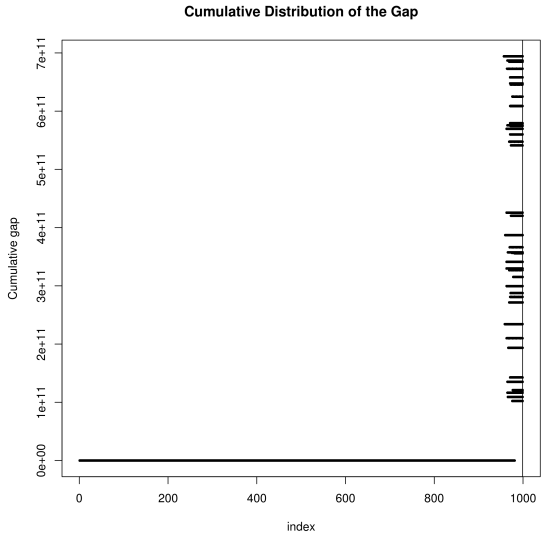
where t is the tolerance threshold. Any singular value σ_i less than the threshold will be replaced by 0 during the inversion.

By selecting an appropriate t, tSVD can effectively regularize the ill-conditioned matrix and help to find a reasonable approximation to x. A problem with tSVD however is that it only work when matrix $\boldsymbol{A}$ has a well determined numeric rank (Hansen, 1987). That is, tSVD work best when there is a large gap between σ_k and σ_{k+1}. If a matrix has ill-conditioned rank, then $\sigma_k - \sigma_{k+1}$ will be small. For any threshold t, a small error can change whether if σ_{k+1} and subsequent singular values should be truncated, leading to unstable results.

According to Hansen (1987), matrix where its rank has meaning will have well defined rank. The LD matrix is the correlation matrix between each individual SNPs, thus the rank of the LD matrix is the maximum number of linear independent SNPs in the region. Because the rank has a meaning, the LD matrix is likely to have well-defined rank.

The easiest way to test whether if the threshold t and if the matrix $\boldsymbol{A}$ has

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



well-defined rank is to calculate the "gap" in the singular value:

$$gap = \sigma_k / \sigma_{k+1} \tag{2.33}$$

a large gap usually indicate a well-defined gap. In this study, we adopt the threshold as defined in MATLAB, NumPy and GNU Octave: $t = \epsilon \times \max(m, n) \times \max(\mathbf{\Sigma})$ where ϵ is the machine epsilon (the smallest number a machine can define as non-zero). And we perfomed a simulation study to investigate the performance of tSVD under the selected threshold. Ideally, if the "gap" is large under the selected threshold, then tSVD will provide a good regularization to the equation.

1,000 samples were randomly simulated from the HapMap (Altshuler et al., 2010) Northern Europeans from Utah (CEU) population with 1,000 SNPs randomly select from chromosome 22. The LD matrix and its corresponding singular value were calculated. The whole process were repeated 50 times and the cumulative distribution of the "gap" of singular values were plotted (fig. 2.1). It is clearly show that the LD matrix has a well-defined rank with a mean maximum "gap" of 466,198,939,298. Therefore the choice of tSVD for the regularization is appropriate.

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

2.2.6 Implementation

Our algorithm was implemented using C++ programming languages (version C++11) and the matrix algebra was performed using the EIGEN C++ header library (Guennebaud and Jacob, 2010). In spite of the fact that the Armadillo library (Sanderson, 2010) is much faster in the calculation of SVD when compared to EIGEN (Ho, 2011), it is dependent on additional libraries such as OpenBLAS. The use of EIGEN therefore simplify the programme installation, making it more user friendly.

Although tSVD allow one to approximate the ill-posed eq. (2.16), it is an $O(n^3)$ algorithm, making the computation run time prohibitive when the number of SNPs is large. Unfortunately, the number of SNPs in a GWAS is generally large, making it impossible for one to calculate the tSVD of the whole genome at once.

If we consider eq. (2.28), the matrix $\boldsymbol{U}$ and $\boldsymbol{V}$ are the eigenvectors of $\boldsymbol{A}\boldsymbol{A}^t$ and $\boldsymbol{A}^t\boldsymbol{A}$ respectively. So for any symmetric matrix such as that of the LD matrix, $\boldsymbol{U}$ and $\boldsymbol{V}$ should be the same. Thus eq. (2.28) reduce into the problem of eigenvalue decomposition where the singular values are the magnitude of the eigenvalues. Although the eigenvalue decomposition is still an $O(n^3)$ algorithm, it has a smaller constant, therefore has a faster run time when compared to the computation of SVD.

However, even with the use of eigenvalue decomposition in place of SVD, the size of the LD matrix is still too big for a feasible computation. Given that it is unlikely for inter chromosomal LD to exists or for SNPs 1megabase (mb) apart to be in LD with each other, one can safely assume SNPs more than 1mb apart or SNPs on different chromosomes are independent of each other. We therefore

separate SNPs into 1mb bins where start of each bin are at least 1 mb away from each other. Three bins are then combined to form one window, and we perform the decomposition on each windows using eq. (2.16) and only update the $\hat{\boldsymbol{h}^2}$ for the bin forming the center of the window. We then transverse the genome with step size of 1 bin until $\hat{\boldsymbol{h}^2}$ for all bins were computed. By breaking down the genome into windows, we were able to reduce the matrix dimension which makes the analysis believable. Users can also choose distance other than 1mb as the distance between bins, allowing for a more flexible usage of the algorithm.

2.2.7 Comparing with LD SCore regression

Conceptually, the fundamental hypothesis of LDSC and our algorithm were quite different. LDSC were based on the "global" inflation of test statistic and its relationship to the LD pattern. LDSC hypothesize that the larger the LD score, the more likely will the SNP be able to "tag" the causal SNP and the heritability can then be estimated through the regression between the LD score and the test statistic.

On the other hand, our algorithm focuses more on the per-SNP level. Our main idea was that the individual test statistic of each SNPs is a combination of its own effect and effect from SNPs in LD with it. Thus, based on this concept, our algorithm aimed to "remove" the inflation of test statistic introduced through the LD between SNPs and the heritability can be calculated by adding the test statistic of all SNPs after "removing" the inflation.

Mathematically, the calculation of LDSC and our algorithm were also very different. LDSC take the sum of all R^2 within a 1cM region as the LD score and regress it against the test statistic to obtain the slope and intercept which represent the heritability and amount of confounding factors respectively. In their model,

LDSC assume that each SNPs will explain the same portion of heritability

$$\mathrm{Var}(\beta) = \frac{h^2}{M}\boldsymbol{I} \tag{2.34}$$

M = number of SNPs

β = vector containing per normalized genotype effect sizes

I = identity matrix

h^2 = heritability

As for our algorithm, the whole LD matrix were used and inverted to decompose the LD from the test statistic. There were no assumption of the amount of heritability explained by each SNPs. However, our algorithm does assumed that the mean of the χ^2 test statistic to be one (e.g. no inflation in the summary statistics), thus our estimation might inflates shall there be any confounding factors in the GWAS summary statistics.

2.3 Comparing Different LD correction Algorithms

Another important consideration in our algorithm is the bias in LD. In reality, one does not have the population LD matrix, instead we have to estimate the LD based on various reference panels such as those from the 1000 genome project (Project et al., 2012) or the HapMap project (Altshuler et al., 2010). These reference panels were a subsamples from the whole population and therefore LD estimated from the reference panels usually contains sampling bias. Under normal circumstances, because the symmetric nature of sampling error, one would expect there to be little to no bias in the estimated LD. However, in our algorithm , the R^2 is required for the estimation of heritability (eq. (2.17)). Because we were using the squared LD, the sampling error will also be squared, generating a positive bias.

On average, there were around 500 samples for each super population from the 1000 genome project reference panel. Given the relatively small sample size, the sampling bias might be large, therefore lead to systematic bias in the heritability estimation in our algorithm.

To correct for the bias, we would like to apply a LD correction algorithm to correct for the bias in the sample LD. Different authors (Weir and W G Hill, 1980; Z. Wang and Thompson, 2007) have proposed methods for the correction of sample R^2 and can be applied for the correction of sample bias in LD. Therefore we considered the following R^2 correction algorithms:

$$\text{Ezekiel} : \tilde{R^2} = 1 - \frac{n-1}{n-2}(1 - \hat{R^2}) \tag{2.35}$$

$$\text{Olkin-Pratt} : \tilde{R^2} = 1 - \frac{(n-3)(1 - \hat{R^2})}{n-2}(1 + \frac{2(1 - \hat{R^2})}{n}) \tag{2.36}$$

$$\text{Pratt} : \tilde{R^2} = 1 - \frac{(n-3)(1 - \hat{R^2})}{n-2}(1 + \frac{2(1 - \hat{R^2})}{n-3.3}) \tag{2.37}$$

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

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

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

In order to assess the performance of each individual correction methods, we perform simulations to compare the performance of our algorithm using different LD bias correction algorithms. Most importantly, we would like to assess the performance of different algorithms not only under one specific LD range, but also under the complex LD structure observed in real life scenarios. First, 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 allow us to simulate genotypes with LD structures comparable to those observed in CEU population. Of those 5,000 SNPs, 100 of them were randomly selected as the causal variant. Orr (1998) suggested that the exponential distribution can be used to approximate the genetic architecture of adaptation. As a result of that, we used the exponential distribution with $\lambda = 1$ as an approximation to the effect size distribution:

$$\begin{aligned} \theta &= \exp(\lambda = 1) \\ \beta &= \pm\sqrt{\frac{\theta \times h^2}{\sum \theta}} \end{aligned} \tag{2.40}$$

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 size ($\boldsymbol{\beta}$), we can simulate a phenotype with target heritability of h^2 as

$$\begin{aligned} \epsilon_i &\sim N(0, \mathrm{Var}(\boldsymbol{X\beta})\frac{1-h^2}{h^2}) \\ \boldsymbol{\epsilon} &= (\epsilon_1, \epsilon_2, ..., \epsilon_n)^t \\ \boldsymbol{y} &= \boldsymbol{X\beta} + \boldsymbol{\epsilon} \end{aligned} \tag{2.41}$$

To simulate the whole spectrum of heritability, we varies the target h^2 from 0 to 0.9 with increment of 0.1.

The summary statistics of association between the genotype and phenotype were then calculated using PLINK (Shaun Purcell et al., 2007). Resulting summary statistic were then input to our algorithm to estimate the heritability, using different LD correction algorithms. An independent 500 samples, a size roughly correspond to the average sample size of each super population form the 1,000 genome project,

were simulated as a reference panel for the calculation of LD matrix. This is because in reality, one usually doesn't have assess to the sample genotype and has to rely on an independent reference panel for the calculation of LD matrix. Thus this simulation procedure should provide a realistic representation of how the algorithm will be commonly used in real life scenario.

The whole process will be repeated 50 times such that a distribution of the estimate can be obtained. In summary, we simulate a large population of samples (e.g. $50 \times 1,000 + 500 = 50,500$) where 500 samples were randomly selected as a reference panel. In the subsequent iteration of simulation, 1,000 samples were randomly selected from the population *without replacement* and estimation were performed.

1. Randomly select 5,000 SNPs with maf$>$ 0.1 from chromosome 22
2. Simulate 500 samples using HAPGEN2 and used as a reference panel
3. Randomly generate 100 effect size with following eq. (2.40)
4. Randomly assign the effect size to 100 SNPs with heritability from 0 to 0.9 (increment of 0.1)
5. Simulate 1,000 samples using HAPGEN2 and calculate their phenotype according to eq. (2.41)
6. Perform heritability estimation using our algorithm with different ways of LD correction
7. Repeat step 5-6 50 times

2.4 Comparison with Other Algorithms

After identifying the optimal LD correction algorithm, we would like to compare our algorithm to existing methods for the performance in estimating the SNP-heritability. It is important for us to consider most if not all conditions in our simulation. Therefore, we would like to simulate quantitative traits and case control studies with different number of causal SNPs; quantitative traits with extreme effect sizes; and last but not least, quantitative traits with extreme phenotype sampling.

Currently, the only other algorithm that is capable to estimate the SNP-heritability using only summary statistic from GWAS is the LDSC (B. K. Bulik-Sullivan et al., 2015) whereas GCTA (J Yang et al., 2011) is the most commonly used programme for the estimation of SNP-heritability from GWAS data. Therefore, we choose to compare the performance of our algorithm to that of LDSC and GCTA. It is important to note we did not simulate any confounding factors yet for LDSC, the default intercept estimation function allows it to estimate and correct for confounding factors with an increase in SE. The simulation will therefore be unfair to LDSC with intercept estimation, as the SE is increased yet there are little confounding factors for it to correct. Thus, we also compared results from LDSC with a fixed intercept (--no-intercept) parameters to avoid bias against LDSC.

2.4.1 Sample Size

One important consideration in our simulation was the sample size of the simulated GWAS study. The sample size was the most important parameter in determining the standard error of the estimate. As sample size increases, study will be more representative of the true population. The increased number of information also means a better estimation of parameters, therefore a smaller standard error (SE). Based on

information from GWAS catalog (Welter et al., 2014), we calculate the sample size distribution using simple text mining and exclude studies with conflicting sample size information in multiple entries. The average sample size for all GWAS recorded on the GWAS catalog was 7,874, with a median count of 2,506 and a lower quartile at 940 (fig. 2.2). We argue that if the algorithm works for studies with a small sample size (e.g lower quartile sample size), then it should perform even better when the sample size is larger. Thus, we only simulate 1,000 samples in our simulation, which roughly represent the lower quartile sample size range.

2.4.2 Number of SNPs in Simulation

Another consideration in the simulation was the number of SNPs included. In a typical GWAS study, there are usually a larger number of SNPs when compared to the sample size. For example, in the PGC schizophrenia GWAS, more than 9 million SNPs were included, with around 700,000 SNPs on chromosome 1. In reality, the estimation of SNP heritability based on 700,000 SNPs can be done quickly.

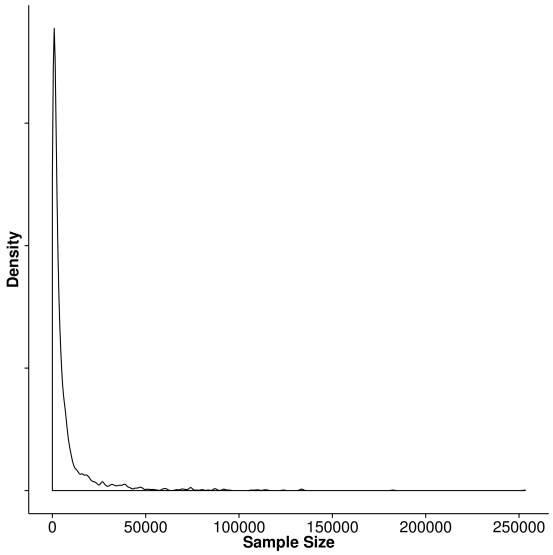


Figure 2.2: GWAS sample size distribution.

However, in our simulation, we will repeat the calculation 50(iteration) $\times$ 10(number of heritability) = 500 times for *each* condition tested. The time required to finish all the simulation quickly becomes infeasible given the large amount of SNPs. To compromise, we simulate a total of 50,000 SNPs from chromosome 1 as a balance between run time of simulation and the total SNPs simulated. With

50,000 SNPs, there are roughly 200 SNPs within a 1 mb region.

2.4.3 Genetic Architecture

Of all simulation parameter, the genetic architecture was the most complicated and important parameter. The LD pattern, the number of causal SNPs, the effect size of the causal SNPs and the heritability of the trait were all important factors contribute to the genetic architecture of a trait.

First and foremost, because the aim of the algorithm was to estimating the heritability of the trait, it is important that the algorithm works for traits from different heritability spectrum. We therefore simulated traits with heritability ranging from 0 to 0.9, with increment of 0.1.

Secondly, in real life scenario, the "causal" variant might not be readily included on the GWAS chip and were only "tagged" by SNPs included on the GWAS chip. However, to simplify our simulation, all "causal" variants were included in our simulation (e.g. perfectly "tagged").

Thirdly, to obtain a realistic LD pattern, we simulate the genotypes using the HAPGEN2 programme (Su, Marchini, and Donnelly, 2011), giving the 1000 genome CEU haplotypes as an input. In a typical GWAS , one usually only have power in detecting "common variants", defined as variants with maf ≥ 0.05. We therefore only consider scenario with "common" variants and only use SNPs with maf ≥ 0.05 in the CEU haplotypes as an input to HAPGEN2 to simulate 1,000 samples.

Finally, we would like to simulate traits with different inheritance model such as oligogenic traits and polygenic traits. We therefore varies the number of causal SNPs (k) with $k \in \{5, 10, 50, 100, 500\}$. The effect size were then simulated using eq. (2.40) and the phenotype were simulated using eq. (2.41).

For GCTA, the sample genotypes were provided to calculate the genetic relationship matrix and the sample phenotypes were used in combination with the genetic relationship matrix to estimate the heritability.

On the other hand, for LDSC and our algorithm, an independent 500 samples were simulated as the reference panel for the calculation of LD scores and LD matrix, mimicking real life scenario where an independent reference panel were used. The genotype association test statistics calculated from PLINK and the LD score / LD matrix were then used for the estimation of SNP heritability for LDSC and our algorithm respectively.

The whole process will be repeated 50 times such that a distribution of the estimate can be obtained. 10 independent population were simulated and the whole processed were repeated. 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. (2.40), 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. (2.41)
6. Perform heritability estimation using our algorithm, GCTA, LDSC with fixed intercept and LDSC with intercept estimation.
7. Repeat step 5-6 50 times
8. Repeat step 1-7 10 times

2.4.4 Extreme Effect Size

On top of the original quantitative trait simulation, another condition we were interested in was the performance of the algorithms when there is a small amount of SNPs with a much larger effect size. This can be observed in disease such as Hirschsprung's disease. The Hirschsprung's disease is a congenital disorder where deleterious mutations on *RET* account for ≈50% of the familial cases yet there is still missing heritability, suggesting that there might be more variants with small effects that have not been identified (Gui et al., 2013).

To simulate extreme effect size, we consider scenarios where m SNPs accounts 50% of all the effect size with $m \in \{1, 5, 10\}$. The effect size was then calculated as

$$\begin{aligned} \beta_{eL} &= \pm\sqrt{\frac{0.5h^2}{m}} \\ \beta_{eS} &= \pm\sqrt{\frac{0.5h^2}{100-m}} \\ \beta &= \{\beta_{eL}, \beta_{eS}\} \end{aligned} \tag{2.42}$$

The effect size were then randomly assigned to 100 causal SNPs and phenotype will be calculated as in eq. (2.41). The simulation procedure then becomes

1. Randomly select 50,000 SNPs with maf > 0.05 from chromosome 1
2. Simulate 500 samples using HAPGEN2 and used as a reference panel
3. Randomly generate 100 effect size where m has extreme effect, following eq. (2.42), 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. (2.41)

6. Perform heritability estimation using our algorithm, LDSC with fixed intercept, LDSC with intercept estimation and GCTA
7. Repeat step 5-6 50 times
8. Repeat step 1-7 10 times

2.4.5 Case Control Studies

The simulation of case control studies was similar to the simulation of quantitative trait. However, there were two additional parameters to consider: the population prevalence and the observed prevalence. These parameters were required to simulate the samples under a liability model for case control studies.

Although there were only two additional parameter, it is significantly more challenging for to simulate when compared to the simulation of quantitative traits. It is mainly because of the number of samples required to simulate under the liability threshold model. Take for example, if one like to simulate a trait with population prevalence of p and observed prevalence of q and would like to have n cases in total, one will have to simulate $\min(\frac{n}{p}, \frac{n}{q})$ samples. Considering the scenario where the observed prevalence is 50%, the population prevalence is 1%, if we want to simulate 1,000 cases, a minimum of 100,000 samples will be required.

Given limited computer resources, it will be infeasible for us to simulate 1,000 cases with 50,000 SNPs when the population prevalence is small (e.g. 1%). To simplify the simulation and reduce the burden of computation, we limited the observed prevalence to 50% and varies the population prevalence p such that $p \in \{0.5, 0.1, 0.05, 0.01\}$. Most importantly, we reduce the number of SNPs simulated to 5,000 on chromosome 22 instead of 50,000 SNPs on chromosome 1. The change from chromosome 1 to chromosome 22 allow us to reduce the number of SNPs without significantly changing the SNP density. We acknowledged that the

current simulation was relatively brief, however, it should serves as a prove of concept simulation to study the performance of the algorithms under the case control scenario.

In the case control simulation, we randomly select 5,000 SNPs from chromosome 22 with maf ≥ 0.05 in the CEU haplotypes as an input to HAPGEN2. We then randomly select k SNPs where $k \in \{10, 50, 100, 500\}$, each with effect size simulated based on eq. (2.40). In order to simulate a case control samples with 1,000 cases, we then simulate $\frac{1{,}000}{p}$ samples and calculate their phenotype using eq. (2.41). The phenotype was then standardized and cases were defined as sample with phenotype passing the liability threshold with respect to p. An equal amount of samples were then randomly selected from samples with phenotype lower than the liability threshold and defined as controls.

Finally, the case control simulation were performed as:

1. Randomly select 5,000 SNPs with maf> 0.05 from chromosome 22
2. Simulate 500 samples using HAPGEN2 and used as a reference panel
3. Randomly generate k effect size following eq. (2.40) 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. (2.41)
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

2.4.6 Extreme Phenotype Sampling

With a limited budget, it is usually difficult to obtain adequate sample size for a GWAS, leading to studies with insufficient power. A possible approach was to perform the extreme phenotype sampling which only select samples with phenotypes on the extreme end of the distribution. This allow one to obtain the same degree of power using a smaller sample size. It is therefore interesting to see how the selection of extreme phenotype will affect the performance of the SNP heritability estimation.

Herein, we performed simulations on extreme phenotype sampling. 50,000 SNPs with maf > 0.05 were selected from chromosome 1 and were used as an input for HAPGEN2 similar to that in the quantitative trait simulation. 500 samples were first simulated to serves as the reference panel.

From the 50,000 SNPs we randomly select 100 as the causal SNPs and their effect was simulated based on eq. (2.40). We then simulate $\frac{1000}{K\times 2}$ samples where K is the portion of extreme samples selected (e.g. 0.1 or 0.2). Phenotype of the individuals were then simulated using eq. (2.41) and were standardized. 500 samples were selected at both end of the phenotype distribution (500 top and 500 bottom, total of 1,000) and were used for the statistical analysis. To compare the performance of extreme phenotype sampling and the general random sampling strategies, we also drawn 1,000 samples from the $\frac{1000}{K\times 2}$ samples at random and perform statistic analysis on them. At the end, we compare the heritability estimated from samples using the two different strategies and the whole procedure was repeated 50 times.

It was noted that the extreme phenotype sampling were not supported by the LDSC and GCTA. To allow comparison in such scenario, we apply the extreme phenotype adjustment from Pak C Sham and Shaun M Purcell (2014) to the

estimation obtained from LDSC and GCTA. 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 a reference panel
3. Randomly generate 100 effect size following eq. (2.40), with heritability ranging from 0 to 0.9 (increment of 0.1)
4. Randomly assign the effect size to 100 SNPs
5. Simulate $\frac{1{,}000}{K\times 2}$ samples using HAPGEN2 where K is the portion of extreme samples selected and $K \in \{0.1, 0.2\}$
6. Phenotype of the samples were calculated according to eq. (2.41) and were standardized
7. Top 500 and bottom 500 samples (ranked by phenotype) were selected, representing the extreme phenotype sample selection strategy
8. 1,000 samples were also randomly selected to represent the general random sampling strategy
9. Perform heritability estimation using our algorithm, GCTA, LDSC with fixed intercept and LDSC with intercept estimation.
10. Adjust the estimation from LDSC and GCTA by the extreme phenotype adjustment factor as proposed by Pak C Sham and Shaun M Purcell (2014)
11. Repeat step 5-10 50 times
12. Repeat step 1-11 10 times

2.5 Application to Real Data

To test the performance of our algorithm under real life scenario, we apply our algorithm to the PGC data, including Bipolar (Psychiatric GWAS Consortium Bipolar Disorder Working Group, 2011), Major depression disorder (Stephan Ripke, Naomi R Wray, et al., 2013), and schizophrenia (Stephan Ripke, B. M. Neale, et al., 2014). We also performed LDSC alongside our algorithm to compare the results from the two algorithm. Unfortunately, as the sample genotypes were not provided, we cannot perform GCTA analysis. For the bipolar and major depression data, we performed liftover (Hinrichs et al., 2006) to convert the genomic coordinates to genome version hg19 such that it is compatible with the data from 1000 genome.

