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 20, 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 thanks 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.
EGF	epidermal growth factor.
ERCC	External RNA Controls Consortium.
FGF	fibroblast growth factor.
GD	Gestation Day.
GO	Gene Ontology.
GWAS	Genome Wide Association Study.
IL-6	Interleukin-6.
LD	Linkage Disequilibrium.
LDSC	LD SCore regression.
LRT	likelihood ratio test.
maf	minor allele frequency.
MAPK	mitogen-activated protein kinase.
MIA	maternal immune activation.
MSigDB	Molecular Signatures Database.
NGS	next generation sequencing.
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.
SHREK	SNP HeRitability Estimation Kit.
SNP	Single Nucleotide Polymorphism.

Contents

Abstract **i**

Declaration **v**

Acknowledgments **vii**

Abbreviations **ix**

Contents **xi**

1 Introduction **1**

1.1 Schizophrenia 1

1.2 Understanding the Disease Mechanism 3

1.2.1 Broad Sense Heritability 3

1.2.2 Narrow Sense Heritability 4

1.2.3 Liability Threshold 7

1.2.4 Adoption Study 9

1.2.5 Twin Studies 9

1.3 Schizophrenia Genetics 11

1.3.1 The Human Genome Project and HapMap Project 13

1.3.2 Genome Wide Association Study 14

1.3.3 Contribution of Common SNPs 17

1.3.4 Rare Variants in Schizophrenia 26

1.4 Environmental Risk Factors of Schizophrenia 28

1.4.1 Prenatal Infection 30

1.4.2 RNA Sequencing 34

1.5 Summary 34

2 Heritability Estimation **35**

2.1 Introduction 35

2.2 Methodology 37

2.2.1 Heritability Estimation 38

2.2.2 Calculating the Standard error 43

2.2.3 Case Control Studies 46

2.2.4 Extreme Phenotype Sampling 47
2.2.5 Inverse of the Linkage Disequilibrium matrix 48
2.2.6 Implementation 52
2.2.7 Comparing with LD SCore regression 54
2.3 Comparing Different LD correction Algorithms 55
2.4 Comparison with Other Algorithms 58
2.4.1 Sample Size 59
2.4.2 Number of SNPs in Simulation 60
2.4.3 Genetic Architecture 61
2.4.4 Extreme Effect Size 63
2.4.5 Case Control Studies 64
2.4.6 Extreme Phenotype Sampling 66
2.5 Application to Real Data 68
2.6 Result 70
2.6.1 LD Correction 70
2.6.2 Comparing with Other Algorithms 71
2.6.3 Extreme Phenotype Simulation 86
2.6.4 Application to Real Data 92
2.7 Discussion 93
2.7.1 LD Correction 94
2.7.2 Simulation Results 97
2.7.3 Application to Real Data 104
2.7.4 Limitations and Improvements 107
2.8 Supplementary 110

3 n-3 Polyunsaturated Fatty Acid Rich Diet in Schizophrenia 119

3.1 Introduction 119
3.2 Methodology 123
3.2.1 Sample Preparation 123
3.2.2 RNA Extraction, Quality Control and Sequencing 124
3.2.3 Sequencing Quality Control 126
3.2.4 Alignment 126
3.2.5 Differential Expression Analysis 127
3.2.6 Functional Annotation 129
3.2.7 Partitioning of Heritability 129
3.2.8 Designing the Replication Study 130
3.3 Results 130
3.3.1 Sample Quality 130
3.3.2 Differential Expression Analysis 132
3.3.3 Functional Annotation 132
3.3.4 Designing the Replication Study 137
3.4 Discussion 137
3.4.1 Limitation 141

3.5 Supplementary . . . 144

4 Conclusion **147**

4.1 Challenge in SNP-Heritability Estimation . . . 148

4.2 Schizophrenia: Future Perspectives . . . 150

Bibliography **155**

List of Figures

1.1 Lifetime morbid risks of schizophrenia in various classes of relatives of a proband . . . 12
1.2 Enrichment of enhancers of SNPs associated with Schizophrenia 16
1.3 Risk factors of schizophrenia . . . 29
1.4 Hypothesized model of the impact of prenatal immune challenge on fetal brain development . . . 33

2.1 Cumulative Distribution of “gap” of the LD matrix . . . 52
2.2 GWAS Sample Size distribution . . . 60
2.3 Effect of LD correction to Heritability Estimation . . . 70
2.4 Mean of Quantitative Trait Simulation Results . . . 72
2.5 Variance of Quantitative Trait Simulation Results . . . 73
2.6 Estimation of Variance in Quantitative Trait Simulation . . . 74
2.7 Mean of Extreme Effect Size Simulation Result . . . 78
2.8 Variance of Extreme Effect Size Simulation Result . . . 79
2.9 Estimation of Variance in Extreme Effect Size Simulation . . . 80
2.10 Mean of Case Control Simulation Results (10 Causal) . . . 82
2.11 Variance of Case Control Simulation Results (10 Causal) . . . 83
2.12 Estimation of Variance in Case Control Simulation (10 Causal) . . . 84
2.13 Mean of Extreme Phenotype Selection Simulation Results . . . 88
2.14 Variance of Extreme Phenotype Selection Simulation Results . . . 89
2.15 Estimation of Variance in Extreme Phenotype Selection . . . 90
2.16 Effect of LD correction to Heritability Estimation with 50,000 SNPs . . 95
2.17 Effect of Extreme Sampling Design . . . 102
2.18 Mean of Case Control Simulation Results (50 Causal) . . . 110
2.19 Variance of Case Control Simulation Results (50 Causal) . . . 111
2.20 Estimation of Variance in Case Control Simulation (50 Causal) . . . 112
2.21 Mean of Case Control Simulation Results (100 Causal) . . . 113
2.22 Variance of Case Control Simulation Results (100 Causal) . . . 114
2.23 Estimation of Variance in Case Control Simulation (100 Causal) . . . 115
2.24 Mean of Case Control Simulation Results (500 Causal) . . . 116
2.25 Variance of Case Control Simulation Results (500 Causal) . . . 117
2.26 Estimation of Variance in Case Control Simulation (500 Causal) . . . 118

3.1 Sample Clustering 131
3.2 QQ Plot Statistic Results 133
3.3 Normalized Expression of *Sgk1* 138
3.4 Schematic of signalling through the PI3K/AKT pathway 140

4.1 Relationship between Effect Size and Allele Frequency 151

List of Tables

1.1 Top 20 leading cause of years lost due to disability 2
1.2 Enrichment of Top Cell Type of Schizophrenia 25

2.1 MSE of Quantitative Trait Simulation with Random Effect Size 76
2.2 MSE of Quantitative Trait Simulation with Extreme Effect Size 81
2.3 MSE of Case Control Simulation 87
2.4 Comparing the MSE of Extreme Phenotype Sampling and Random Sampling 91
2.5 Heritability Estimated for PGC Data Sets 92
2.6 Heritability Estimated for PGC Data Sets without Intercept Estimation 105

3.1 Sample Information 125
3.2 Significant Pathways When Comparing Effect of Diet in PolyI:C Exposed Mouse 135
3.3 Significant Pathways When Comparing Effect PolyI:C in Mouse Given n-6 polyunsaturated fatty acid (PUFA) Rich Diet 136
3.4 Design for Follow Up Study 145

Chapter 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 schizophrenia, 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 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 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.