The reference genome were downloaded from 1000 genome (Project et al., 2012) and were converted to plink binaries using plink --vcf function. We used the European super population which contain a total of 503 samples where singleton and non-biallelic SNPs were filtered out. To filter related samples, genotypes were first pruned before the identity by descent (IBD) were calculated. Samples pairs with pi hat larger than 0.125 were considered related, which roughly correspond to third degree relativeness. Samples were removed on a stepwise fashion where samples related to most samples were removed first, until none of the samples were related. In total, 57 samples were removed, leaving us with 446 reference samples. For LDSC, we calculated the LD score based on the 446 samples using a 1mb window size and filter out SNPs with maf < 0.1. To allow for the adjustment of confounding factors, we performed the intercept estimation with LDSC.

As only summary statistics were available, there is no way for us to determine the male to female ratio in the samples. This makes the analysis on the sex chromosome problematic, thus we only performed the SNP heritability estimation on the autosomal chromosomes.

All the studies were case control GWAS, thus the population prevalence of the trait has to be provided in order to adjust for the attenuation bias. Therefore we used prevalence of 0.15 for major depression disorder and 0.01 for schizophrenia and bipolar disorder.

Unfortunately, the density of the SNPs in the PGC schizophrenia samples were too high, making it impossible for SHREK to finish the analysis with the current available computation resources using the default window size, even if we separate the analysis to individual chromosome. To facilitates the analysis, we reduce the distance between each bin to 50,000 bp instead of the original 1mb distance. This might leads to inflation in the estimates and therefore the heritability estimates from SHREK should only be considered as an upper bound of the true SNP heritability.

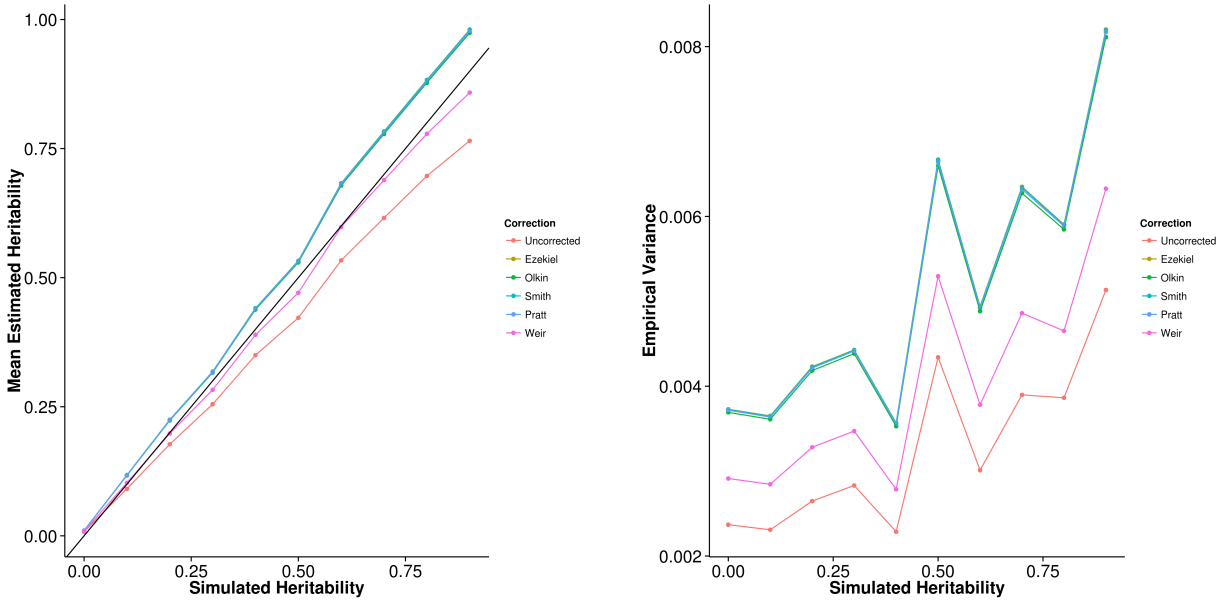
2.6 Result

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

2.6.1 LD Correction

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

Different bias correction algorithms were applied and their performance was compared (fig. 2.3a). From the graph, it was observed that when no bias correction was applied, the mean estimation were in general downwardly biased.



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

Figure 2.3: Effect of LD correction to Heritability Estimation. We compared the performance of our algorithm when different R^2 bias correction algorithm was used. When no bias correction was carried out, a downward bias was observed. After the application of the bias correction algorithms, the mean estimations of all except in the case of Weir eq. (2.39) algorithms leads to an overestimation of heritability. On the other hand, the corrections all lead to increase in variance of the estimation.

This was consistent with our expectation of a general upward bias in sample R^2 which will downwardly penalize the resulting heritability estimation. On the other hand, the bias correction algorithms all worked as expected where they increases the mean estimation of heritability as removing the upward bias in the sample R^2, the heritability estimation should increase. However for most algorithms except for Weir's formula (eq. (2.39)) an over adjustment were observed, leading to a general upward bias in the estimation. Taking into account of the variance of estimation (fig. 2.3b), Weir's formula was the most suitable for SHREK where not only it reduces the bias in the final heritability estimation, it does not introduce too much additional variance into the estimation. As a result of that, we selected the Weir's formula as our default LD correction algorithm.

2.6.2 Comparing with Other Algorithms

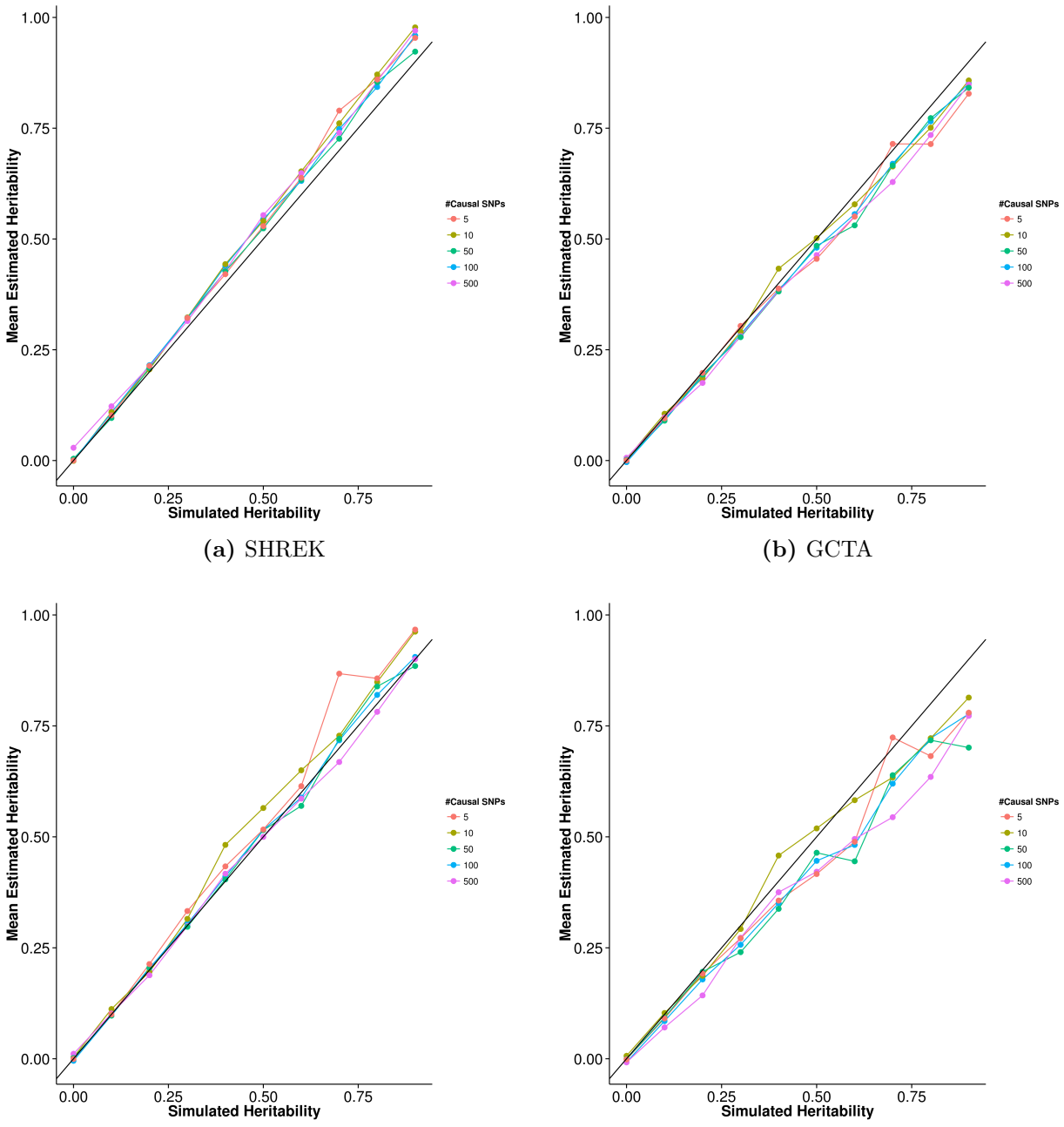
Having selected the optimal LD correction algorithm, we then compared the performance of SHREK with existing algorithms to understand the relative performance of these algorithms under different conditions. First, we examined the performance of the algorithms under the quantitative trait scenario where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

Quantitative Trait Simulation

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

First, the mean heritability estimation were compared to the simulated heritability in order to identify the bias in estimation for each algorithms. From the graph (fig. 2.4), it was observed that the mean estimations of SHREK has a small upward bias (fig. 2.4a). However, the bias was insensitive to the change in number of causal SNPs suggesting that SHREK is relatively robust to trait complexity. On the other hand, estimations form GCTA were moderately biased downward (fig. 2.4b), similar to the estimations from LDSC with intercept estimation (fig. 2.4d), but with a smaller variability. Finally, when the intercept is fixed, LDSC has the smallest bias when the trait is polygenic but an upward bias is also observed when the number of causal SNPs is small.

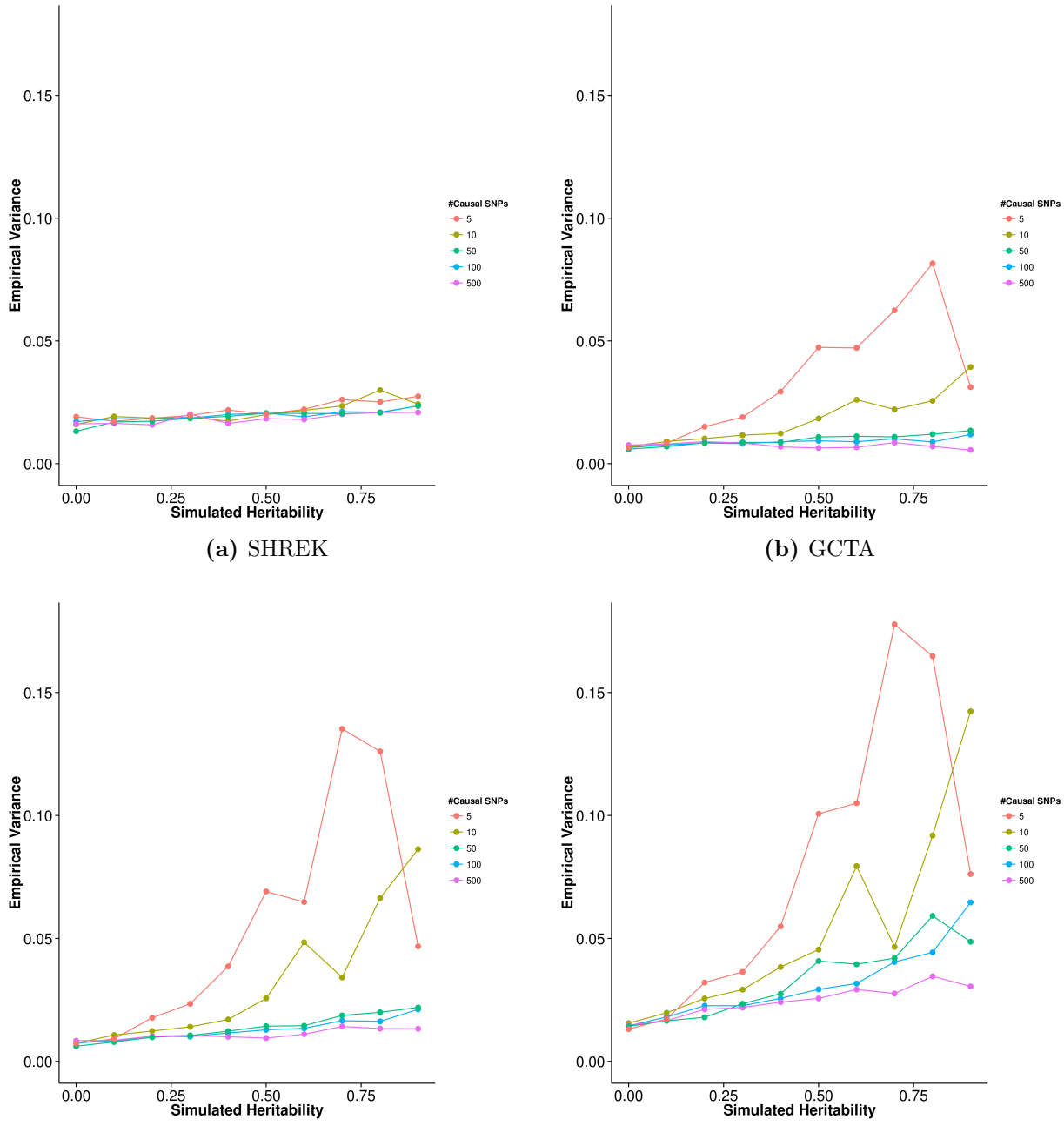
Furthermore, while comparing the empirical variance of the estimates (fig. 2.5), variance of estimations from LDSC were sensitive to the number of causal SNPs. As the number of causal SNPs decreases (figs. 2.5c and 2.5d), the variance of LDSC



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

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

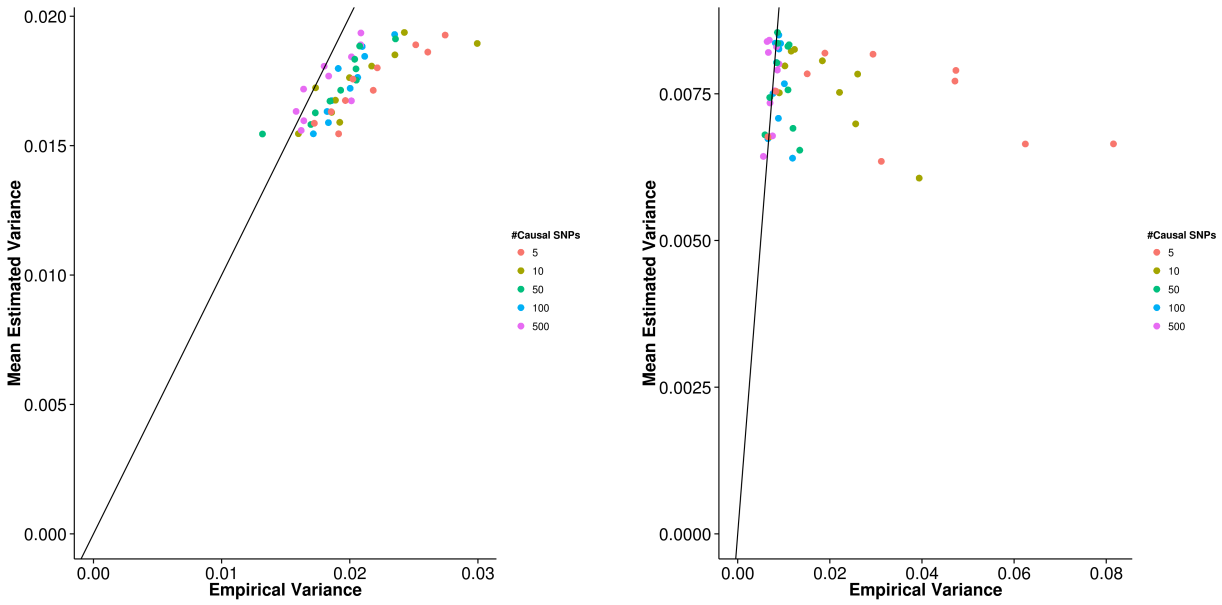
Figure 2.4: Mean of results from quantitative trait simulation with random effect size simulation. Estimations form SHREK were slightly biased upwards whereas GCTA and LDSC with intercept estimations both biased downwards. On the other hand, LDSC with fixed intercept provides least biased estimates under polygenic conditions. However, when the number of causal SNPs is small (e.g. 5 or 10), an upward bias was observed.



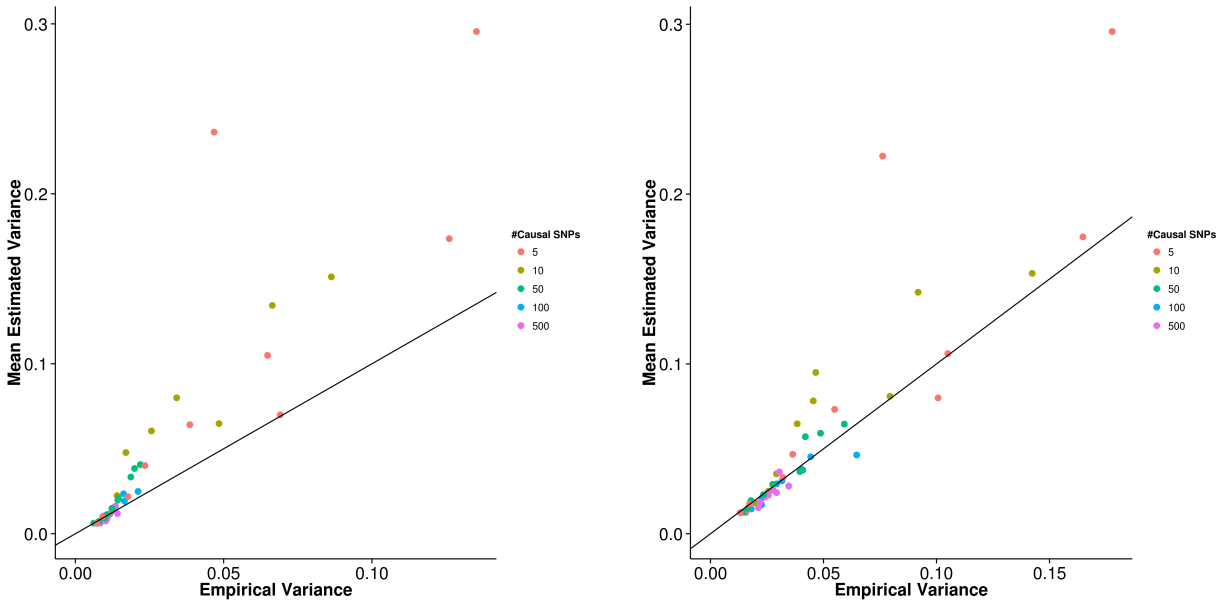
(a) SHREK **(b)** GCTA

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

Figure 2.5: 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 2.6: Estimated variance of results from quantitative trait simulation with random effect size simulation when compared to the empirical variance. GCTA has the best estimate of its empirical variance under the polygenic conditions whereas SHREK tends to under-estimate its empirical variance. On the other hand, LDSC to over-estimate the variance especially when the number of causal SNPs is small.

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 out perform LDSC with intercept estimation when there are minimal confounding variables.

estimates increases, similar to what was reported by B. K. Bulik-Sullivan et al. (2015). The variance were also higher when intercept estimation was performed. On the other hand, although the variance of SHREK was relatively higher when compared to LDSC when the intercept was fixed, the variation of its estimations was insensitive to the number of causal SNPs. When the number of causal SNPs was small, the variance of estimates from SHREK can even be lower than LDSC (fig. 2.5a). Finally, of all the algorithms, the estimations from GCTA has the lowest variation when compared to other algorithm (fig. 2.5b), except when it was the case of 5 causal SNPs where it has a slightly higher variance in comparison to SHREK when the simulated heritability was high (e.g. ≥ 0.8).

Another important factor to consider was the estimation of the SE. Of all the algorithms, GCTA (fig. 2.6b) has the best estimate, follow by SHREK (fig. 2.6a). However, it was noted that a consistent underestimation of variance was observed with SHREK whereas GCTA only underestimate the variance when the number of causal SNPs is small. On the other hand, when the intercept was fixed (fig. 2.6c), LDSC cannot accurately estimate its variance and tends to overestimate, especially when the number of causal SNPs were small. When intercept estimations

was performed (fig. 2.6d), the estimation of variance was relatively better yet the overestimation were still observed when the number of causal SNPs is small.

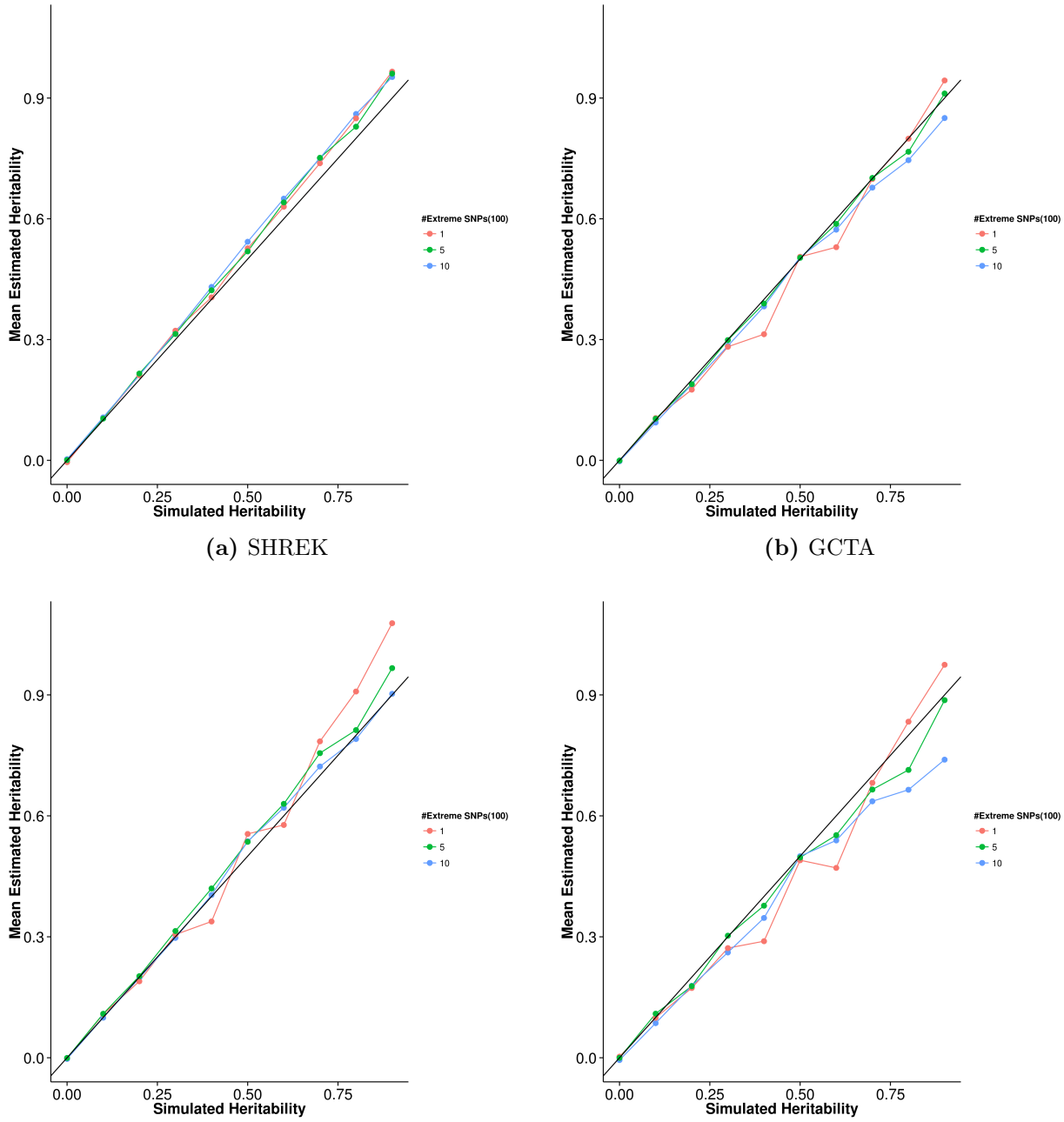
By taking into consideration of both the bias and variance of the estimates, GCTA has the best overall performance. Under the oligogenic condition (e.g. number of causal SNPs ≤ 10), SHREK has relatively better performance when compared to LDSC. Whereas under the polygenic condition, LDSC has better performance.

Quantitative Trait Simulation with Extreme Effect Size

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

When assessing the mean estimation of heritability (fig. 2.7), the performance of the algorithms were similar to that in the quantitative trait simulation. The only exception was when 1 SNP with large effect was simulated, the mean estimation of LDSC and GCTA fluctuates (figs. 2.7b to 2.7d). The same fluctuation was not observed in SHREK (fig. 2.7a). Similarly, the empirical variance of the estimation (fig. 2.8) from GCTA and LDSC increases and fluctuates when only 1 SNP with large effect was simulated. It was most obvious in the case of LDSC where the variance increased drastically as the heritability is high (fig. 2.8c). However, SHREK does not seems to be affected and were robust to the number of SNPs with large effect.

The estimated variance were also affected by the number of SNPs with large effect where the larges discrepancy between the estimated and empirical variance was observed when only 1 SNP with large effect was simulated. It was observed that



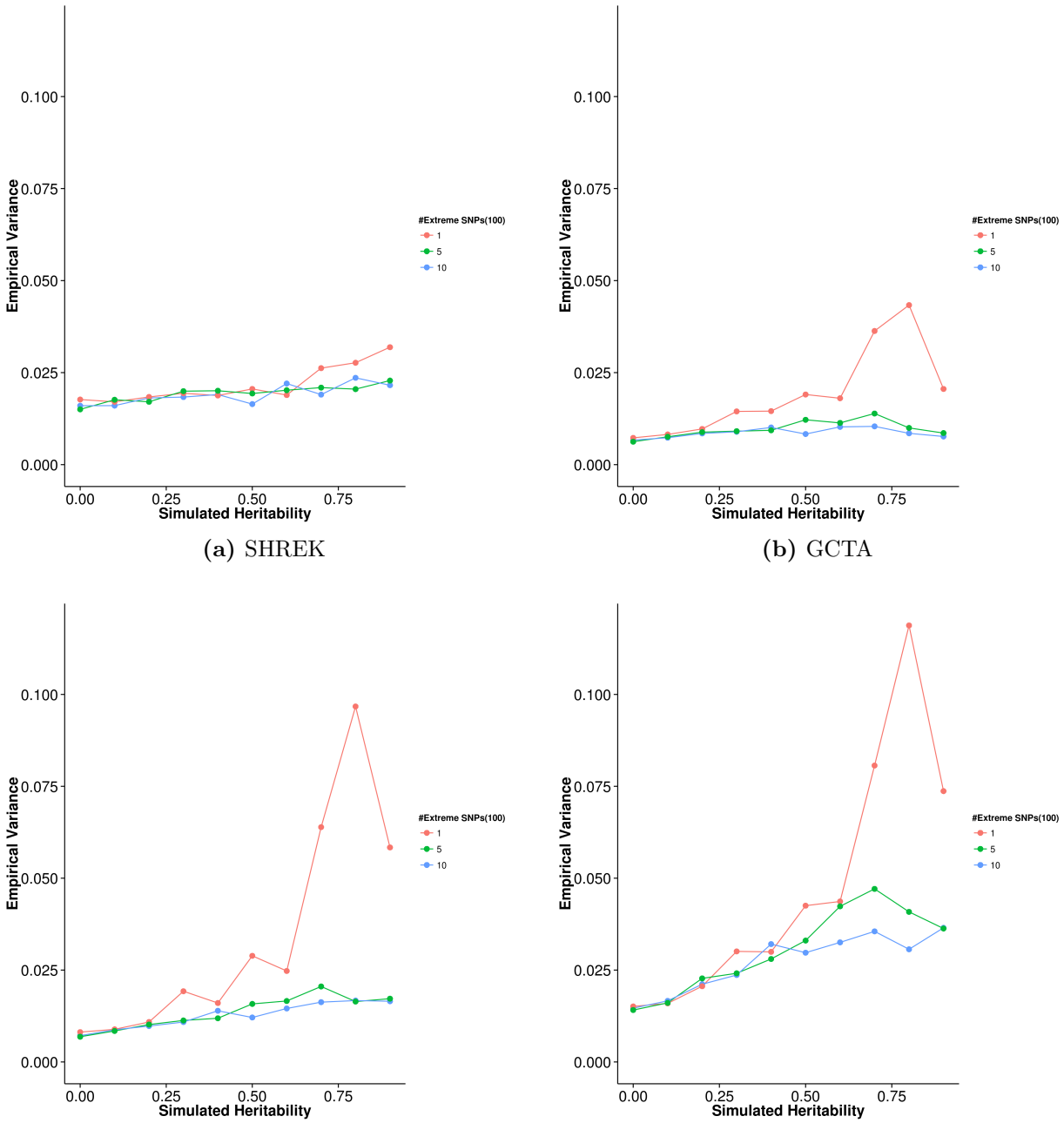
(a) SHREK

(b) GCTA

(c) LDSC with fix intercept

(d) LDSC with intercept estimation

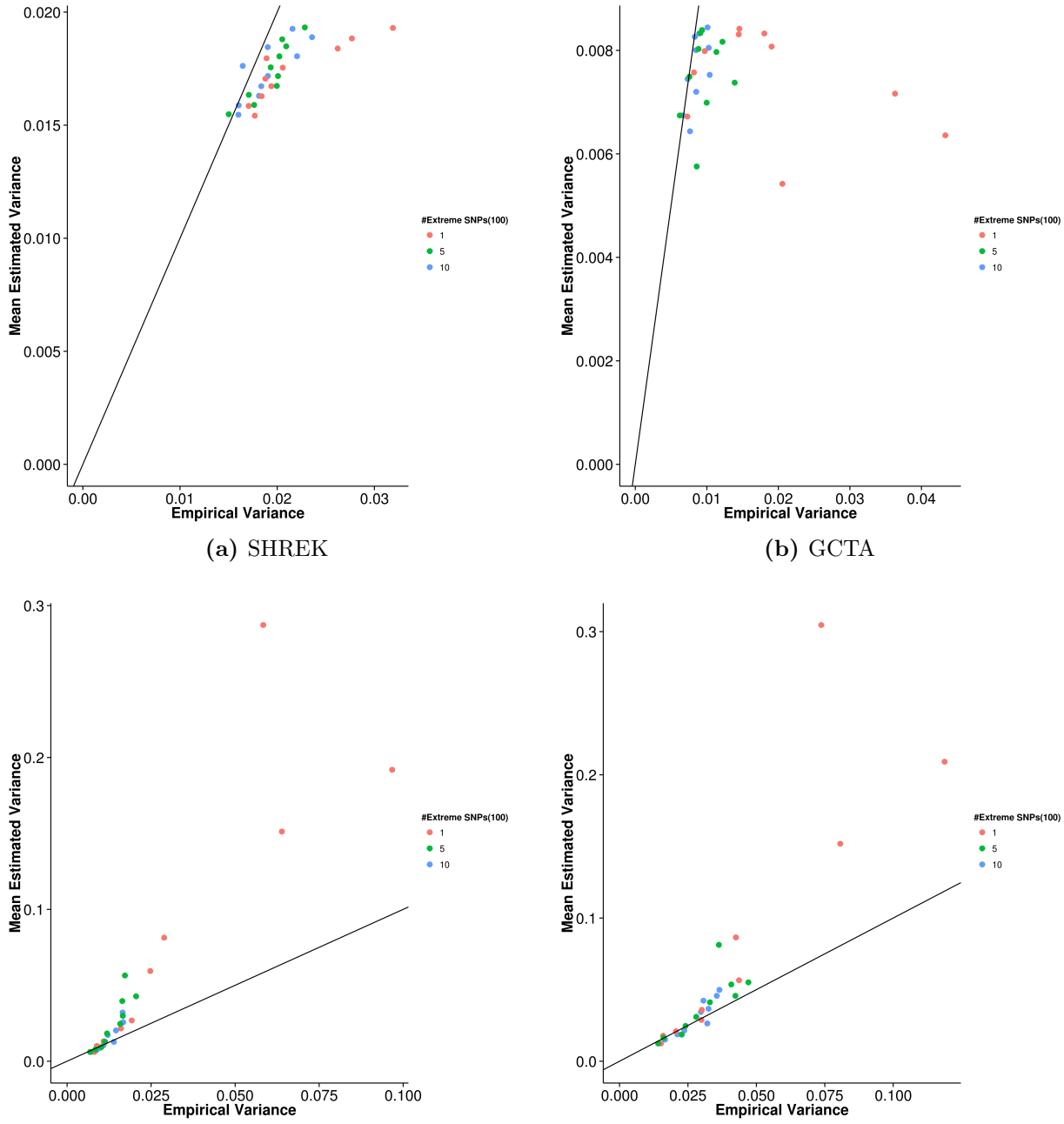
Figure 2.7: Mean of results from quantitative trait simulation with extreme effect size simulation. It was observed that the mean estimation of heritability of SHREK is not affected by the number of SNP(s) with large effect but with slight upward bias. On the other hand, the mean estimation of LDSC and 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 2.8: 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 increase slightly when the simulated heritability is large and with only 1 SNP with extreme effect. Suggesting that it 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 2.9: Estimated variance of results from quantitative trait simulation with extreme effect size simulation when compared to the empirical variance. 100 causal SNPs were simulated. SHREK and GCTA generally under-estimate the variance with the magnitude of bias being the highest when there is only 1 SNP with extreme effect. On the other hand, LDSC tends to over-estimate the variance and it can overestimate the variance by more than 3 folds when there is only 1 SNP with extreme effect.

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

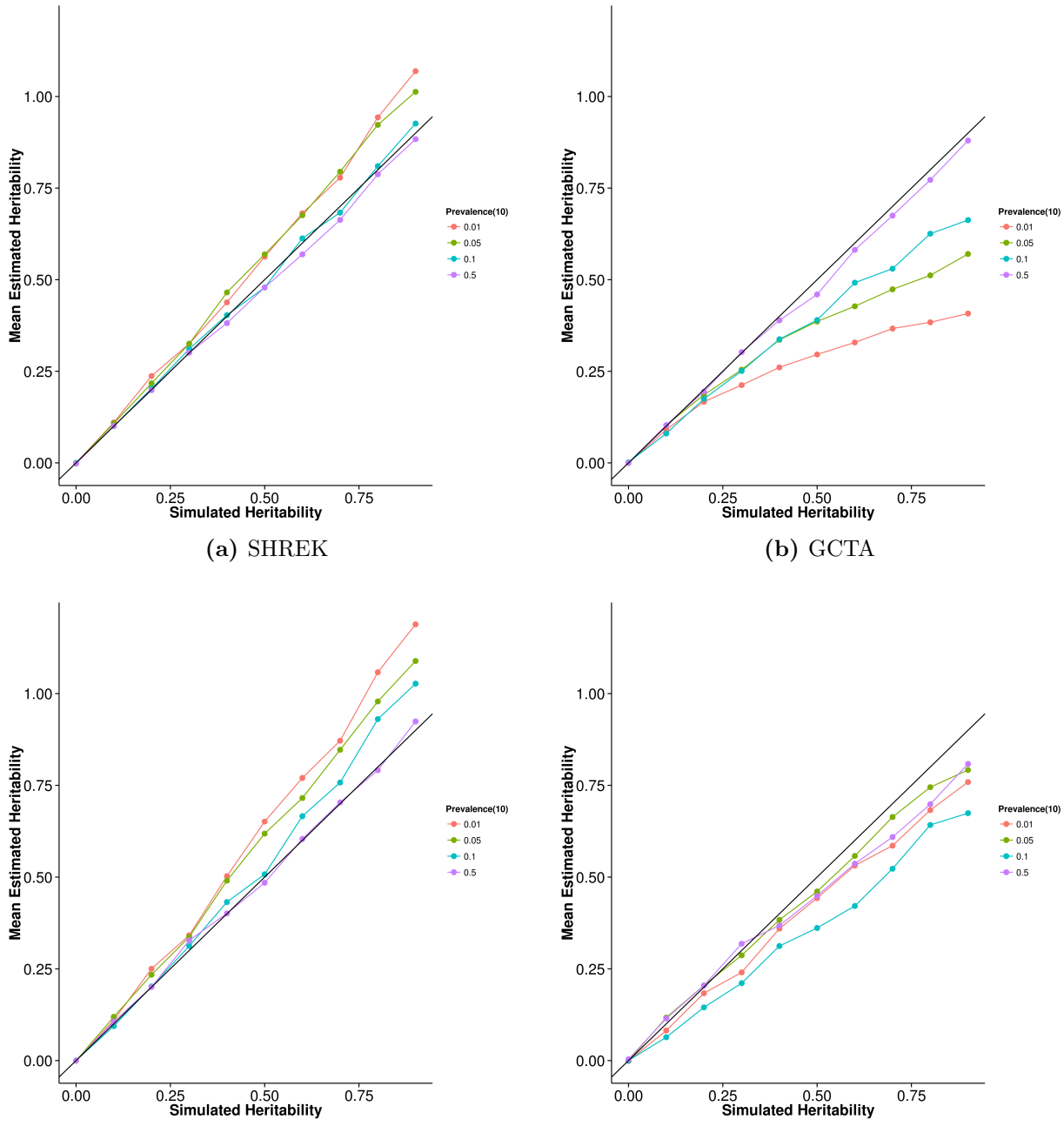
Table 2.2: MSE of quantitative trait simulation with extreme effect size. Of all the algorithms, GCTA has the lowest MSE in all situations. When comparing the performance of SHREK and LDSC, SHREK only has a better performance when there is one SNP with large effect. For other scenarios, LDSC with fixed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

both SHREK and GCTA tends to underestimates their empirical variance whereas LDSC tends to overestimates the empirical variance. The difference between the estimated and empirical variance for LDSC with fixed effect can be as much as 3 fold.

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

Case Control Simulation

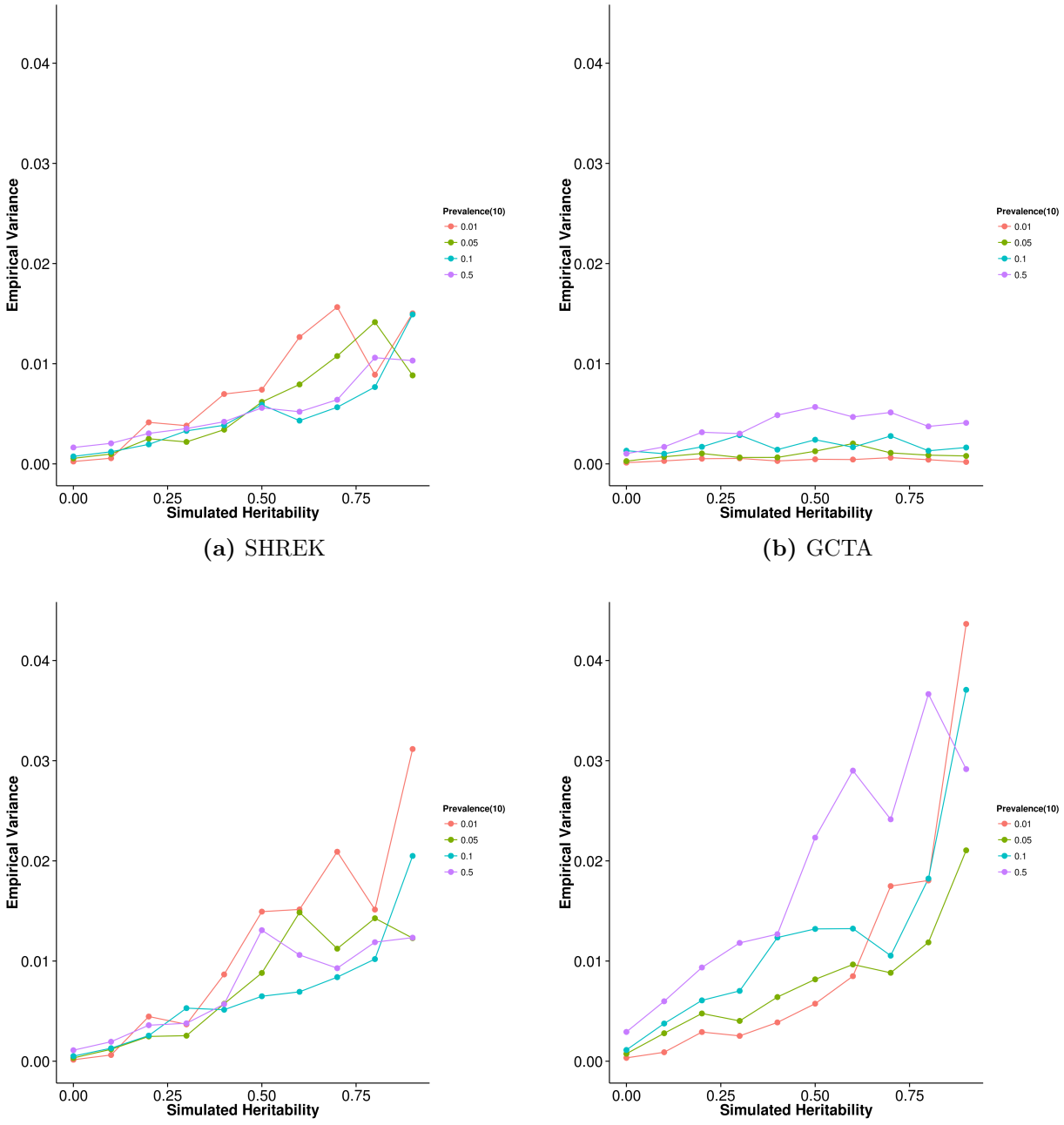
Nowadays, most of the GWAS are Case Control studies, thus it is important to test the performance of the algorithms when dealing with case control samples. In the case control simulation, we varied the population prevalence and the trait heritability. We also varied the number of causal SNPs to assess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



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

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

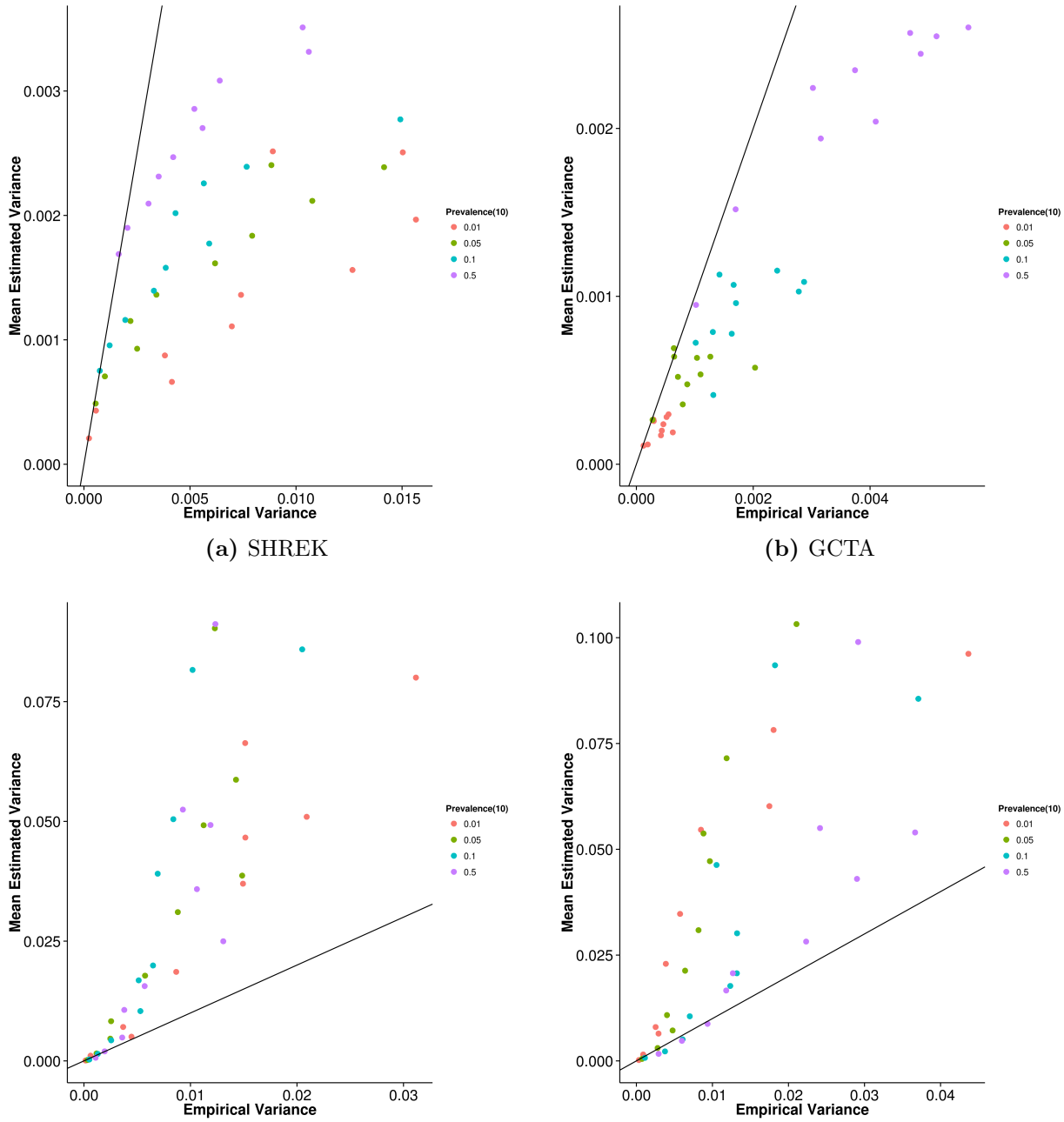
Figure 2.10: Mean of results from case control simulation with random effect size simulation with 10 causal SNPs. The performance of GCTA was as suggested by Golan, Eric S Lander, and Rosset (2014) where there was an underestimation as prevalence decreases. On the other hand, the upward bias of both 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 2.11: 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 2.12: 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.

First, we simulated traits with 10 causal SNPs. From the graph, it is clear that the population prevalence has a significant impact to the performance of the algorithms (fig. 2.10). The performance of GCTA was as suggested by Golan, Eric S Lander, and Rosset (2014) where the degree of underestimation increases as the prevalence decreases. On the other hand, the opposite effect was observed for SHREK and LDSC with fixed intercept. Interestingly, when allow the estimate the intercept, the heritability estimated from LDSC becomes underestimated. The magnitude of the bias also decreases, suggesting that the intercept estimation might have corrected for part of the bias of LDSC. The same pattern were also observed when the number of causal SNPs increases (figs. 2.18, 2.21 and 2.24), suggesting that the effect of number of causal SNPs were not the main contributor to the difference in bias.