1.2.1 Broad Sense Heritability

Heritability is defined as the *proportion* of total variance of a trait in a population explained by variation of genetic factors in the population. One can partition observed phenotype 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 broad sense heritability can then be defined as the ratio between the variance of the observed phenotype and the variance of the genetic effects

$$H^2 = \frac{\sigma^2_G}{\sigma^2_P}$$

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 time-point 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^2_A), variance of dominant genetic effects (σ^2_D) and other epistatic genetic effects (σ^2_I) such that

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

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 (Visscher, Hill, and 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

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

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

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

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

$$h^2 = \frac{1}{2}\frac{V_A}{\sigma_P^2} \qquad (1.4)$$

If we consider the simple linear regression equation $Y = X\beta + \epsilon$, its slope can be calculated as

$$\beta_{XY} = \frac{\mathrm{Cov}_{XY}}{\sigma_X Y} \qquad (1.5)$$

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

$$h^2 = 2\beta_{OP} \qquad (1.6)$$

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

$$h^2 = \frac{\beta_{XY}}{r} \qquad (1.7)$$

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 calculation 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 (Visscher, Hill, and Wray, 2008). The variance of phenotype can therefore be calculated as the variance under the normal distribution. However, such is not the case for disease 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.

One can then estimate the heritability of the 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

$$\begin{aligned}
T_G &= \text{Liability threshold of the general population} \\
T_R &= \text{Liability threshold of relatives of the index case} \\
q_G &= \text{Prevalence in the general population} \\
q_R &= \text{Prevalence in relatives of the index case} \\
L_a &= \text{Mean Liability of the index case}
\end{aligned}$$

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 eventhough 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 Neale (2003). Based on data obtained from 12 published schizophrenia twin studies, Sullivan, Kendler, and 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 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. 1.1) (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).

Not only does Risch (1990a) supports the polygenic model for schizophrenia,

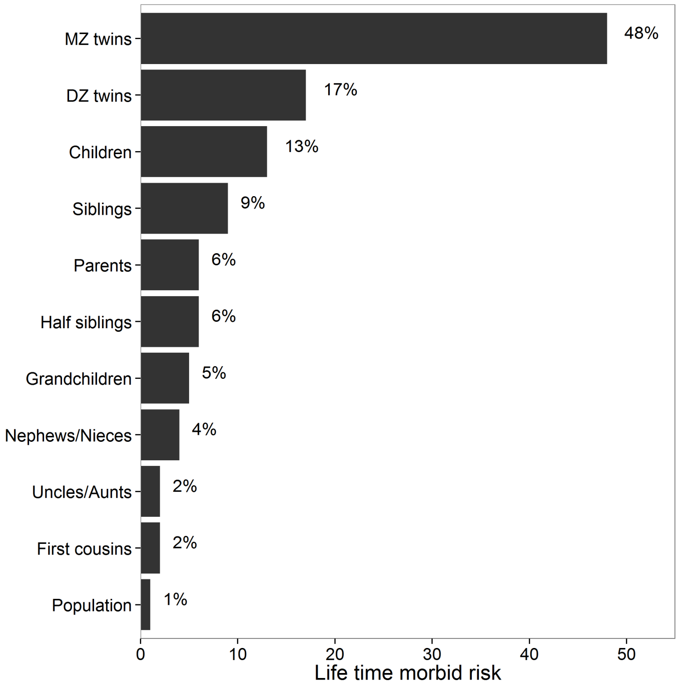


Figure 1.1: Lifetime morbid risks of schizophrenia in various classes of relatives of a proband. It was noted that the morbid risk of monozygotic (MZ) twins were only 48%, much lower than one would expect if schizophrenia follows a Mendelian pattern. Reproduced with permission from journal (Riley and Kendler, 2006).

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 (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

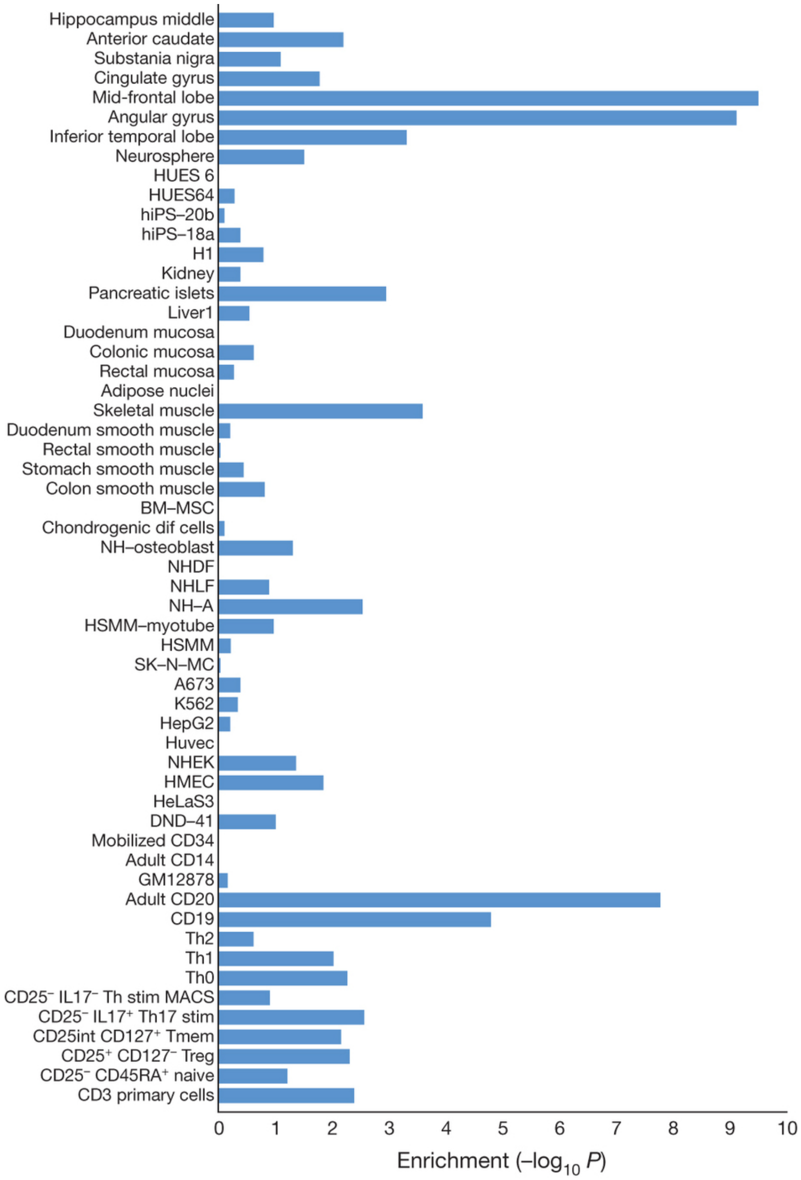
In GWAS, genome-wide genotyping array were commonly used to systematically detect common genetic variants such as SNP and copy number variation (CNV). For quantitative traits, the association between the trait and frequency of the variants were calculated using methods such as linear regression. On the 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 34,241 schizophrenia samples and 45,604 controls (Stephan Ripke et al., 2014). By combining the samples with those obtained by deCODE genetics, a total of 36,989 schizophrenia samples and 113,075 controls were used for the largest meta-analysis of schizophrenia. In their study (Stephan Ripke et al., 2014), 128 linkage-disequilibrium-independent SNPs were found to exceeded the genome-wide significance (p-value $\leq 5\times10^{-8}$), 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. 1.2)(Stephan Ripke et al., 2014).

Figure 1.2: 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 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 (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.

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)

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

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

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), 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. 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):

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

When confounding factors were present in the study (e.g. population stratification), eq. (1.15) can instead be defined as

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

where a is the contribution of confounding bias.

By considering eq. (1.16) as a regression model, 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 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, 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, Bulik-Sullivan et al. (2015) estimated the SNP heritability of schizophrenia using the summary statistics from the PGC schizophrenia GWAS (Stephan Ripke et al., 2014). By applying the liability threshold adjustment, 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 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 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 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 po-tentiation, 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 (≤ 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.