As one inspect the empirical variance of the algorithms, GCTA clearly has the smallest average empirical variance among the algorithms (fig. 2.11b) where LDSC with intercept estimation has the largest empirical variance (fig. 2.11d). Unlike the quantitative trait simulation, the empirical variance of the estimates from SHREK (fig. 2.11a) seems to be very close to that of LDSC with fixed intercept (fig. 2.11c). When the heritability of the trait is high, the empirical variance of SHREK is even lower than that of LDSC with fixed intercept. As one increases the number of causal SNPs, the empirical variance of all algorithms decreases (figs. 2.19, 2.22 and 2.25) agreeing with the results from the quantitative trait simulation.

On the other hand, both SHREK (fig. 2.12a) and GCTA (fig. 2.12b) underestimates their empirical variance whereas LDSC overestimates its empirical variance no matter if the intercept estimation was performed (fig. 2.12). As the number of causal SNPs increases (figs. 2.20, 2.23 and 2.26), the bias of variance estimation remain unchanged for SHREK. However, for LDSC, the magnitude of bias of variance estimation reduces as the number of causal SNPs increases and were able to

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

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

provide a relatively accurate estimation of its empirical variance when there were 500 causal SNPs (fig. 2.26c).

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

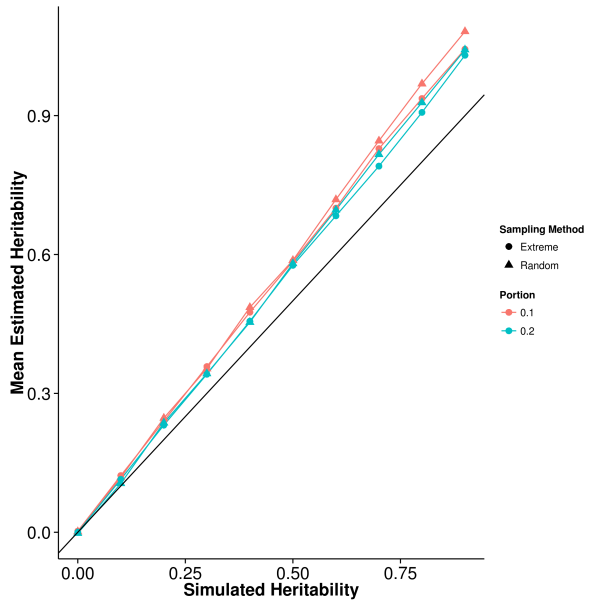
In general, the effects of the number of causal SNPs in the case control simulation agrees with what was observed in the quantitative trait simulations where as the number of causal SNPs increases the MSE tends to decrease for all algorithms, with SHREK least sensitive. Finally, it is important to note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

2.6.3 Extreme Phenotype Simulation

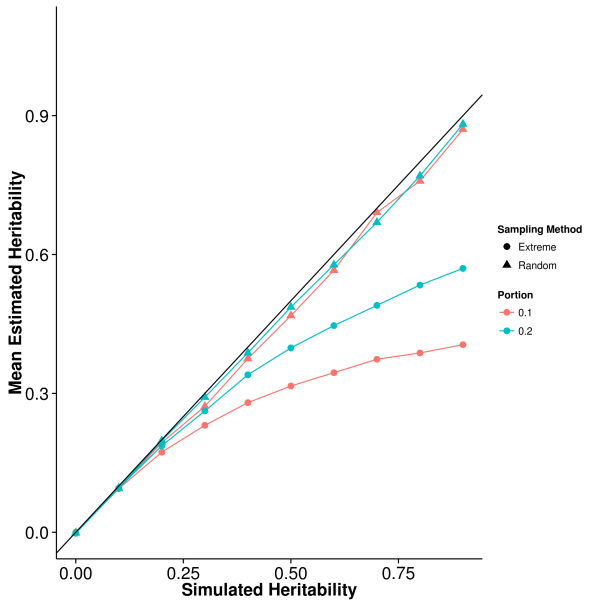
Sometimes, when budget is limited, it is not possible to include all samples in the final GWAS. By using appropriate sampling strategy, such as that of extreme phenotype sampling (Peloso et al., 2015), one can increase the power of the association study. Here we perform simulations using extreme phenotype sampling and study the effect of this selection on the performance of heritability estimations. The random sampling procedure were also performed in our simulations such that a clear comparison can be made between the power of extreme phenotype sampling and the traditional random sampling.

From the graph (fig. 2.13), it was observed that performance of SHREK and LDSC were similar to what was observed in the quantitative trait simulation, where when the random sampling strategy were used, higher estimation were usually obtained. Interestingly, GCTA performs poorly when extreme phenotype sampling was performed. As the portion of sample sampled decreases, the bias of the estimates from GCTAincreases (fig. 2.13b).

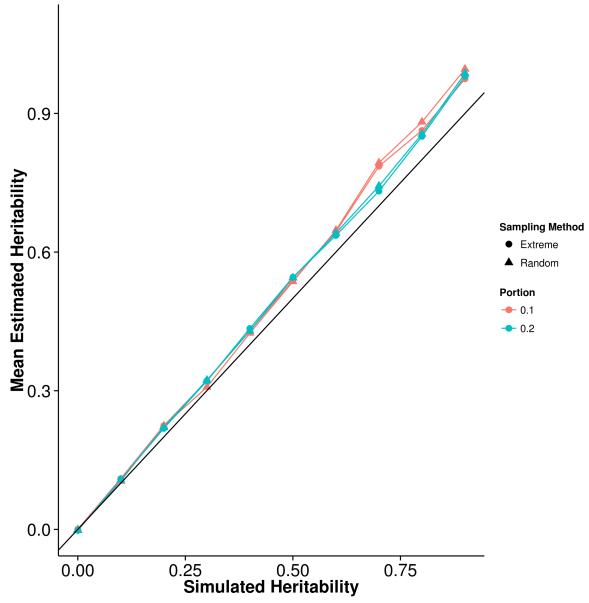
When comparing the empirical variance, the random sampling strategy



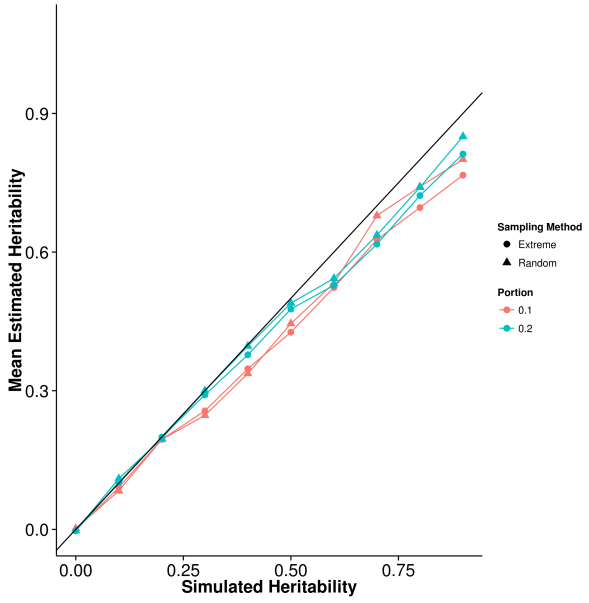
(a) SHREK



(b) GCTA

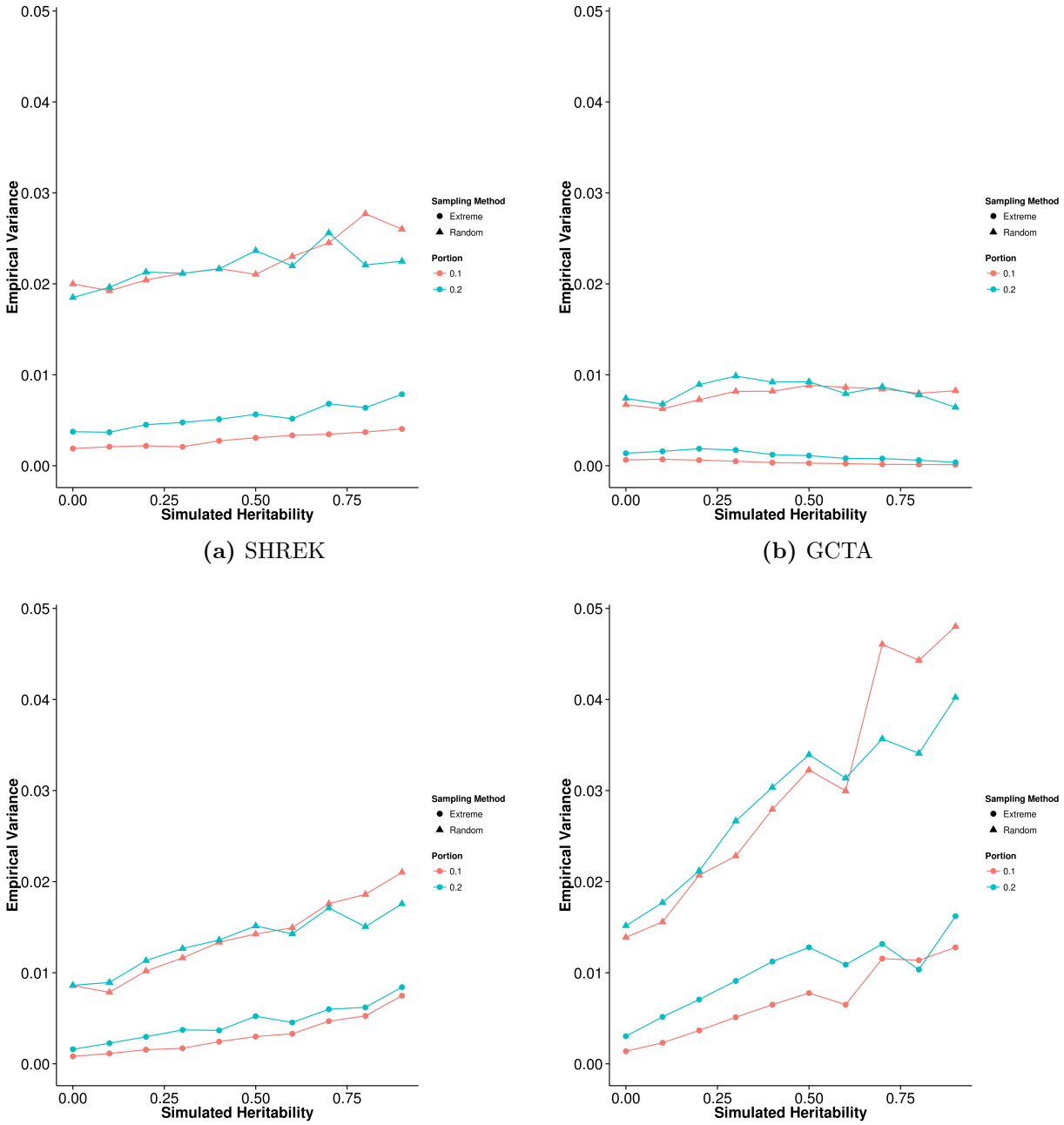


(c) LDSC with fix intercept



(d) LDSC with intercept estimation

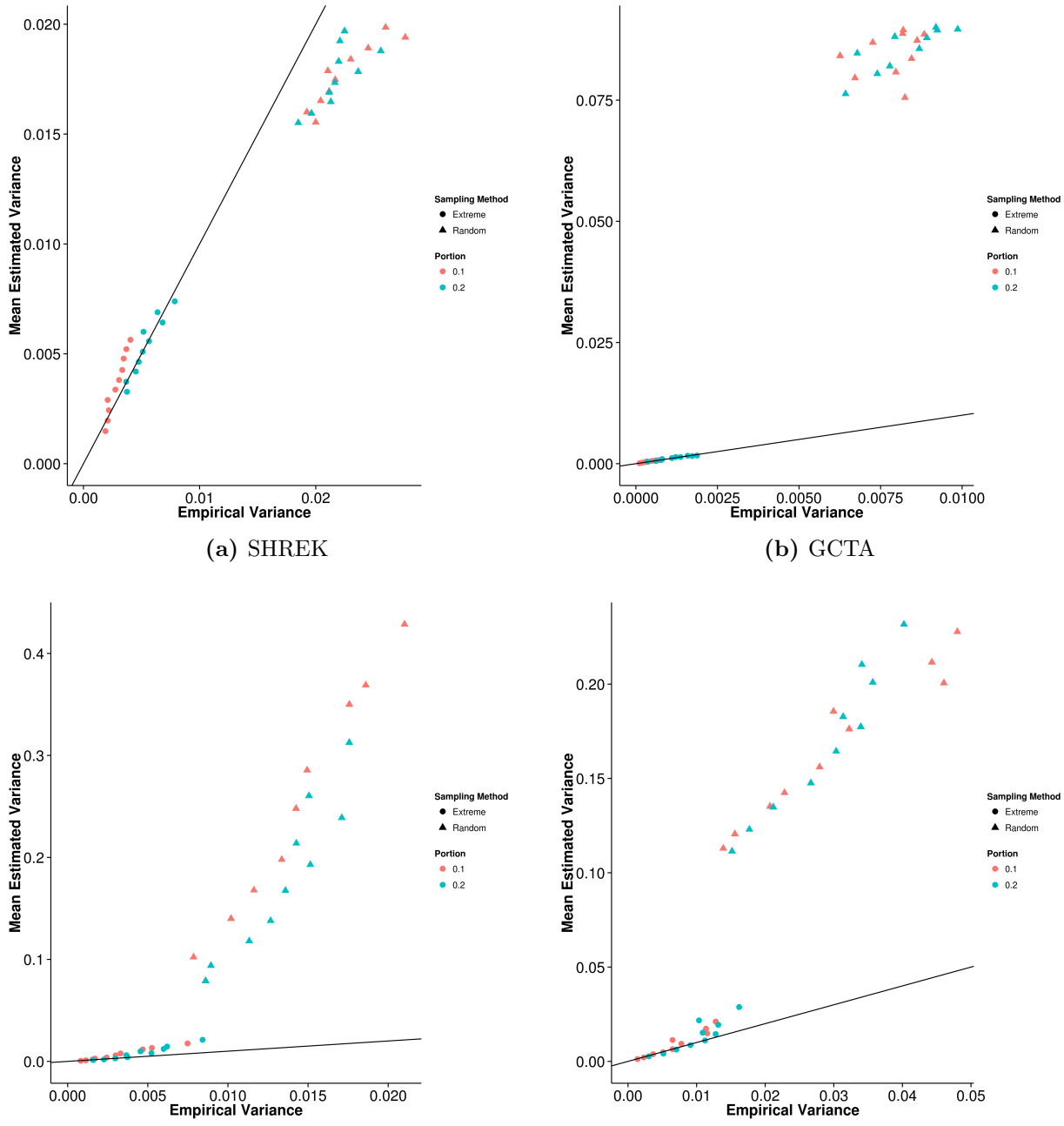
Figure 2.13: Mean of results from extreme phenotype simulation. The performance of the algorithms when random sampling was performed were similar to what was observed in the quantitative trait simulation. However, when extreme phenotype was performed, a larger under estimation was observed for GCTA and 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 2.14: 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 2.15: Estimated variance of results from extreme phenotype selection when compared to empirical variance. Surprisingly, except for SHREK, the estimated variance from LDSC and GCTA under the random sampling condition was much higher than the empirical variance. It is much different from the estimated variance from the quantitative trait simulation and further investigations are required to understand this discrepancy.

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

Table 2.4: Here, we compared the MSE of random sampling (Rand) against the MSE of Extreme phenotype sampling (Extreme). With the exception of GCTA, the extreme phenotype selection generally produce a smaller MSE when compared to random sampling. However, for GCTA, because of the large bias introduced by extreme phenotype sampling (fig. 2.13b), the MSE is much higher when extreme phenotype sampling was performed.

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

Strangely, although the empirical variance under the random sampling strategy is the same as what was observed in the quantitative trait simulation, there was a large discrepancy in the estimated variance where a tenfold overestimation was observed for LDSC and GCTA (fig. 2.15). More surprisingly, SHREK was unaffected. We are uncertain of the origin of such problem and further investigations are required.

2.6.4 Application to Real Data

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

It was estimated that the heritability for major depression disorder is around 0.256 by SHREK and 0.161 by LDSC whereas the heritability of bipolar was estimated to be around 0.312 by SHREK and 0.185 by LDSC (table 2.5). As for schizophrenia, the heritability was estimated to be around 0.133 by LDSC

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

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

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

2.7 Discussion

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

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

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

2.7.1 LD Correction

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

When comparing impact of different bias correction algorithm on the performance of SHREK, it was observed that majority of the algorithms, except that of eq. (2.39), inflates the heritability estimated, suggesting that there was an overestimation, whereas when the sampling bias left uncorrected, the estimates were biased

downward, as one would expect. The superior performance of eq. (2.39) leads us to use it as our default LD sampling bias correction algorithm.

What was surprising was that in the quantitative trait simulation, an overestimation of heritability was observed despite using eq. (2.39) for LD correction. This overestimation was similar to what was observed in the previous LD correction simulations where 5,000 SNPs on chromosome 22 were simulated. It is possible that despite the superior performance of eq. (2.39), small imprecisions were still introduced to the LD matrix during the bias correction. When the number of SNPs increases, these imprecisions cumulates, thus leads to bias in the final heritability estimates.

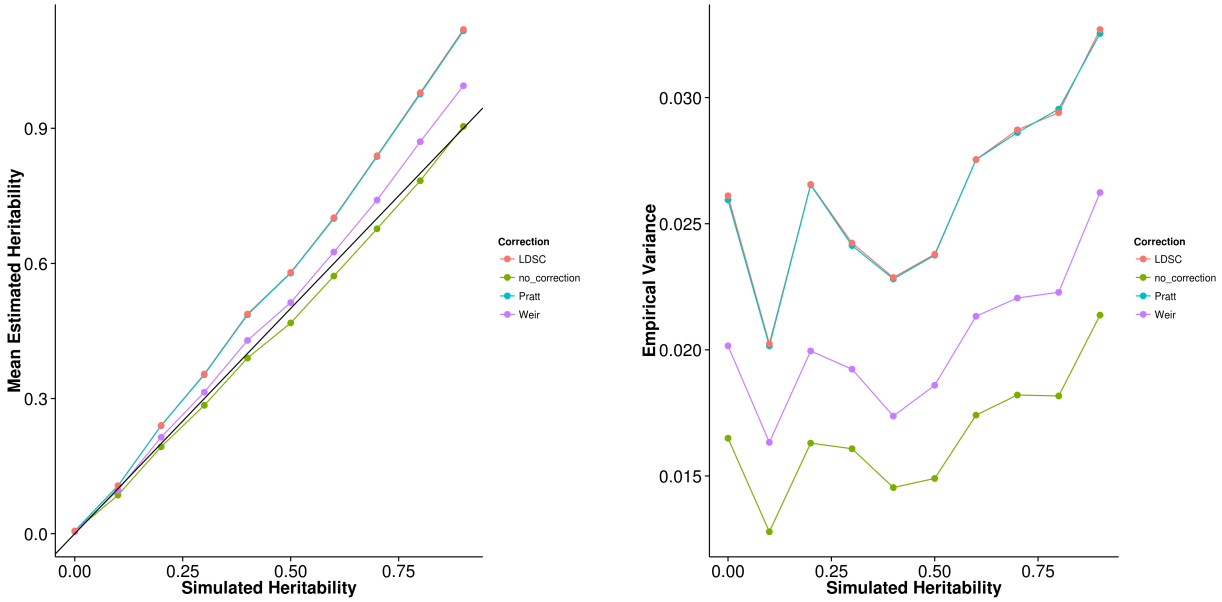
Intriguingly the same overestimation was not observed in LDSC. When inspecting the algorithm of LDSC, it was observed that LDSC also correct for the sampling bias in R^2 using:

$$\text{LDSC} : \tilde{R^2} = \hat{R^2} - \frac{1 - \hat{R^2}}{n - 2} \tag{2.43}$$

which was not tested in our previous LD correction simulation.

An interesting analysis will be to test the performance of the LD correction algorithm when the number of SNPs is higher (e.g 50,000 SNPs on chromosome 1) and whether if eq. (2.43) produce a better results. We therefore repeated the LD correction simulation by increasing the number of simulated SNPs to 50,000 on chromosome 1. To reduce the run time of the simulation, we only compared the performance of SHREK when eq. (2.43), eq. (2.39) and eq. (2.37) were used for the LD correction.

From the results (fig. 2.16), it was clear that all LD correction algorithms inflates the heritability estimation from SHREK in oppose to the underestimation observed when no LD correction was performed. The underestimation was as expected because the positive sampling bias in R^2 will lead to an "over correction"



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

Figure 2.16: Effect of LD correction to Heritability Estimation when 50,000 SNPs were simulated. As an overestimation was observed in the quantitative trait simulation, we performed a short simulation to assess the impact of LD correction to the heritability of SHREK when the there is a larger number of SNPs. From the graph, it was observed that all LD correction algorithms inflate the heritability estimation when large number of SNPs were simulated. In fact, the bias was the smallest when no LD correction was performed.

of the collinearity, thus result in lower estimates. As mentioned, it is possible that eq. (2.39) does introduce small imprecision (e.g. overcorrection) to the LD matrix which accumulates as the number of SNPs increases, leading to overestimation of the heritability. Our simulation results do support that as one of the possible explanation. What was interesting though is that the MSE is the lowest when no LD correction was performed, suggesting that when the number of SNPs increases, these LD correction algorithms actually has a negative impact to the performance of SHREK.

It was noted that most LD correction algorithm assumes the correlation was calculated on normally distributed data. However, genomic data follows a binomial distribution, which might violates the assumption, leading to a biased

correction. This will be an important area for further research. Without a good bias correction algorithm, the estimates from SHREK will most likely be biased downward, especially when the reference panel is small. Meanwhile, we allow users the freedom to disable the LD correction in SHREK.

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

2.7.2 Simulation Results

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

Quantitative Trait Simulation

In the quantitative trait simulation, it was clear that for most situation, GCTA has the best performance. By using the genetic relationship matrix, the estimation from GCTA were more accurate when compared to LDSC and SHREK. However, when the sample genotypes are unavailable, it is not possible to calculate the ge-

netic relationship matrix required by GCTA. Thus one can only rely on LDSC and SHREK.

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

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

It will be interesting to assess the performance of these algorithms when there is confounding effects such that one can test the importance of the intercept estimation function in the correction of confounding effects. However, the simulation of population and, especially cryptic relationship, is nontrivial. For example, although one can provide haplotype from different population to HAPGEN2, there is a lot of uncertainties in the simulation of the individual phenotypes: Should one standardize the genotype of the two population independently in the calculation of phenotype? Should the two population have the same causal SNPs? If not, should we limit the causal SNPs within the same biological pathway / function?

Moreover, heritability is dependent on the environment and genotype frequency. Theoretically, it is possible for different population to have a different heritability for a particular trait. The possible combinations and the complexity of

the problem is beyond the scope of this thesis but we do acknowledge that it is an important subject and further research is required.

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

Case Control Simulation

More often than not, researchers are interested in case control studies where "affected" and "normal" samples were compared. This is particular useful for the studies of disease traits such as schizophrenia. However, the heritability estimation is not as straight forward and requires the adaptation of the liability threshold model. It was known that GCTA, the most widely adopted algorithm for heritability estimation in GWAS was unable to provide accurate estimates in case control scenarios and its estimates are affected by the population prevalence and sample size of the studies (Golan, Eric S Lander, and Rosset, 2014). Our simulation results agree with the observation of Golan, Eric S Lander, and Rosset (2014), suggesting that as the population prevalence decreases, the magnitude of bias in the estimates of GCTA increases.