In order to understand the $G \times E$ interaction, one might need to first understand how environmental factors, especially that of prenatal infection, participate in the development of schizophrenia.

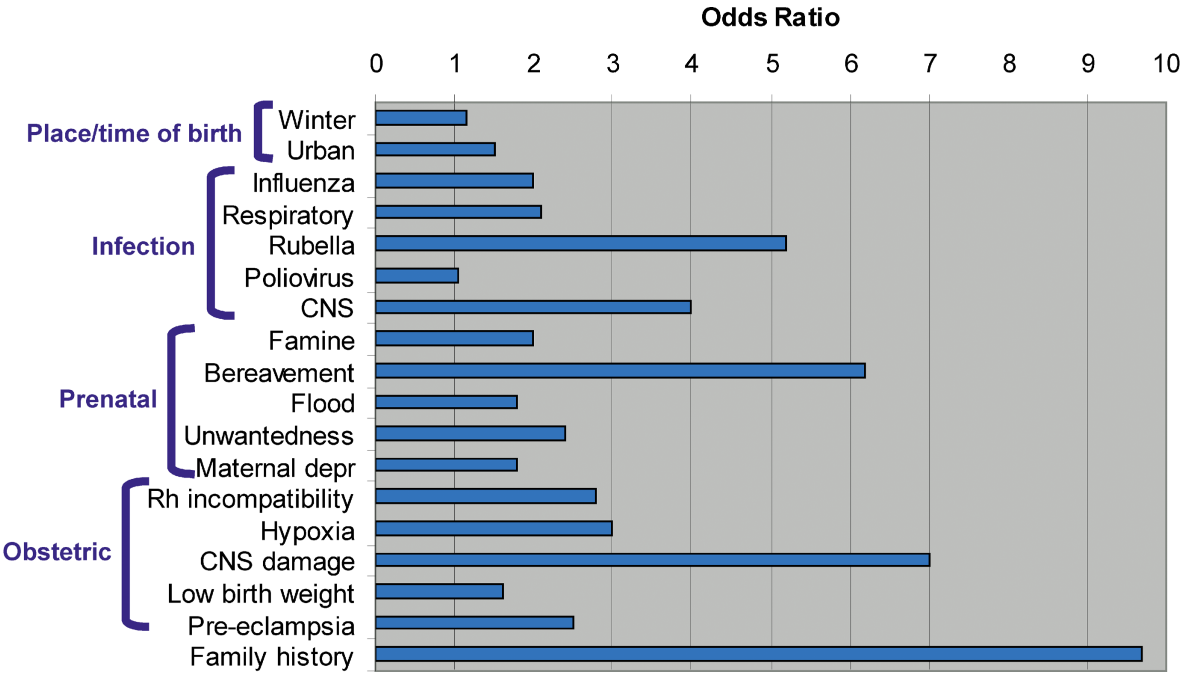


Figure 1.3: Risk factors of schizophrenia. It was observed that family history of schizophrenia was the largest risk factors. Risk of schizophrenia can be more than 9 times higher than the general population for individual with a family history of schizophrenia

1.4.1 Prenatal Infection

Prenatal infection has always been an important risk factor of schizophrenia, being the single largest non-genetic risk factor of schizophrenia (fig. 1.3)(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 neuroprotective 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