According to Golan, Eric S Lander, and Rosset (2014), in case control

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

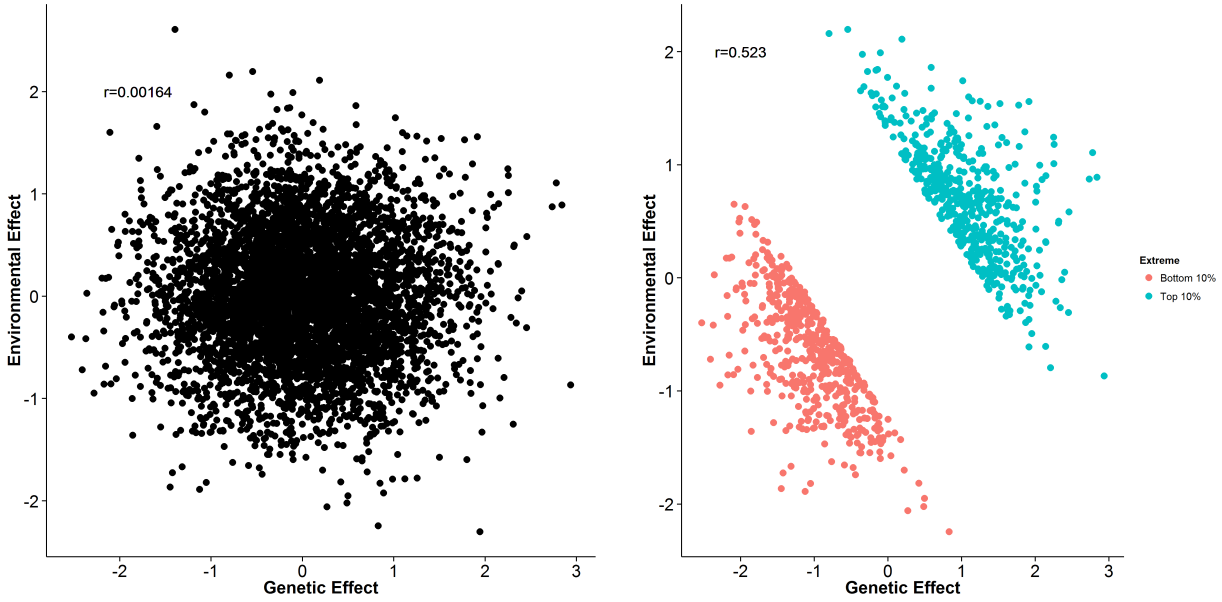
The question then is whether if this artificial correlation will affect the performance of SHREK and LDSC. First, it was observed that as the population prevalence decreases, the magnitude of bias for both LDSC with fixed intercept and SHREK increases suggesting that the population prevalence and the sampling bias might indeed be influential to the estimates of LDSC and GCTA. However, the direction of bias was opposed to what was observed in GCTA where a smaller population prevalence leads to a larger *overestimation* in the heritability. Considering that for SHREK, we adjust the estimates by multiplying eq. (2.23) to the estimates. An overestimation might suggest that we have an under correction of the bias. Of course the bias introduced by the LD correction is another factor to be considered, but considering that only 5,000 SNPs were simulated, the bias introduced by LD sampling bias should be relatively small as suggested by our LD correction simulation. To understand the effect of LD correction in case control scenario, we will need to increase the number of SNPs simulated yet that is only possible when additional computation resources are made available.

What was most surprising in the case control simulation is the performance of LDSC with intercept estimation. As we did not simulate any confounding factors, we would expect the performance of LDSC with intercept estimation would be worst compared to LDSC with fixed intercept because of the unnecessary additional degree of freedom in the estimation. However, it was observed that unlike SHREK and

LDSC with fixed intercept, the bias of LDSC with intercept estimation was robust to the change in population prevalence (figs. 2.10, 2.18, 2.21 and 2.24), thus when the population prevalence is small, the bias of LDSC with intercept estimation is relatively smaller when compared to LDSC with fixed intercept.

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

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



(a) No Extreme sampling

(b) Extreme Sampling

Figure 2.17: Effect of extreme sampling design. Here we simulated the genetic and environmental effect independently. When no extreme sampling was performed, there is no correlation between the environmental effect and genetic effect as expected. However, when extreme sampling was performed, an artificial correlation was observed. This might be the main reason why the estimates from GCTA are downward biased.

Extreme Phenotype Sampling

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

Herein, we simulated the situation where an extreme selection design was

performed to assess the performance of the heritability estimation algorithms. We are also interested in comparing the performance between extreme phenotype sampling and random sampling strategy. First, it was observed that when extreme phenotype sampling was performed, the estimates from GCTA were biased downward. This observation was similar to what was observed in the case control simulation. It was noted that although we were simulating independent environmental and genetic effects, the extreme phenotype sampling strategy does introduced an artificial correlation between the two effects, similar to what was observed in case control scenario (fig. 2.17). This might therefore affect the performance of GCTA where as the portion of sample selected decreases, the magnitude of bias increases, similar to the change of population prevalence in case control studies.

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

However, although the empirical variance observed in the random sampling for all the algorithm was the same as what was observed in the quantitative trait simulation with 100 causal SNPs, the estimated variance for GCTA and LDSC was much worst. A larger upward bias was observed in the estimates from LDSC with fixed intercept, suggesting there might be some difference between the simulation of random sampling and the simulation of quantitative trait, even though most of

the parameter for simulation are the same. The only difference in the two simulation was the standardization of genotype when calculating the phenotype. For the quantitative trait simulation, the genotype was standardized based on the genotype of 1,000 samples of which all were included in the analysis. However, in the simulation of the random sampling design, 5,000 samples were used to standardize the genotype, of which only 1,000 out of 5,000 were included in the final analysis. It was uncertain how this affects the performance of the algorithm and further analysis might be required.

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

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

2.7.3 Application to Real Data

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

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

$$\mathrm{E}[\chi^2|l_j] = Nl_j\frac{h^2}{M} + Na + 1 \tag{2.44}$$

where l_j is the LD score of variant j, N is the sample size, a is the contribution of confounding biases, h^2 is the heritability and M is the number of SNPs. When contact the author about the discrepancy of the estimation between our run of LDSC and the estimates shown in the supplementary table, B. Bulik-Sullivan (2015) replied that the estimated from the supplementary table define M as the total number of SNPs in the reference panel used to estimated LD score whereas the current version of LDSC defines M as the number of SNPs with maf$>$ 5% in the reference panel used to estimate LD score which they deem more appropriate based on new data they observed after their original paper was published. Based on the caption of their supplementary, they stated that "... if the average rare SNP explains less phenotypic variance than the average common SNP, then a smllaer value of M would be more appropriate, and the estimates in the supplementary table will be biased upwards." This explain the smaller estimates from our run.

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

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

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

it is likely that the difference in table 2.5 is a direct result of the estimation of the intercept.

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

Based on our estimation, the PGC schizophrenia GWAS can at most account for $\sim 20\%$ of the heritability of SCZ despite the amount of samples included. When compared to the heritability estimated from twin studies, there are around $40\% \sim 60\%$ of missing heritability unaccounted for. This suggested that rare vari-

ants or other factors (e.g. CNV) other than common SNPs can account for the remaining heritability of schizophrenia.

If one would like to estimates the contribution of rare variants to SCZ, new algorithms might be required. This is mainly because the LD estimates form the rare variants usually have a large variability and might not be reliable, thus leading to unreliable estimates from SHREK and LDSC. Thus special care are required if one would like to include the rare variants in the estimation process. Also, it was noted that we only performed the estimation on the autosomal chromosomes. The main reason behind was that there are large difference between male and female on the sex chromosomes (e.g. 2 X for female and XY for male). The proportion of male and female were usually not given. This leads to difficulties in eq. (2.14) where the sample size is an important factors. Special consideration might therefore be required if one would like to estimates the contribution of variants on the sex chromosome to schizophrenia.

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

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

2.7.4 Limitations and Improvements

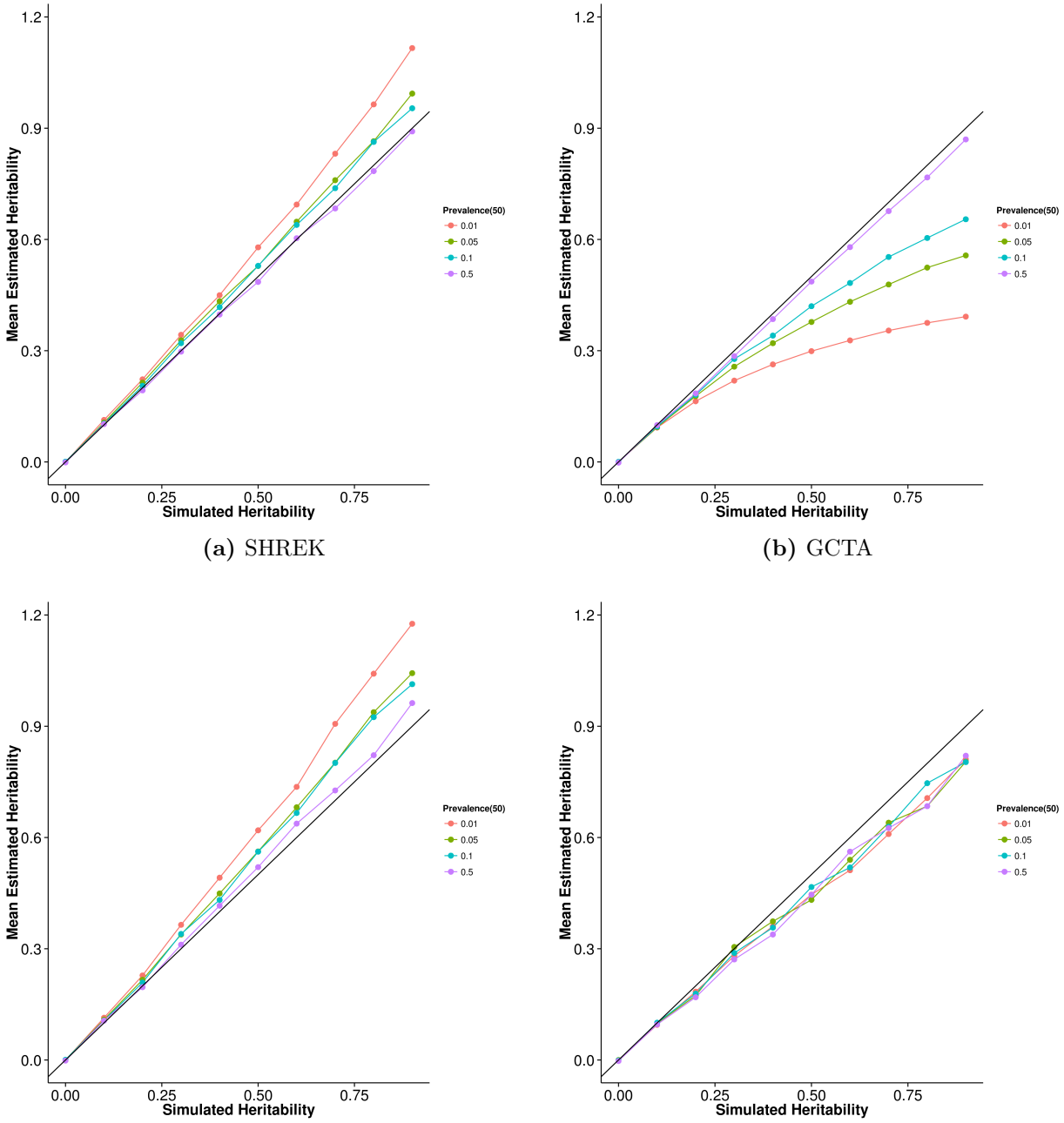
One of the biggest disadvantage of SHREK is its speed when compared to LDSC. To estimate the heritability, SHREK requires the calculation of the inverse of the LD matrix, which is an $O(n^3)$ operation. Although the use of sliding window has significantly reduce the time requirement, the run time will still increase substantially as the density of the SNPs increases. For example, it can take more than 2 days to process one chromosome of the PGC schizophrenia data set, where there can be more than 5,000 SNPs per window. When applying SHREK to the real data, the computation resources required to estimates heritability of the PGC SCZ GWAS was too high, forcing us to reduce the window size for the analysis. To make the use of SHREK feasible, further development might be required to improve the speed of SHREK. An obvious choice might be to use the Armadillo library (Sanderson, 2010) together with the OpenBLAS library which can be more than 3 times faster when compared to the EIGEN C++ library (Ho, 2011).

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

Finally, we do acknowledge that we have not exhaust all possible combinations of genetic architectures in our simulation. For example, one can also test the performance of the algorithms when the observed prevalence was different (e.g.

not 50%). It is also possible for one to investigate the effect of number of causal SNPs on the performance of the algorithms when extreme phenotype sampling was performed. However, we do argue that we have performed a substantial amount of simulations and should be able to provide a general concept as to how the performance of SHREK, LDSC and GCTA are affected in the general scenarios.

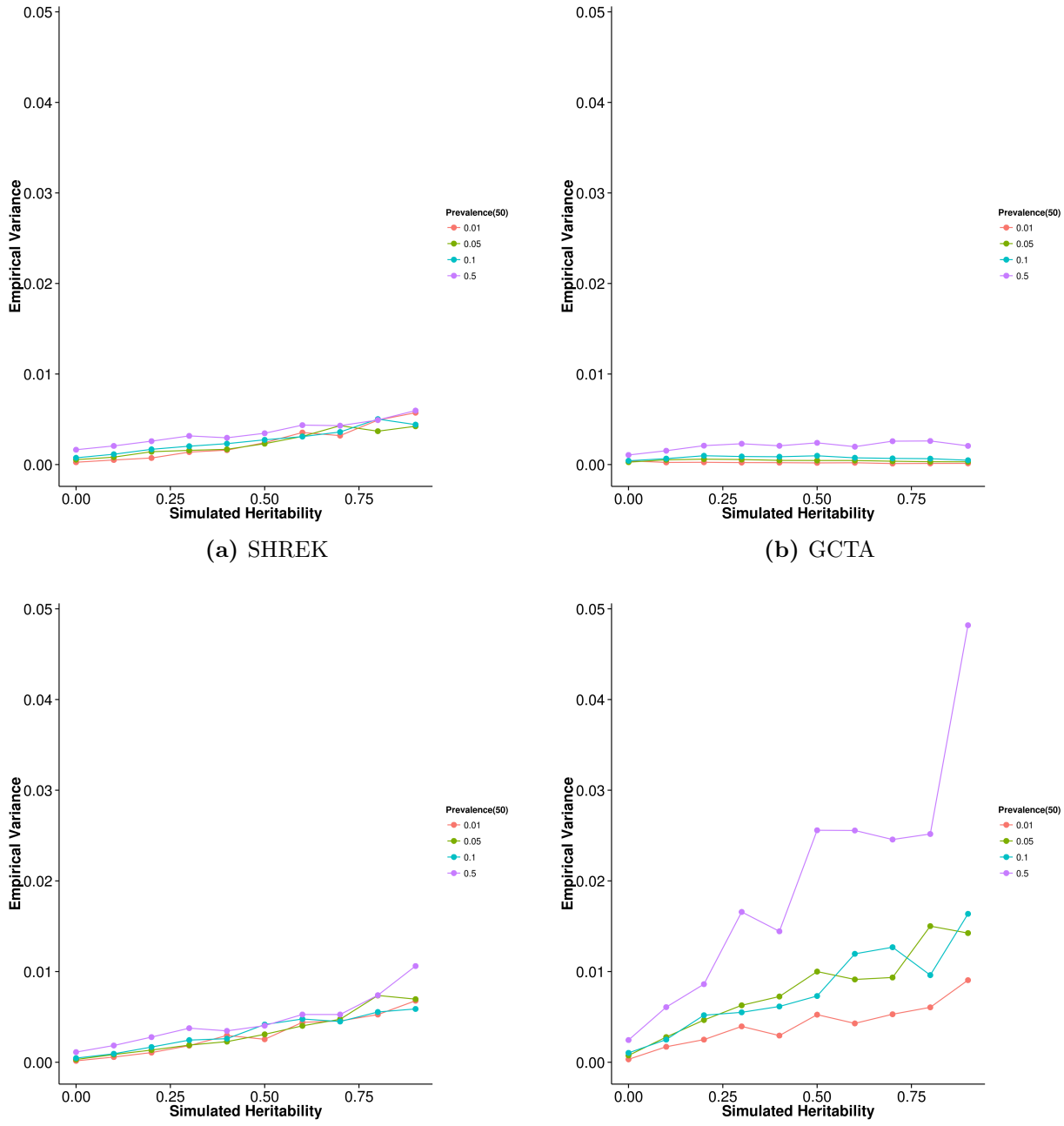
2.8 Supplementary



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

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

Figure 2.18: 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.



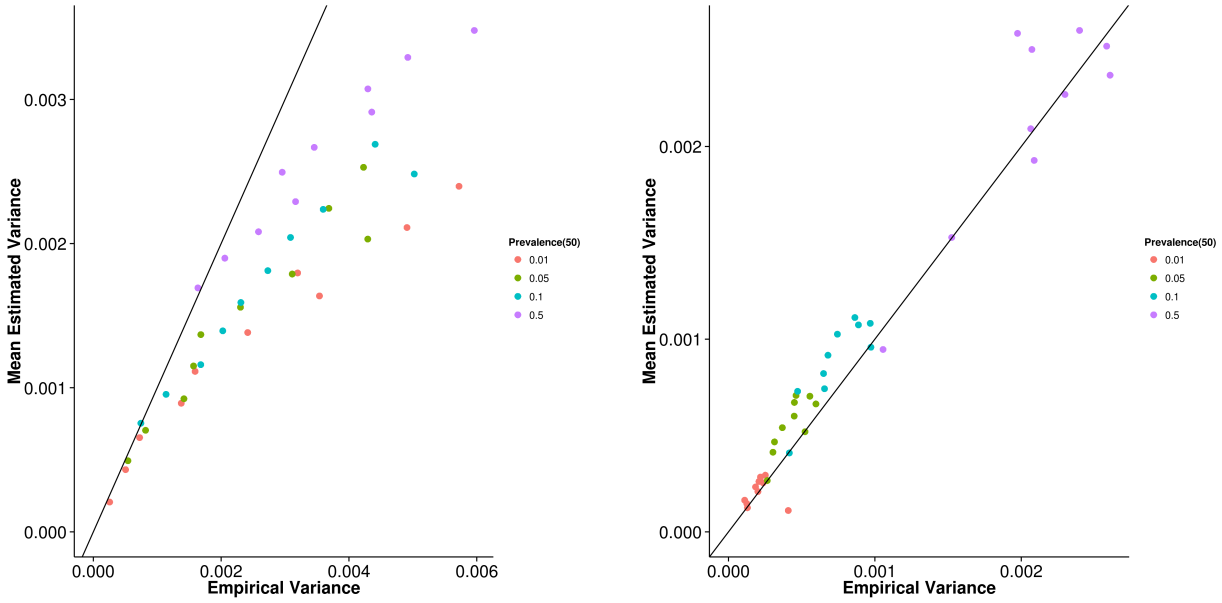
(a) SHREK

(b) GCTA

(c) LDSC with fix intercept

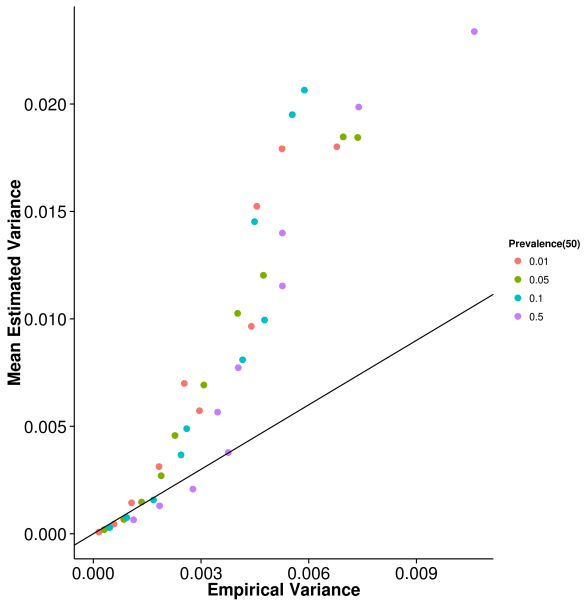
(d) LDSC with intercept estimation

Figure 2.19: 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.

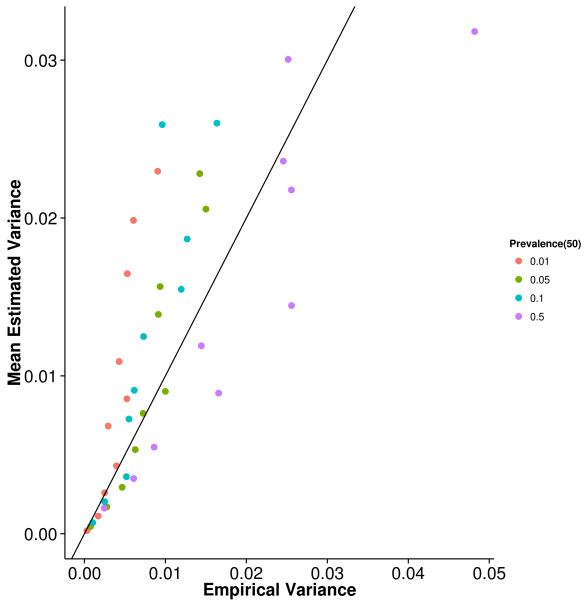


(a) SHREK

(b) GCTA

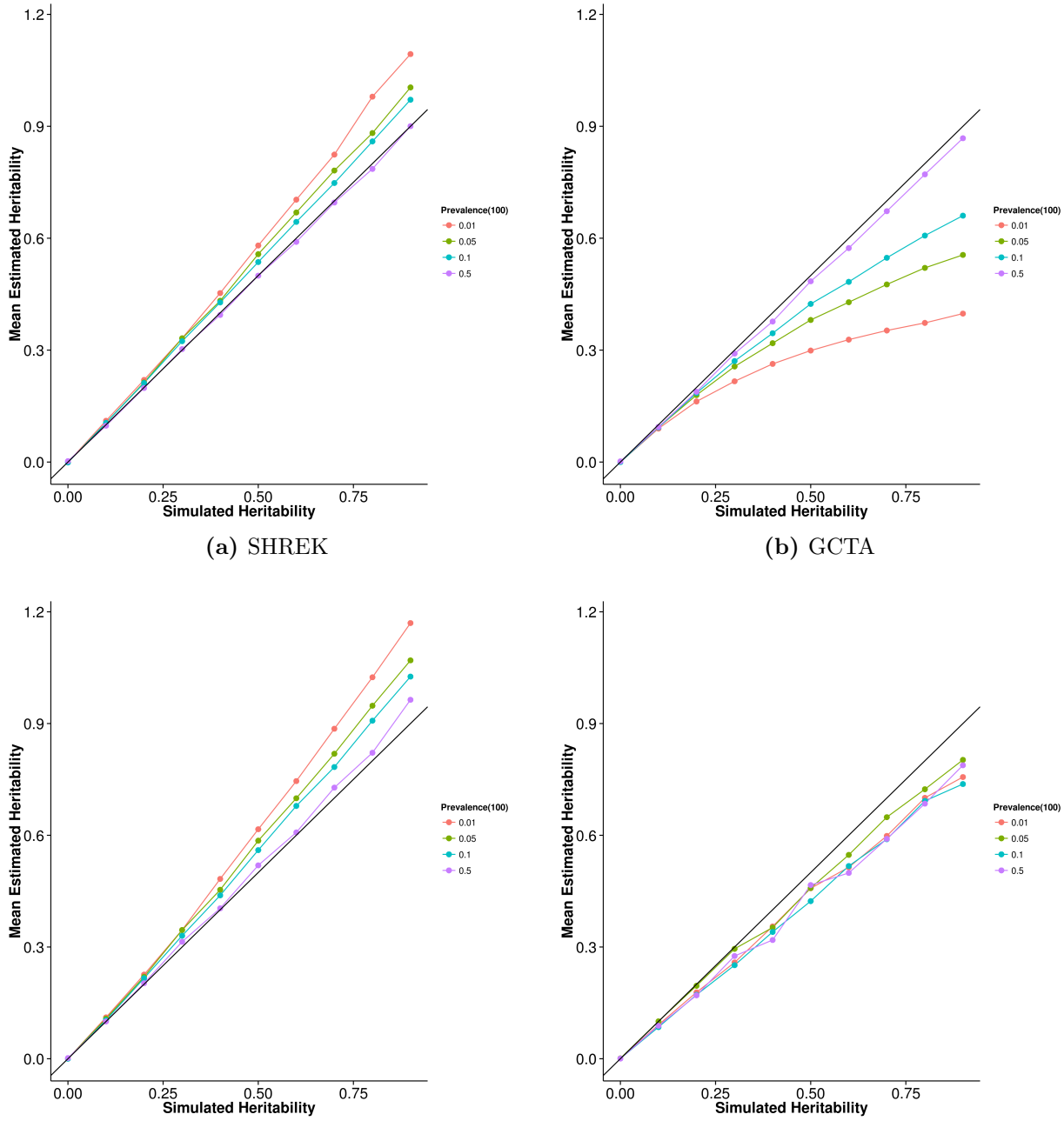


(c) LDSC with fix intercept



(d) LDSC with intercept estimation

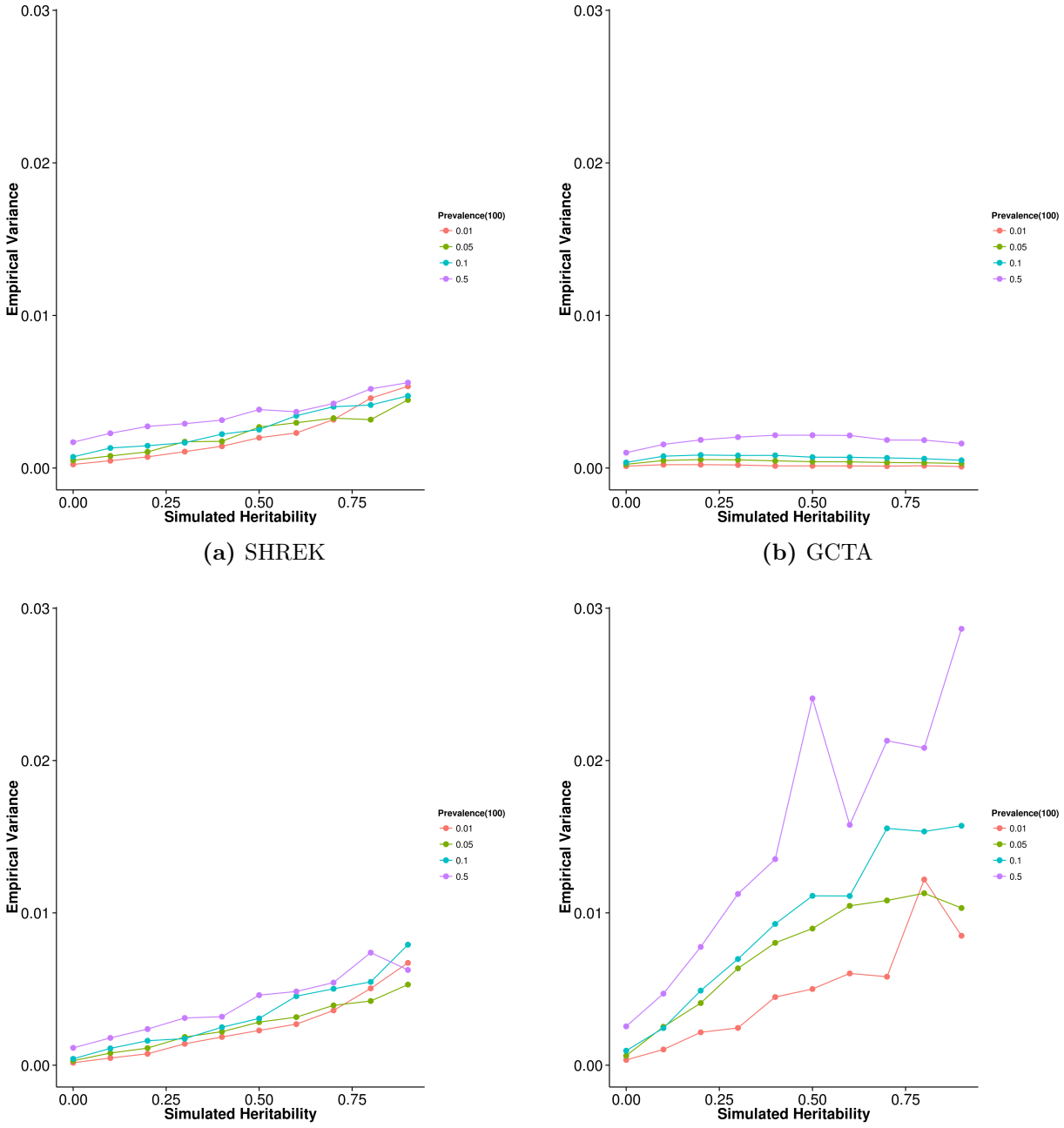
Figure 2.20: Estimated variance of results from case control simulation with random effect size simulation when compared to empirical variance when 50 causal SNPs was simulated. Again, the estimation of variance from SHREK tends to be downwardly biased and LDSC with fixed intercept tends to be upwardly 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 2.21: 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.



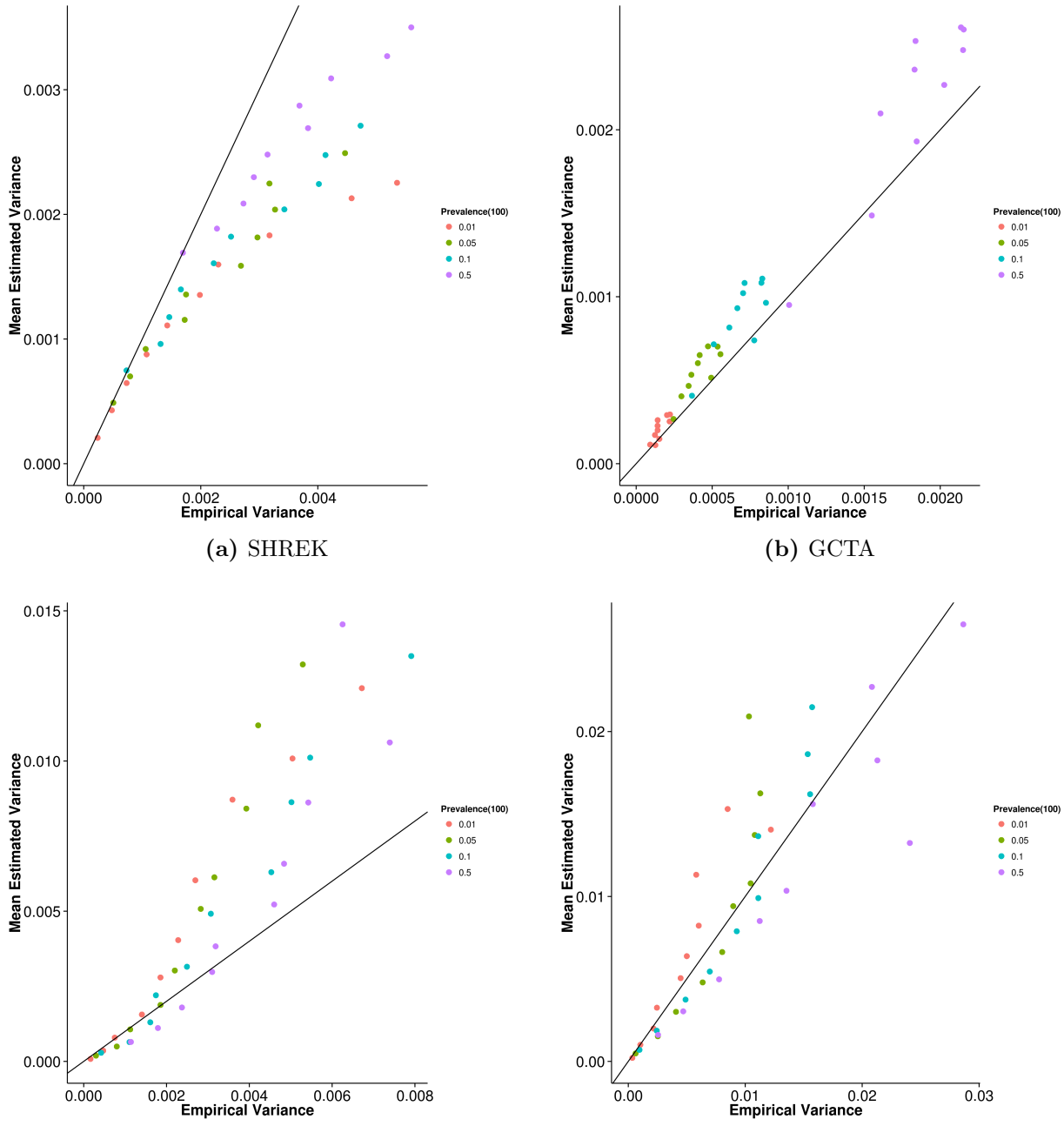
(a) SHREK

(b) GCTA

(c) LDSC with fix intercept

(d) LDSC with intercept estimation

Figure 2.22: 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.



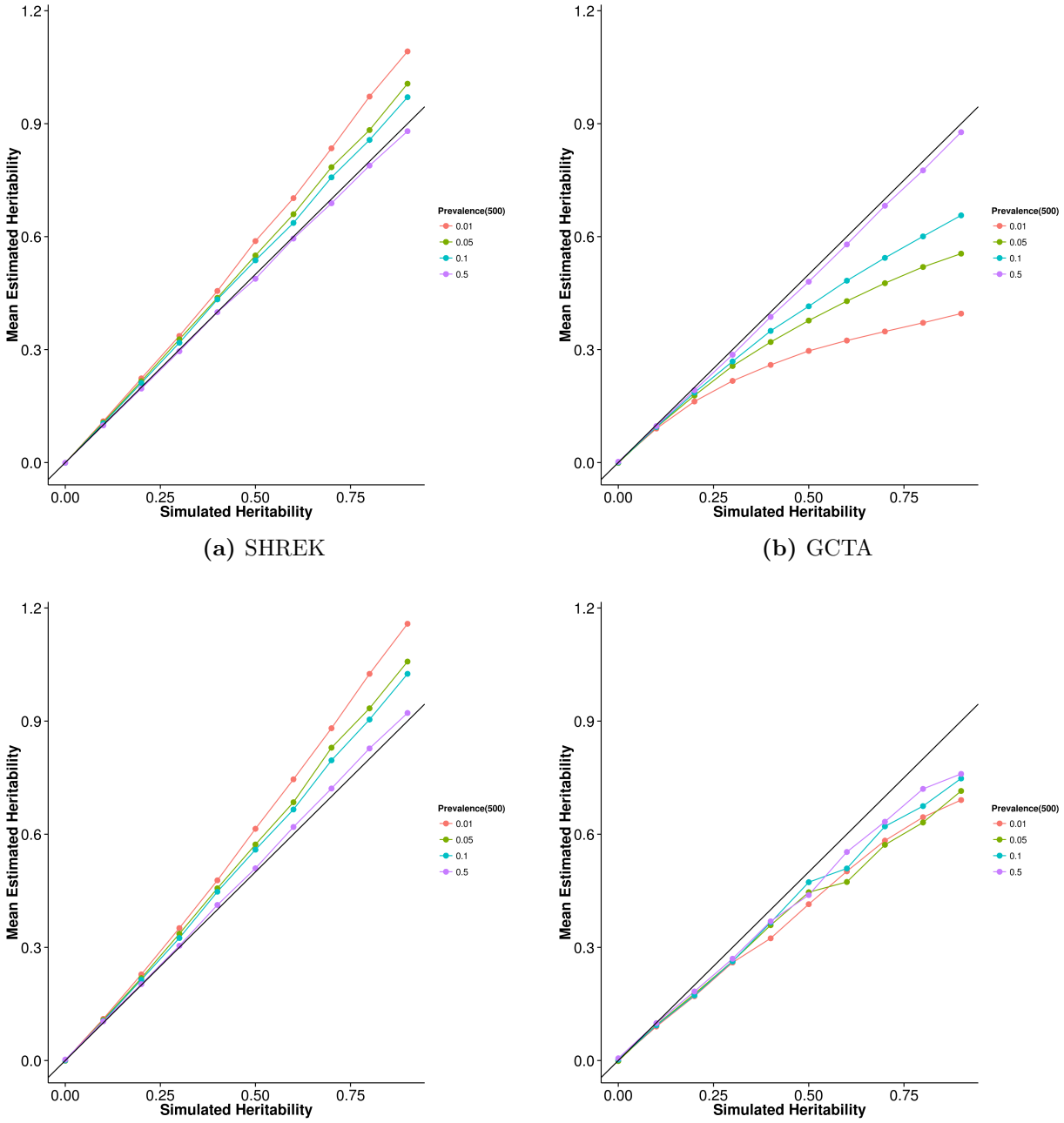
(a) SHREK

(b) GCTA

(c) LDSC with fix intercept

(d) LDSC with intercept estimation

Figure 2.23: 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.



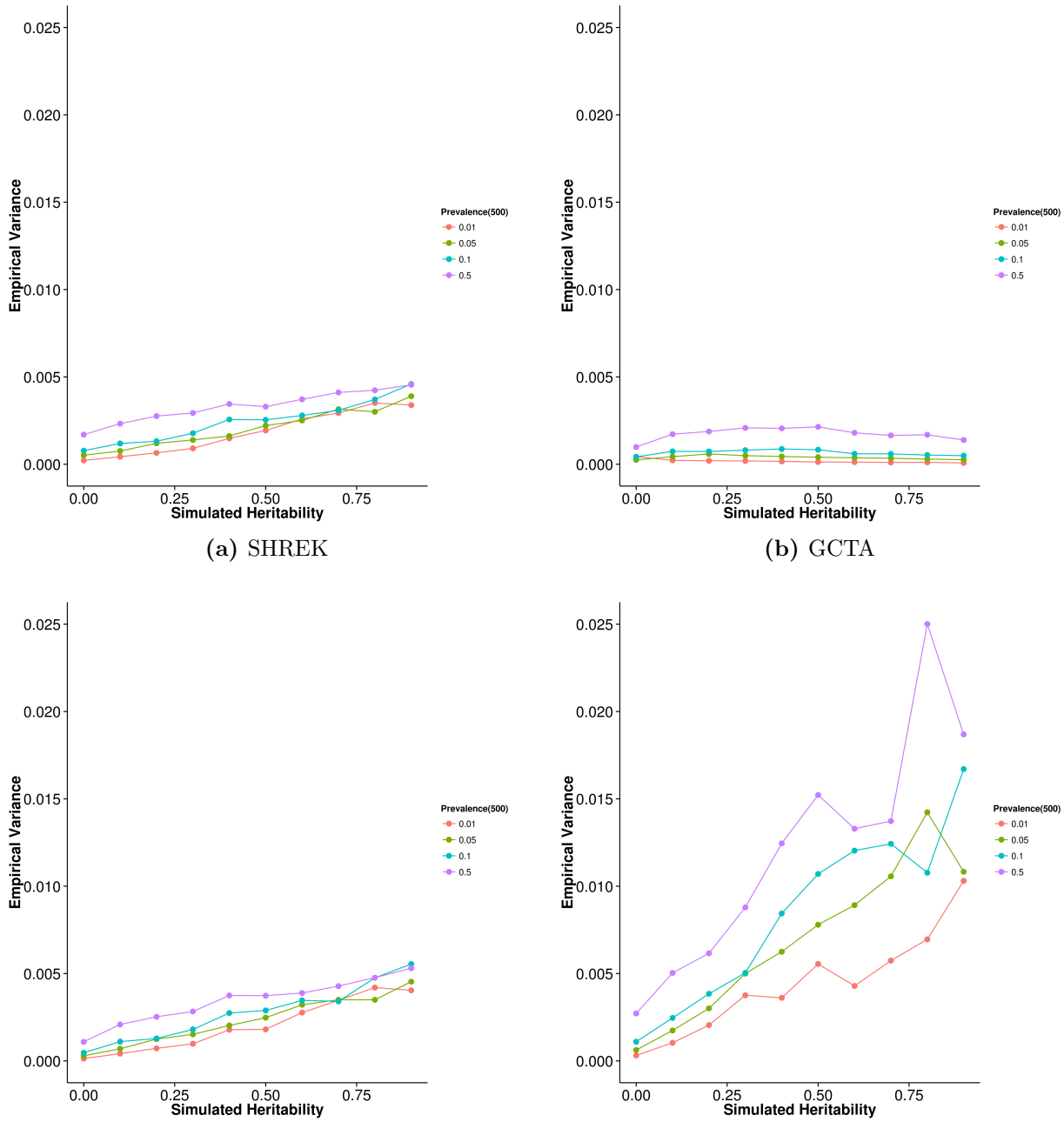
(a) SHREK

(b) GCTA

(c) LDSC with fix intercept

(d) LDSC with intercept estimation

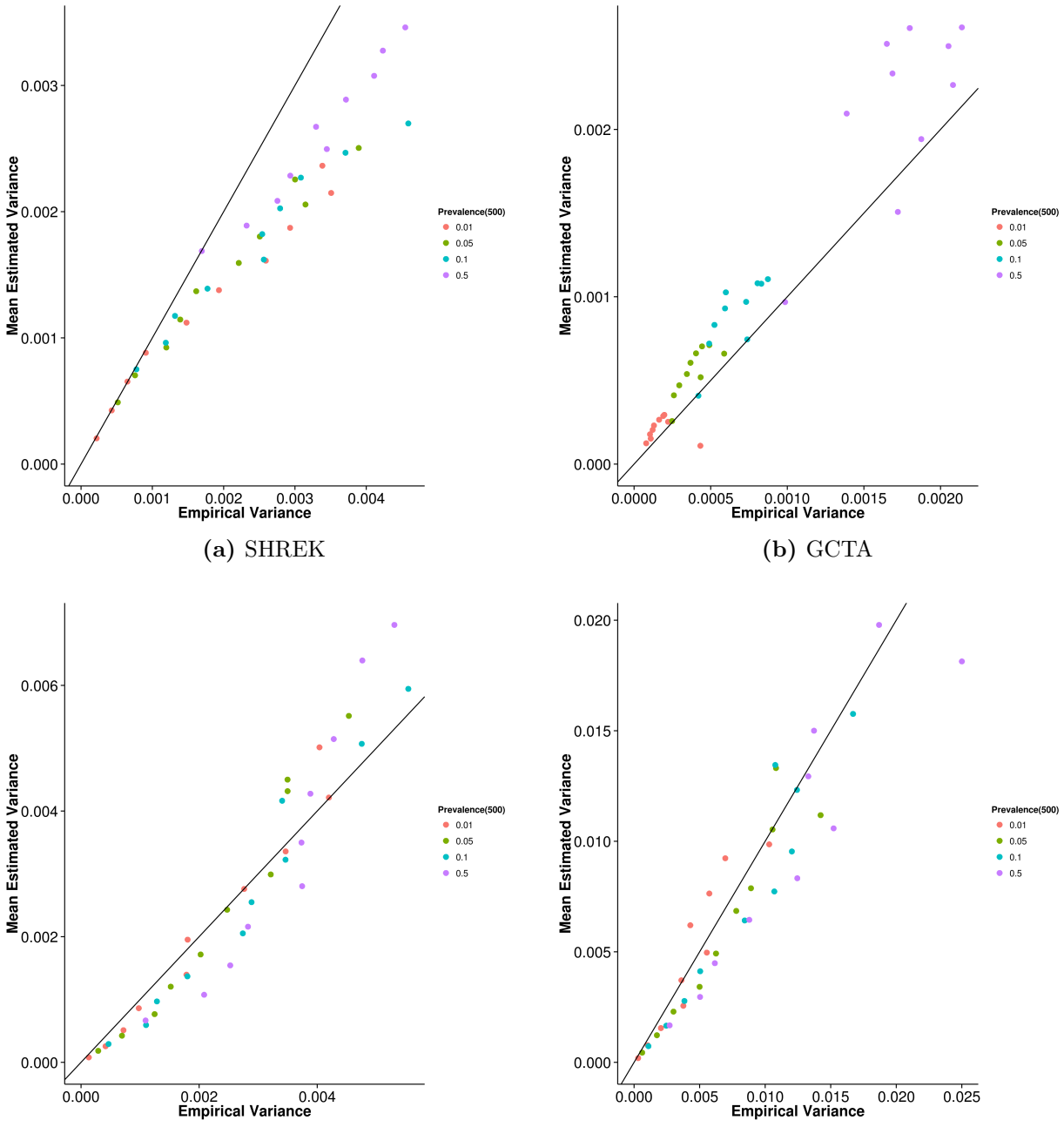
Figure 2.24: Mean of results from case control simulation with random effect size simulation with 500 causal SNPs. Again, a clear pattern of underestimation was observed for GCTA and LDSC with intercept estimation whereas estimations from SHREK and LDSC with fixed intercepts tends to be upwardly biased, 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 2.25: 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 2.26: 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 was much smaller than expected, accounting for only 20% of the variance in schizophrenia. This suggest that other factors such as rare variants and epigenetic factors might have contributed to the heritability of schizophrenia. Another possibility will be the gene environmental interaction ($G \times E$).

Previous studies have suggested there might be interaction between prenatal infection and genetic variations in the development of schizophrenia (Tienari et al., 2004; Clarke et al., 2009). Evidences now suggest that the effect of prenatal infection was mainly mediated by maternal immune response instead of the specific infection (A S Brown and Derkits, 2010) therefore it is likely that the perturbation induced by maternal immune activation (MIA) are interacting with genetic variations in the development of schizophrenia. With the development of LDSC, we may now perform the partitioning of SNP heritability using summary statistics from GWAS. This allow one to investigate whether if a particular functional pathway perturbed by early MIA contributes disproportionately to the heritability of schizophrenia.

On the other hand, one of the main goal in schizophrenia research is to identify effective treatments for schizophrenia such that the quality of life of schizophrenic patients can be improved. Based on the MIA model, one possible candidate might be the n-3 polyunsaturated fatty acid (PUFA) rich diet. It has been suggested that n-3 PUFA can inhibits the production of IL-6 (Trebble et al., 2003), which is a major mediator in MIA (Smith et al., 2007). Moreover, n-3 PUFA also plays a critical role in the development of central nervous system (Clandinin, 1999) and it has robust anti-inflammatory properties (Trebble et al., 2003). Therefore it is possible that a n-3 PUFA rich diet can help to alleviate the symptoms of schizophrenia. Indeed, previous study from our lab suggested that an n-3 PUFA rich diet can help to reduce the schizophrenia-like phenotype in mice exposed to early MIA insults (Q. Li, Leung, et al., 2015).

Herein, we introduce a hypothesis generation study aiming 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. We would also like to investigate whether if functional pathways perturbed by MIA or changed by diet contributes more to the heritability of schizophrenia using LDSC.

In this study we selected the cerebellum as the target tissue for our experiment. Although hippocampus (Velakoulis et al., 2006; Nugent et al., 2007) and prefrontal cortex (Knable and Weinberger, 1997; Perlstein et al., 2001) were the two most studied region in schizophrenia, the cerebellum has also been reported to be related to schizophrenia (Yeganeh-Doost et al., 2011; Andreasen and Pierson, 2008). Moreover, the cerebellum plays a central role in the cortico-cerebellar-thalamic-cortical neuronal circuit which is important to schizophrenia. Positron emission tomography (PET) studies have shown that a dysfunction in this circuit can contribute to "cognitive dysmetria", e.g. impaired cognition and other symptoms of

schizophrenia (Yeganeh-Doost et al., 2011). Altogether, this makes the cerebellum an interesting target to investigate.

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} PolyI:C in an injection volume 5ml kg^{-1}, prepared on the day of injection was administered to pregnant mice on 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.

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

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 1M_2 and 1M_3 which came from the same litter but were given a different diet.

Stranded mRNA-Seq Kit. All samples were sequenced using Illumina HiSeq 1500 at 2 lanes (2×101 base pair (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 were rather standardized where FastQC (Andrews, n.d.) is the most widely adopted tools. It can generate the required per base QC and provide a general picture of how well the sequencing were done.

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), we trim the adapter sequences from the sequence reads and

only retain reads that were at least 75 bp long 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 of all the aligners investigated taking into account of accuracy and speed. 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 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 author of DESeq2 were very active in providing supports for the package, we selected DESeq2 (version 2.1.4.5) (Love, Wolfgang Huber, and Simon Anders, 2014) as the statistic package for the differential gene expression analysis.

Perhaps one of the most controversial study in RNA Sequencing was the mouse ENCODE paper by Yue et al. (2014) where Gilad and Mizrahi-Man (2015) demonstrated that most of the findings from Yue et al. (2014) was confounded by lane and batch effect. 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 differs across different batches, we were unable to fully balance out the batch effect. Therefore, in our analysis, we must control for the batch effect.

In our study, we were interested in the following comparisons:

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

To obtain the desire comparison, and also control for batch effect, we used $\sim Batch + Condition + Diet + Condition : Diet$ as our model of statistical analysis where Condition is the MIA exposure status. We did not incorporate the RIN into our statistic model because it was advised against by the author.

We would also like to see if the batch effect can leads to false positive results. Therefore we performed the likelihood ratio test (LRT). 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, we removed all genes with base mean count < 10 to reduce the noise associated with low expression and the Benjamini and Hochberg method were then used to correct for multiple testing.

3.2.6 Functional Annotation

One of the most important aim of the current study is to investigate whether if functional pathways perturbed by MIA or affected by diet contribute to larger amount of SNP heritability. It is therefore important to perform functional annotation of the differentially expressed genes (DEGs) in order to identify pathways that were affected (e.g. enriched by the DEGs).

Because of the limited number of DEG identified, we performed the Wilcoxon Rank Sum test to test whether if genes within the pathway are more significant the genes outside pathway. The canonical pathways annotations obtained from the Molecular Signatures Database (MSigDB) (v5.0 updated April 2015) (Subramanian et al., 2005) were used for our analysis. To avoid testing overly narrow or broad functional pathways, pathways with more than 300 genes or less than 10 genes were removed from our analysis. Pathways with adjusted p-value < 0.05 (using Benjamini and Hochberg adjustment) were considered as significant.

3.2.7 Partitioning of Heritability

We then tried to perform partitioning of heritability using the significant pathways as our annotation. Two "super pathways" were also generated which included all the genes in the pathways perturbed in MIA or genes in pathways affected by diet.

SNPs from the 1000 genome was first associated with genes using SnpEFF (Cingolani et al., 2012) with the GRCh37.75 annotation, where SNPs within $\pm$5kb region of a gene is considered to be associated. The annotation file required by LDSC was then generated by identifying SNPs participated in the significant pathways from section 3.2.6. If a SNP was found to be associated with more than 1 genes, then it was considered to be within all pathways where its associated genes were part of.

Finally, we performed the partitioning of heritability using LDSC (B. K. Bulik-Sullivan et al., 2015) --annot and --overlap-annot options using 1000kb window size and the reference panel generated from section 2.5. Pathways with positive proportion of heritability explained and an enrichment p-value > 0.05 were considered significant.

3.2.8 Designing the Replication Study

Another important goal of our current study is to provide information for further replication studies. In order to estimate the power and required samples for the replication studies, we performed the power estimation using Scotty (Busby et al., 2013). We provided the count data from our pilot samples to Scotty to estimate the minimal required samples for our replication study if we would like to detect at least 90% of the genes that are differentially expressed by a 2× fold change at p< 0.01 and that at least 80% of genes has at least 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. A quality score at 30 represents the probability of having an incorrect base call is less than 1 in 1,000. After removing the adapter sequences from the reads, more than 97% of the reads remains. Over 90% of the trimmed reads could be uniquely mapped to the *Mus musculus* reference genome (mm10, Ensembl GRCm38.82) using the STAR aligner (version 2.5.0a) (Dobin et al., 2013). To obtained the expression count, we used the featureCounts (version 1.5.0) (Liao, Gordon K Smyth, and Shi, 2014) to generate the count matrix required

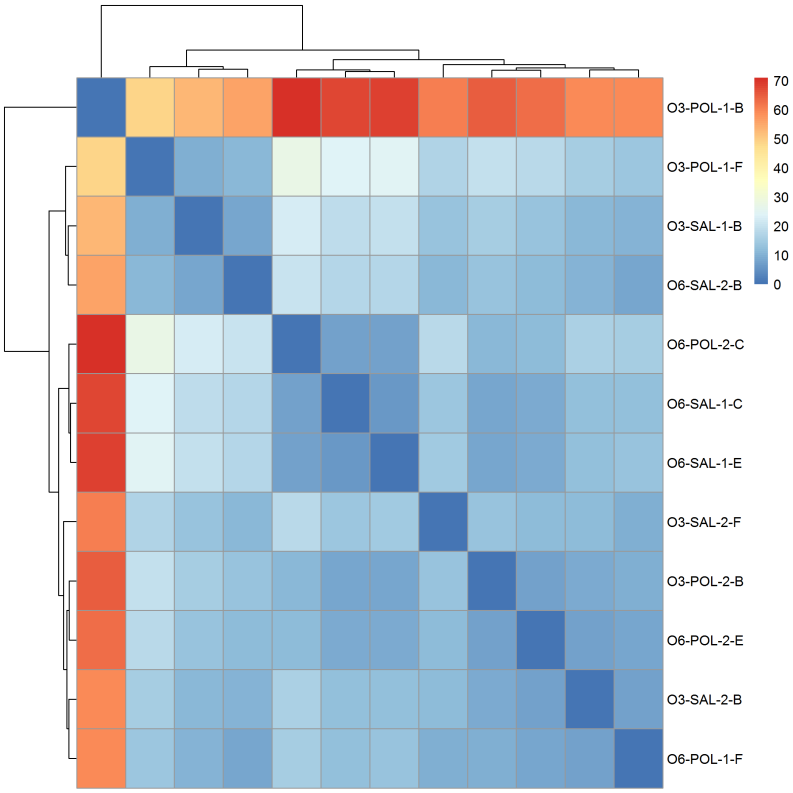


Figure 3.1: Sample Clustering results. It was observed that there was no clear clustering for lane or batch effects. However, one sample from the n3-PUFA-PolyI:C group was found to be substantially different from all other samples.
It was unclear whether if the difference was due to sample contaminations or was due to sample mis-label. To avoid problems in down-stream analysis, we excluded this sample from subsequent analyses

for downstream analysis.

Next, we were interested in whether if there are any contamination of samples or series confounding effect of bath or lane. We therefore performed an unsupervised clustering on the sample count data. It was observed that none of the samples were clustered by lane or batch, suggesting that there were no serious batch or lane effect presented in our samples. However, one sample from the n3-PolyI:C group was found to be substantially different from all other samples (fig. 3.1). It was unclear whether if the difference was due to sample contaminations (from other source) or was due to sample mis-label. To avoid problems in down-stream analysis, we excluded this sample from subsequent analyses

3.3.2 Differential Expression Analysis

After excluding the problematic samples, we performed the DESeq2 analysis. Of the 16,747 genes that passed through quality control, only one gene, *Sgk1* (p-adjusted=0.00186) was found to be significantly differentiated when we examines the effect of n-3 PUFA rich diet on the gene expression in the cerebellum of PolyI:C exposed mice (fig. 3.2c). No genes were found to be significant for the other two comparisons (figs. 3.2a and 3.2b).

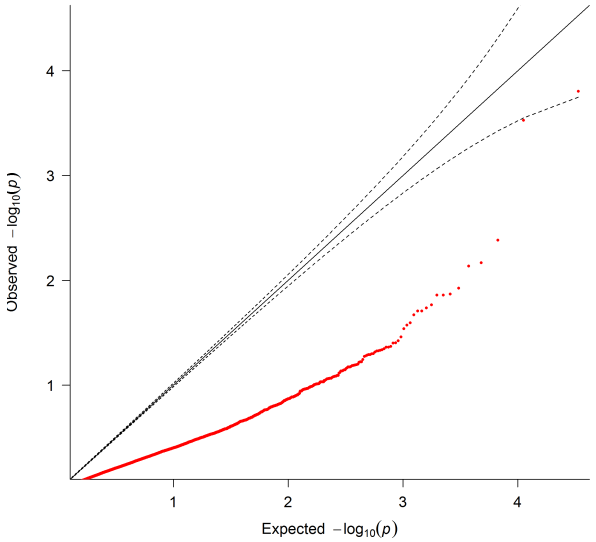
We also performed the LRT to compare test the effect of batch on our analysis. A total of 178 genes were found to be significant differentiated (fig. 3.2d), suggesting that the "Batch" is indeed an important factor to consider in our analysis.

3.3.3 Functional Annotation

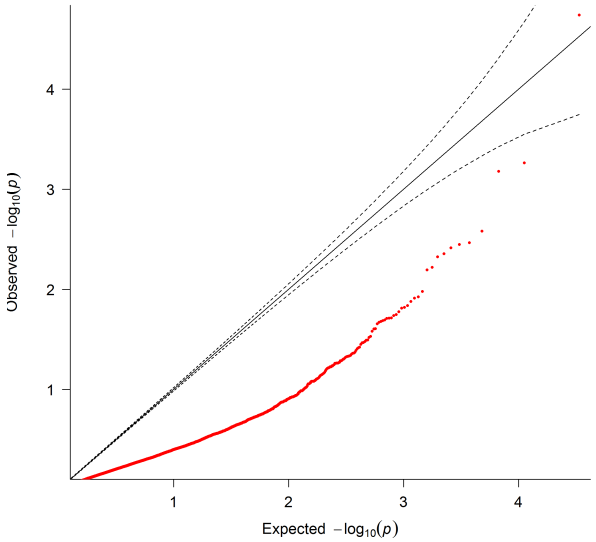
It is common practice to try and perform functional annotation to the DEGs. However, in most of our analysis, there were either no DEG or only 1 DEG, making it difficult to perform functional annotation. We therefore used the Wilcox rank sum test to analysis whether if a pathway contain genes that are more significant than genes not within the pathway instead of trying to identify pathways that were *enriched* by DEGs.

None of the pathway were found to be significant when comparing the effect of the n-3 rich diet in Saline exposed mice. On the contrary, 17 pathways were found significant when comparing the effect of n-3 PUFA rich diet in PolyI:C exposed samples (table 3.2) where 4 pathways were related to growth factors such as fibroblast growth factor (FGF) or epidermal growth factor (EGF) and 4 others were related to kinases such as phosphatidylinositol 3-kinase (PI3K) or MAPK.

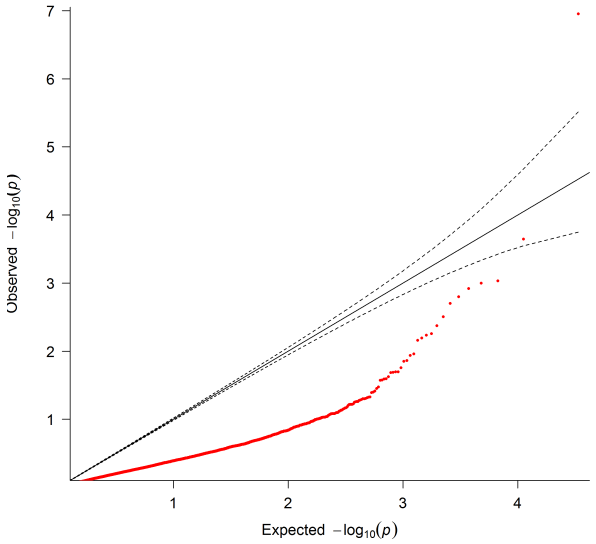
Finally, 12 pathways were found to significant when comparing Saline and



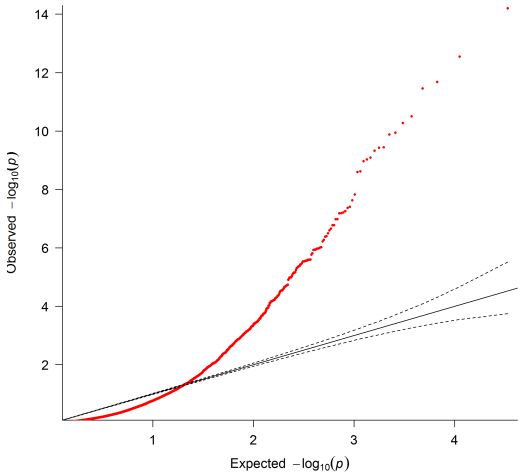
(a) O6-Saline mice vs O6-PolyI:C mice



(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 QQ plot, it was observed that most of the observed p-value was less than what would have been expected. This is likely due to the small sample size of our study which leads to an under powered association. The only exception was the analysis of batch effect were a large amount of genes were found to be significant. This demonstrate the importance of adjusting for batch effect

PolyI:C exposed mice given the n-6 PUFA rich diet (table 3.3) with pathways such as neuroactive ligand-receptor interaction (p-adj = 1.27×10^{-3}), calcium signaling pathway (p-adj = 2.79×10^{-3}) and genes involved in Neuronal System (p-adj=0.00153) among the significant pathways.

3.3.4 Partitioning of Heritability

Given the significant pathways, we performed the partitioning of heritability using LDSC (B. K. Bulik-Sullivan et al., 2015). In total, 29 unique pathways were included in the analysis were 26 of them were found to have non-negative contribution to the heritability of schizophrenia. 8 out of the 26 pathways were found to be significant including the MIA super pathway (table 3.4). Of the 8 significant pathways, only 2 of those (M3008 and M5884) were affected by diet and they were both also perturbed by MIA.

ID	Size	Source	Description	Adjusted P-Value
M508	78	REACTOME	Genes involved in Signaling by SCF-KIT	0.00671
M570	44	REACTOME	Genes involved in PI3K events in ERBB2 signaling	0.0242
M3008	196	NABA	Genes encoding structural ECM glycoproteins	0.0309
M1090	112	REACTOME	Genes involved in Signaling by FGFR	0.0309
M563	109	REACTOME	Genes involved in Signaling by EGFR in Cancer	0.0309
M17776	100	REACTOME	Genes involved in Downstream signaling of activated FGFR	0.0309
M1076	83	REACTOME	Genes involved in Amyloids	0.0309
M850	56	REACTOME	Genes involved in PI-3K cascade	0.0309
M10450	38	REACTOME	Genes involved in GAB1 signalosome	0.0309
M16227	24	REACTOME	Genes involved in Cholesterol biosynthesis	0.0309
M5872	17	KEGG	Steroid biosynthesis	0.0309
M16334	10	BIOCARTA	Eph Kinases and ephrins support platelet aggregation	0.0309
M5884	275	NABA	Ensemble of genes encoding core extracellular matrix including ECM glycoproteins, collagens and proteoglycans	0.0456
M635	127	REACTOME	Genes involved in Signaling by FGFR in disease	0.0456
M568	38	REACTOME	Genes involved in PI3K events in ERBB4 signaling	0.0456
M165	32	PID	Syndecan-4-mediated signaling events	0.0456
M1262	15	REACTOME	Genes involved in GRB2:SOS provides linkage to MAPK signaling for Intergrins	0.0456

Table 3.2: Significant Pathways when comparing effect of diet in PolyI:C exposed mice. The pathway IDs are the systematic name from MSigDB. Most of the significant pathways were related to the kinase such as PI3K and MAPK or growth factors such as FGF and EGF.

ID	Size	Source	Description	Adjusted P-Value
M13380	272	KEGG	Neuroactive ligand-receptor interaction	1.27×10^{-3}
M2890	178	KEGG	Calcium signaling pathway	2.79×10^{-3}
M12289	188	REACTOME	Genes involved in Peptide ligand-binding receptors	0.00118
M5884	275	NABA	Ensemble of genes encoding core extracellular matrix including ECM glycoproteins, collagens and proteoglycans	0.00119
M735	279	REACTOME	Genes involved in Neuronal System	0.00153
M15514	186	REACTOME	Genes involved in Transmission across Chemical Synapses	0.00401
M4904	121	REACTOME	Genes involved in G alpha (s) signalling events	0.0127
M3008	196	NABA	Genes encoding structural ECM glycoproteins	0.0131
M752	137	REACTOME	Genes involved in Neurotransmitter Receptor Binding And Downstream Transmission In The Postsynaptic Cell	0.0131
M10792	267	KEGG	MAPK signaling pathway	0.0195
M17	59	PID	Notch signaling pathway	0.0406
M18437	184	REACTOME	Genes involved in G alpha (q) signalling events	0.0406

Table 3.3: Significant Pathways When Comparing Effect PolyI:C in Mouse Given n-6 PUFA Rich Diet. The pathway IDs are the systematic name from MSigDB. Interestingly, we observed a lot of neural related pathways and even got significant signal in the calcium signaling pathway, which was reported to be associated with schizophrenia (S M Purcell et al., 2014).

ID	Size	Source	Description	Proportion of h^2	SE	Enrichment P-Value
M5884	275	NABA	Ensemble of genes encoding core extracellular matrix including ECM glycoproteins, collagens and proteoglycans	0.00317	0.00302	0.00210
M735	279	REACTOME	Genes involved in Neuronal System	0.0290	0.00627	0.00775
M15514	186	REACTOME	Genes involved in Transmission across Chemical Synapses	0.0212	0.00523	0.0197
M2890	178	KEGG	Calcium signaling pathway	0.0273	0.00825	0.0224
MIA		Super Pathway	All genes from pathways pertrubed by MIA	0.0716	0.0110	0.0334
M10792	267	KEGG	MAPK signaling pathway	0.0246	0.00799	0.0403
M3008	196	NABA	Genes encoding structural ECM glycoproteins	0.00374	0.00252	0.0429
M752	137	REACTOME	Genes involved in Neurotransmitter Receptor Binding And Downstream Transmission In The Postsynaptic Cell	0.0156	0.00451	0.0454

Table 3.4: Pathways significantly contributes to SNP heritability of schizophrenia. Only two significant pathways from were affected by the diets and they were also affected by MIA. This suggest that it is likely for the differential gene expression induced by MIA and genetic mutation might have act upon the same functional pathway in the development of schizophrenia.

3.3.5 Designing the Replication Study

Other than generation of hypothesis, we would also like to use the current information to help designing subsequent replication studies. Using Scotty (Busby et al., 2013), given that we would like to detect at least 90% of the genes that are differentially expressed by a 2× fold change at p< 0.01 and that at least 80% of genes has at least 80% of the maximum power, we will need at least 10 samples per group in the replication study given the current sequencing depth.

3.4 Discussion

In this hypothesis generation study, we demonstrated that *Sgk1* might be affected by n-3 PUFA rich diet in the cerebellum of MIA exposed mice. *Sgk1* is a serine/threonine kinase activated by PI3K signals and studies have shown 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 where transfection of *Sgk1* mutant DNA impairs the water maze performance in rat.

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 contributes 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). Moreover, *Sgk1* also up-regulates the glutamate transporters such as EAAT4 (Bohmer et al., 2004). The glutamate receptors are vital for clearance of glutamate from the synaptic cleft.

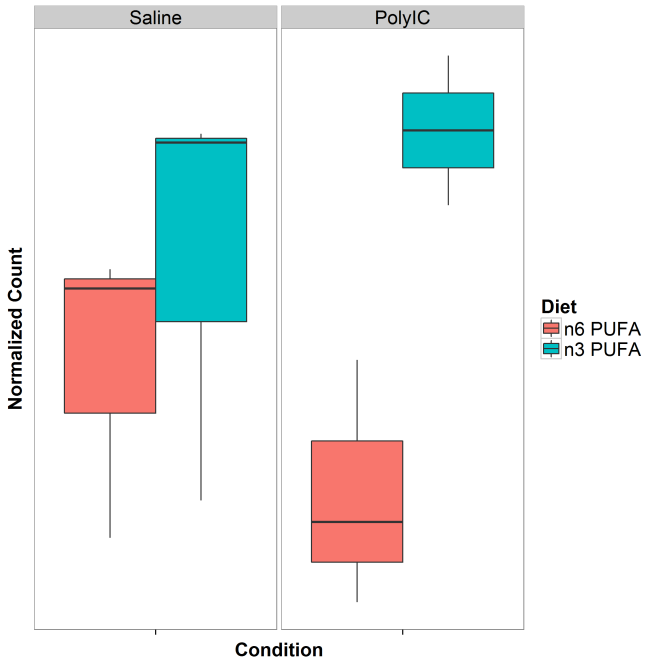


Figure 3.3: Normalized Expression of *Sgk1*. It was observed that the expression level of *Sgk1* increases after the mice was given a n3-PUFA rich diet where a significant increase was observed in mice exposed to PolyI:C.

This prevents excessive glutamate accumulation and therefore help to prevent the neurotoxic effects of glutamate (Lang, Strutz-Seebohm, et al., 2010). Considering the complexity of the glutamatergic system and the conflicting role of *Sgk1*, it is likely for more genes to play a role in the tight regulation of the glutamatergic system. However, it is likely that the disruption of *Sgk1* might be have an impact to the normal functioning of the glutamatergic system.

In our study, it was observed that upon given the n-3 PUFA rich diet, the *Sgk1* expression in the cerebellum increases (fig. 3.3). Although the increase was not significant in the saline mice, a significant up-regulation was observed in the PolyI:C exposed mice (fig. 3.3). Additionally, it was also observed that PI3K pathways and pathways related to FGF receptors and EGF receptors were significant when studying the effect of n-3 PUFA rich diet to PolyI:C exposed mice. Upon further investigation, it was found that the FGF receptors and EGF receptors are upstream of the PI3K-Akt pathway (fig. 3.4) which is responsible for the activation of *Sgk1*.

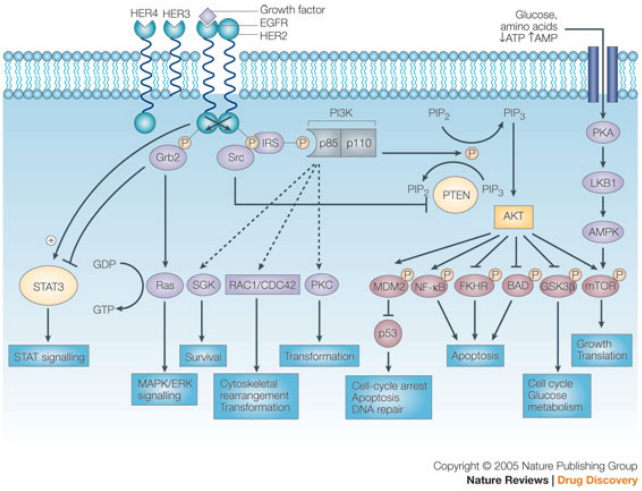


Figure 3.4: Schematic of signalling through the PI3K/AKT pathway. It was observed that the growth factors were upstream of the PI3K/AKT pathway of which *Sgk1* is one of the member of the pathway. Figure adopted from Hennessy et al. (2005) with permission from journal.

Although we were unable to provide direct connection between the expression of *Sgk1* and the improve functioning of the PolyI:C mice given n-3 PUFA diet, our results do suggest a possible effect of the n-3 PUFA rich diet in the expression of genes related to the PI3K/Akt pathway and might affect the expression of *Sgk1*. Further studies are therefore required to understand whether if the change in expression of *Sgk1* can account for the improved functioning of the PolyI:C mice. A possible design will be to induce the expression of *Sgk1* in PolyI:C mice through transfection and examine whether if the PolyI:C mice with higher expression of *Sgk1* display a reduction in schizophrenia-like behaviours.

An important point to note was that previous research of *Sgk1* has been focusing on the hippocampus and not the cerebellum. Thus, it is possible that *Sgk1* might have a different function in the cerebellum. It is therefore vital for subsequent research to study the effect of *Sgk1* in the cerebellum or for one to study the effect of n-3 PUFA rich diet on the gene expression in the hippocampus.

When examine the expression change in mice exposed to PolyI:C, none

of the genes were significantly differentiated. However, we do observe 12 pathways that contains genes that were more significant than genes not within the pathway (table 3.3). Interestingly, of the 12 significant pathways, 5 pathways were related to neuronal functions such as neuroactive ligand-receptor interaction (padj=1.27×10^{-3}), genes involved in neuronal system (padj=0.00153) and genes involved in transmission across chemical synapses (padj=0.00401). It has long been developed that the neuronal system and the neurotransmitter regulation plays a critical role in schizophrenia. For example, the disruption of the GABAergic and glutamtergic neuronal system might leads to excitation/inhibition imbalance which might ultimately lead to schizophrenia (Wassef, Baker, and Kochan, 2003). Moreover, the alteration in balance betweem excitation and inhibition can distort the connectivity patterns between different brain regions, thus leads to developmental and behavioral deficits (Cline, 2005).

Additionally, it was found that the calcium signaling pathway was significant when comparing the effect of MIA. The association of the calcium signaling pathway with schizophrenia was not a new finding (Lidow, 2003; S M Purcell et al., 2014; Stephan Ripke, B. M. Neale, et al., 2014). Previous exome sequencing study of schizophrenia by S M Purcell et al. (2014) has already report the enrichment of non-synonymous variants within the voltage gate calcium ion channel genes in the schizophrenia cases and the PGC schizophrenia GWAS has also found association between genes encoding the calcium channel subunits with schizophrenia. As calcium signaling pathway is the key component of the mechanism responsible for regulating neuronal excitability (Berridge, 2014), the disruption of the calcium signaling pathway is likely to have a profound effect on the neural function. Together, our results suggest that MIA might have disrupted the normal functioning of the neural system in the cerebellum, thus lead to schizophrenia-like behaviours in the adult mice yet follow up studies are required to validate our findings.

Moreover, we performed the partitioning of heritability hoping to see whether if the significant pathways have contributes disproportionately to the SNP heritability of schizophrenia. Interestingly, of the 12 significant pathways that were likely affected by MIA, 7 of them were found to be contributing a significantly higher portion to the SNP heritability. The "super pathway" containing all the genes participating in the MIA related pathways were also found to be significant (table 3.4). Based on our results, it is very likely that the differential gene expression in the cerebellum induced by early MIA events and the genetic variants to act upon the same pathways in the development of schizophrenia.

Specifically, among the 7 pathways found to be significantly contributes to the SNP heritability, 2 of those, both related to the extracellular matrix, were also found to be significantly affected by the n-3 PUFA rich diet in the PolyI:C exposed mouse. Intriguingly, those were also the only 2 pathways from the n-3 PUFA rich diet that were found to be significantly contributing to the SNP heritability of schizophrenia and they were both related to the extracellular matrix.

There are emerging evidence suggesting that the extracellular matrix (ECM) abnormality might be associated with schizophrenia (Berretta, 2012). The ECM glycoprotein Reelin has been reported to have a decreased expression in the cerebellum of schizophrenia patients (Maloku et al., 2010) and were found to be accompanied by decreased expression of glutamic acid decarboxylase 67 (Costa et al., 2001). Studies also suggested that Reelin might have important role in corticogenesis and synaptic maturation and stabilization (Berretta, 2012). Moreover, another ECM molecule, Semaphorin 3A has been reported to be increased in the cerebellum of subjects with schizophrenia (Eastwood et al., 2003). The Semaphroin 3A protein was found to regulates axonal guidance and has a critical role in the regulation of tangential migration of cortical GABAergic interneurons (Zimmer et al., 2010). It was also reported that the elevated Semaphorin 3A is associated with down-regulation of

genes involved in synaptic formation and maintenance (Eastwood et al., 2003). Together, these evidence suggest that the ECM molecules might have critical role in the development of schizophrenia.

It is therefore fascinating for us to observe a significant in the pathways related to the ECM glycoprotein. Not only was it significant when investigating the effect of MIA, it was also significant when we investigate the effect of n-3 PUFA rich diet in PolyI:C exposed samples. These pathways were also the only pathways that were affected by MIA, n-3 PUFA rich diet and also contribute for a significant disproportionated amount to the SNP heritability of schizophrenia. It has been reported that the n-3 PUFA diet can modulate the matrix metalloproteinase (MMP) (Derosa et al., 2009; Kavazos et al., 2015) which can regulates the ECM composition (Stamenkovic, 2003). Therefore it is possible that the n-3 PUFA diet has exerted its effect to the ECM through MMP. However, we acknowledge that our study lack power to develop direct evidence for such interaction. Further investigation are therefore vital and not until then can one understand how the n-3 PUFA direct interacts with the ECM and the effect of such interaction in MIA exposed individuals.

Finally, it is important to note that the current study serves only as a hypothesis generation study and the sample size was modest. We therefore like to use the current results to provide an estimation of sample size required for a replication study. By using Scotty (Busby et al., 2013), we have estimated that the replication study should contain at least 10 samples for each group in order for us to detect at least 80% of genes has at least 80% of the maximum power. We have also demonstrated that the batch effect can have a big impact to the association (fig. 3.2d), therefore one should always control for the batch effect whenever possible. Given the current resources, one of the preferred design for the follow up study are given in table 3.5.

3.4.1 Limitation

We first acknowledge that the sample size of the current study is small and are underpowered. This is reflected in the QQ-plots (fig. 3.2) where the observed p-values were generally smaller than would have expected. A better study design will include more samples yet we are limited by our budget. However, the importance of a pilot study is to identify potential targets for replications, hypothesis generation or to provide guidance for follow up studies. In this study, we have identified *Sgk1* as an interesting candidate gene that might have an important role in the effect of n-3 PUFA in PolyI:C exposed individuals. Our results also suggested that the differential expression induced by early MIA might act on the same functional pathways as genetic variations observed in SCZ. These provide interesting candidates for follow studies and we were able to estimate and design a better replication study based on the current data. Therefore we argue that as a hypothesis generation study, our study is successful.

Second, we examined only the male brains in the current study. The decision to direct experimental resources to males was made because there is evidence that the male fetus is more vulnerable to environmental exposures such as inflammation in prenatal life (Bergeron et al., 2013; Lein et al., 2007). We acknowledge that an interesting follow up study would be to investigate the gender difference in response to MIA and dietary change.

Third, although RNA Sequencing was performed, we have not performed any analysis on possible alternative splicing events or denovo transcript assembly. The reason behind such decision is that our sample size is simply too small. Without sufficient information, denovo transcript assembly can return noisy results. On the other hand, in order to investigate possible alternative splicing events, we would need to perform the analysis on transcript level instead of gene level. This increase the possible candidates from 47,400 genes to 114,083 transcripts. Combined with the

difficulties of the quantification of different isoforms, a much larger power is required for the alternative splicing analysis. On top of that, the functional annotation of transcripts is another difficult aspect to tackle. While there are a lot of information for the annotation of genes, information on functional difference between isoforms of the same gene were generally lacking. The lack of annotation simply leads to difficulties in making sense of the data. Thus although we acknowledge the possible importance of alternative splicing and denovo transcripts, we did not perform any alternative splicing analysis or denovo transcripts assembly. Nonetheless, the use of RNA Sequencing allow us to easily perform these experiments once sufficient samples are obtained.

Forth, it is important to note that 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. However, we do argues that RNA Sequencing can help to identify potential targets for protein assays where detail analysis can be performed on the protein level.

Finally, at the time of this thesis, we have yet completed any real time PCR (rt-PCR) or any functional studies to validate our findings. One of the most vital steps after any RNA Sequencing results is to validate the differential expression findings using the rt-PCR. Ideally, not only should one perform the rt-PCR on the sequenced samples, one should also perform the rt-PCR on an independent set of samples. Moreover, the RNA Sequencing only helps to identify possible candidates that were "associated" with a particular trait. It does not however provide any causal linkage between the phenotype and the differential expression. If one would

like to establish a direct linkage between the phenotype and the gene, one will need to carry out functional studies such as knock-in knock-out mouse design. So take for example, in order to understand the functional impact of the differential expression of *Sgk1*, one might try to examine whether if the pure up-regulation of *Sgk1* through transfection can reduce the schizophrenia-like behavior in PolyI:C exposed mice.

Currently, we are planning to perform the rt-PCR on *Sgk1* on all available samples. Shall the results be validated, we can then perform subsequent functional studies.

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 3.5: Design for follow up study. This will be the idea design for follow up study where litter effect, cage effect, batch effect and lane effects are all balanced out for the conditions. 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

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

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

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

synaptic functioning.

Moreover, because 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 were also interested to investigate 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. Most importantly, we found that pathways related to extracellular matrix (ECM) were affected not only by MIA, but also in PolyI:C samples given different diets. Pathways related to ECM were also found to contributes a disproportionate amount of SNP heritability to schizophrenia, suggesting that it is a potential target for future research.

4.1 Challenge in SNP-Heritability Estimation

Although it is now possible to estimates the SNP heritability based on the summary statistic from GWAS, a lot of questions remain unanswered in the estimation of SNP heritability. One major problem of SHREK and LDSC is that they both heavily relies on the Linkage Disequilibrium (LD) structures from the reference panel. However, GWAS samples can come from large variety of ethnic background thus the LD pattern estimated from the reference panel might not be representative of the sample LD. If the fundamental LD structure was not as expected, both SHREK and LDSC will not be able to provide an accurate estimate. For example, if a GWAS was conducted with 50% European and 50% African, population stratification may confound the results. Even if one control for the population stratification using the principle component analysis (PCA), the question remains whether if one should use the African reference panel or the European reference panel in the estimation of SNP heritability. Moreover, information regarding the population stratification

(e.g. the Principle Component (PC)) were usually unavailable making the problem more complicated. Further researches are therefore required to tackle the problem of population stratification before one can confidently estimate the SNP heritability from summary statistics from GWAS that might contain samples from large variety of ethnic background.

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

As technology advances, researchers can now use the next generation sequencing (NGS) technology to sequence the genome at per base resolution. This brings great prospect in the genetic studies for now we can directly identify the causal variants and can even detect rare causal variants providing sufficient sample size. However, both SHREK and LDSC are designed to work on the summary statistics from GWAS where common SNPs are usually the focus. Because of the huge sampling error associating with rare variants, the LD calculated for rare variants usually has a larger standard error (SE). As SHREK and LDSC are both heavily rely on an accurate LD estimation, they might be unsuitable for the estimation of the contribution of rare variants to schizophrenia. In fact, it was found that when all causal variants are rare (minor allele frequency (maf) $< 1\%$), LDSC will often generate a negative slope, and the intercept will exceed the mean χ^2 statistic (B. K. Bulik-Sullivan et al., 2015). As a result of that, a different algorithm must be developed in order to estimates the heritability from rare variants.

4.2 Schizophrenia: Future Perspectives

With the success of the PGC schizophrenia GWAS, research in schizophrenia genetics has finally entered an era of success. Through international collaboration, the PGC has finally identified 108 genetic loci that were associated with schizophrenia using GWAS approach (Stephan Ripke, B. M. Neale, et al., 2014). The results from the GWAS was based on the statistical association between variants and schizophrenia yet the functional involvement of these variants in the etiology of schizophrenia remains unknown. Functional analysis of these 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 suggest that factors other than common SNP were contributing to the heritability of schizophrenia.

Clear evidences suggest 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 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 be-

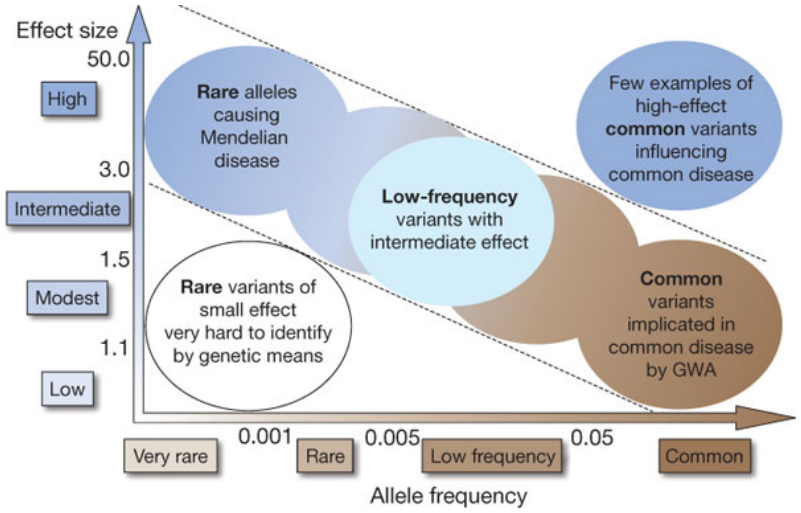


Figure 4.1: Relationship between effect size and allele frequency. It is expected that rare variants with large effect size were actively selected against in the population and therefore should be rare.

tween 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 suggest 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 important of epigenetic in the etiology of schizophrenia.

As a genetic 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 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 a deeper understanding of the complex disease mechanism of schizophrenia be obtained. Hopefully, in the near future, enough information can be gathered to start translating the research findings into clinical applications to help improving the quality of life of schizophrenia patients.

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. 54, 57).

American Psychiatric Association (2013). *Diagnostic and Statistical Manual of Mental Disorders*. American Psychiatric Publishing, p. 991 (cit. on pp. 1, 30).

Anders, S and W Huber (2010). "Differential expression analysis for sequence count data". eng. In: *Genome Biol* 11.10, R106 (cit. on p. 35).

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. 120).

Andrews, S. *FastQC A Quality Control tool for High Throughput Sequence Data* (cit. on p. 123).

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. 141).

Bernstein, Bradley E et al. (2010). "The NIH Roadmap Epigenomics Mapping Consortium." eng. In: *Nature biotechnology* 28.10, pp. 1045–1048 (cit. on p. 23).

Berretta, S (2012). "Extracellular matrix abnormalities in schizophrenia". eng. In: *Neuropharmacology* 62.3, pp. 1584–1597 (cit. on p. 139).

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. 138).

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. 135).

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. 27–29, 33, 119, 152).

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. 30).

Bulik-Sullivan, Brendan (2015). *Replicating MDD heritability Estimation* (cit. on pp. 104, 105).

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. 20, 21, 36, 39, 40, 61, 77, 105, 127, 131, 149).

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. 127, 135, 140).

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. 30).

Cingolani, Pablo et al. (2012). “A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w(1118); iso-2; iso-3”. In: *Fly* 6.2, pp. 80–92 (cit. on p. 126).

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. 120).

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. 27, 28, 36, 119, 152).

Cline, H (2005). “Synaptogenesis: a balancing act between excitation and inhibition”. eng. In: *Curr Biol* 15.6, R203–5 (cit. on p. 138).

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. 153).

Consortium, The International HapMap (2005). “A haplotype map of the human genome”. In: *Nature* 437, pp. 1299–1320 (cit. on p. 13).

Costa, E et al. (2001). “Dendritic spine hypoplasticity and downregulation of reelin and GABAergic tone in schizophrenia vulnerability.” eng. In: *Neurobiology of disease* 8.5, pp. 723–742 (cit. on p. 139).

Derosa, G et al. (2009). “Effects of long chain omega-3 fatty acids on metalloproteinases and their inhibitors in combined dyslipidemia patients.” eng. In: *Expert opinion on pharmacotherapy* 10.8, pp. 1239–1247 (cit. on p. 140).

Deverman, B E and P H Patterson (2009). “Cytokines and CNS development”. eng. In: *Neuron* 64.1, pp. 61–78 (cit. on p. 15).

Dobin, A et al. (2013). “STAR: ultrafast universal RNA-seq aligner”. eng. In: *Bioinformatics* 29.1, pp. 15–21 (cit. on pp. 35, 124, 127).

Eastwood, S L et al. (2003). "The axonal chemorepellant semaphorin 3A is increased in the cerebellum in schizophrenia and may contribute to its synaptic pathology." eng. In: *Molecular psychiatry* 8.2, pp. 148–155 (cit. on pp. 139, 140).

ENCODE Project Consortium (2012). "An integrated encyclopedia of DNA elements in the human genome". In: *Nature* 489.7414, pp. 57–74 (cit. on pp. 23, 153).

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. 124).

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, 6, 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. 25).

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. 23, 25).

Fromer, M et al. (2014). "De novo mutations in schizophrenia implicate synaptic networks". eng. In: *Nature* 506.7487, pp. 179–184 (cit. on p. 27).

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 p. 31).

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. 124).

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. 35).

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. 32).

Golan, David, Eric S Lander, and Saharon Rosset (2014). "Measuring missing heritability: Inferring the contribution of common variants". In: *Proceedings of the National Academy of Sciences* 111.49, E5272–E5281 (cit. on pp. 39, 83, 86, 99, 100).

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, 27).

— (1967b). "A polygenic theory of schizophrenia". In: *Proceedings of the National Academy of Sciences* 58.1, pp. 199–205 (cit. on p. 11).

Guennebaud, Gaël, Benoît Jacob, et al. (2010). *Eigen v3*. http://eigen.tuxfamily.org (cit. on pp. 52, 55).

Guey, Lin T. et al. (2011). "Power in the phenotypic extremes: A simulation study of power in discovery and replication of rare variants". In: *Genetic Epidemiology* 35.4, pp. 236–246 (cit. on pp. 50, 51).

Gui, Hongsheng et al. (2013). "RET and NRG1 interplay in Hirschsprung disease." eng. In: *Human genetics* 132.5, pp. 591–600 (cit. on p. 65).

Hansen, Per Christian (1987). "The truncated SVD as a method for regularization". In: *Bit* 27.4, pp. 534–553 (cit. on p. 53).

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).

Hennessy, Bryan T et al. (2005). "Exploiting the PI3K/AKT Pathway for Cancer Drug Discovery". In: *Nat Rev Drug Discov* 4.12, pp. 988–1004 (cit. on p. 137).

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. 70).

Ho, Nghia (2011). *OPENCV VS. ARMADILLO VS. EIGEN ON LINUX* (cit. on pp. 55, 108).

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. 35).

Jiang, Lichun et al. (2011). "Synthetic spike-in standards for RNA-seq experiments". In: *Genome Research* 21.9, pp. 1543–1551 (cit. on p. 145).

Kavazos, Kristyn et al. (2015). "Dietary supplementation with omega-3 polyunsaturated fatty acids modulate matrix metalloproteinase immunoreactivity in a mouse model of pre-abdominal aortic aneurysm." eng. In: *Heart, lung & circulation* 24.4, pp. 377–385 (cit. on p. 140).

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. 27).

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. 35).

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. 120).

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. 13).

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. 135).

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. 135, 136).

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. 135).

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. 141).

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. 34).

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. 14, 48).

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. 59).

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. 30, 33).

Li, Q, C Cheung, R Wei, E S Hui, et al. (2009). "Prenatal immune challenge is an environmental risk factor for brain and behavior change relevant to schizophrenia: evidence from MRI in a mouse model". eng. In: *PLoS One* 4.7, e6354 (cit. on pp. 30, 33, 121).

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. 37, 120, 148).

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 pp. 124, 127).

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. 11, 22, 147).

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. 138).

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. 124).

Maloku, Ekrem et al. (2010). "Lower number of cerebellar Purkinje neurons in psychosis is associated with reduced reelin expression". In: *Proceedings of the National Academy of Sciences* 107.9, pp. 4407–4411 (cit. on p. 139).

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. 35).

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. 123).

McClellan, Jon M, Ezra Susser, and Mary-Claire King (2007). "Schizophrenia: a common disease caused by multiple rare alleles". In: *The British Journal of Psychiatry* 190.3, pp. 194–199 (cit. on p. 13).

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. 28).

Mednick (1988). "Schizophrenia Following Prenatal Exposure to an Influenza Epidemic". In: *Arch Gen Psychiatry* 45.1 (cit. on p. 29).

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. 30–33, 152).

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. 30).

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. 152).

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. 52).

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. 151).

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. 120).

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. 27).

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. 122).

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. 59).

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. 31).

Peloso, Gina M et al. (2015). "Phenotypic extremes in rare variant study designs." ENG. In: *European journal of human genetics : EJHG* (cit. on p. 88).

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. 120).

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. 57, 70).

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. 70).

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. 26, 133, 138, 150, 151).

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. 14).

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 p. 59).

Reeves, P G, F H Nielsen, and G C Jr Fahey (1993). *AIN-93 purified diets for laboratory rodents: final report of the American Institute of Nutrition ad hoc writing committee on the reformulation of the AIN-76A rodent diet.* eng (cit. on p. 122).

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 p. 12).

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. 15, 16, 21–23, 70, 93, 138, 150).

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. 70).

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. 22).

Risch, N (1990a). “Linkage strategies for genetically complex traits. I. Multilocus models.” In: *American Journal of Human Genetics* 46.2, pp. 222–228 (cit. on pp. 11, 12).

— (1990b). “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 p. 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. 35).

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, 150).

Sanderson, Conrad (2010). *Armadillo: An Open Source C++ Linear Algebra Library for Fast Prototyping and Computationally Intensive Experiments*. Tech. rep. (cit. on pp. 55, 108).

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. 124).

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. 51, 68, 69, 102).

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. 152).

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. 30, 120, 152).

Stamenkovic, Ivan (2003). "Extracellular matrix remodelling: the role of matrix metalloproteinases." eng. In: *The Journal of pathology* 200.4, pp. 448–464 (cit. on p. 140).

Su, Zhan, Jonathan Marchini, and Peter Donnelly (2011). "HAPGEN2: Simulation of multiple disease SNPs". In: *Bioinformatics* 27.16, pp. 2304–2305 (cit. on pp. 58, 63).

Subramanian, Aravind et al. (2005). "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles". In: *Proceedings of the National Academy of Sciences* 102.43, pp. 15545–15550 (cit. on p. 126).

Sullivan, Patrick F (2005). "The Genetics of Schizophrenia". In: *PLoS Med* 2.7, e212 (cit. on p. 28).

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, 11, 22, 147, 150).

Szatkiewicz, J P et al. (2014). "Copy number variation in schizophrenia in Sweden". In: *Mol Psychiatry* 19.7, pp. 762–773 (cit. on pp. 25, 26).

Talkowski, Michael E et al. (2007). "Dopamine Genes and Schizophrenia: Case Closed or Evidence Pending?" In: *Schizophrenia Bulletin* 33.5, pp. 1071–1081 (cit. on p. 15).

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. 27, 36, 119).

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. 35).

Trebble, Timothy et al. (2003). "Inhibition of tumour necrosis factor-alpha and interleukin 6 production by mononuclear cells following dietary fish-oil supple-

mentation in healthy men and response to antioxidant co-supplementation." eng. In: *The British journal of nutrition* 90.2, pp. 405–412 (cit. on p. 120).

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. 135).

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. 120).

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. 142).

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. 31).

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 p. 26).

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. 35).

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. 58).

Wassef, A, J Baker, and L D Kochan (2003). "GABA and schizophrenia: a review of basic science and clinical studies". eng. In: *J Clin Psychopharmacol* 23.6, pp. 601–640 (cit. on p. 138).

Weir, B S and W G Hill (1980). "EFFECT OF MATING STRUCTURE ON VARIATION IN LINKAGE DISEQUILIBRIUM". In: *Genetics* 95.2, pp. 477–488 (cit. on p. 58).

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. 62).

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. 151).

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 pp. 18, 19).

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. 19, 20).

Yang, Jian, Naomi R. Wray, and Peter M. Visscher (2010). "Comparing apples and oranges: Equating the power of case-control and quantitative trait association studies". In: *Genetic Epidemiology* 34.3, pp. 254–257 (cit. on p. 49).

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. 17–19, 61).

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 pp. 120, 121).

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. 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. 15).

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. 34).

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 p. 19).

Zimmer, Geraldine et al. (2010). “Chondroitin sulfate acts in concert with semaphorin 3A to guide tangential migration of cortical interneurons in the ventral telencephalon.” eng. In: *Cerebral cortex (New York, N.Y. : 1991)* 20.10, pp. 2411–2422 (cit. on p. 139).