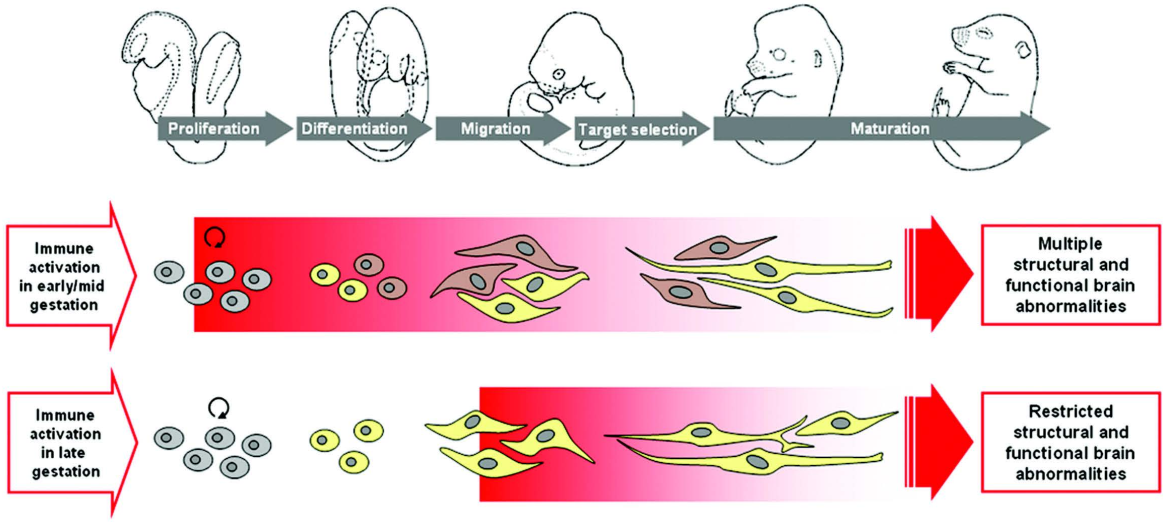

Proliferation
Differentiation
Migration
Target selection
Maturation
Immune activation in early/mid gestation
Multiple structural and functional brain abnormalities
Immune activation in late gestation
Restricted structural and functional brain abnormalities

Figure 1.4: 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)

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. 1.4) (U Meyer, Yee, and J Feldon, 2007).

In a separate study by Giovanoli et al. (2013), mice were exposed to a lower dosage of PolyI:C during early gestation. 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).

1.4.2 RNA Sequencing

1.5 Summary

To conclude, schizophrenia is a complex disorder affecting approximately 1% of the population worldwide. It is now known that the disease is affected by a combination of genetic and environmental factors. Therefore, 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 can act as a trigger for the disorder in apparently healthy individuals.

In this thesis, we would like to develop an algorithm for the estimation of SNP heritability from GWAS summary statistics that is robust to traits with different genetic architectures. We would also like to investigate the effect of case control sampling and extreme phenotype sampling on the performance of LDSC. On the other hand, as prenatal infection is the largest environmental risk for schizophrenia, we would like to understand how prenatal infection triggers schizophrenia through studying the change in global gene expressions in mice cerebellum.

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.

Chapter 4

Conclusion

SNP HeRitability Estimation Kit (SHREK), an algorithm for the estimation of heritability using Genome Wide Association Study (GWAS) test statistics are reported in this thesis. To our knowledge, this is the only algorithm other than LD SCore regression (LDSC) that can perform heritability estimation using summary statistics from GWAS. Our simulation results suggest 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 Single Nucleotide Polymorphism (SNP)-heritability of 0.174 (SD=0.00453), which is similar to the estimate of 0.197 (SD=0.0058) by LDSC.

On the other hand, we report a pilot RNA Sequencing study aiming to investigate the effect of maternal immune activation (MIA) and n-3 polyunsaturated fatty acid (PUFA) rich diet on the gene expression pattern in adult cerebellum. Overall, our results suggest the maternal immune activation (MIA) exposure might disrupt gene expressions that are related to neural function or the calcium ion signaling in the cerebellum of adult mice. In addition, we observed a significant up-regulation of *Sgk1* in polyriboinosinic-polyribocytidilic acid (PolyI:C) exposed mice that were given the n-3 PUFA rich diet. Consider the regulatory role of *Sgk1* in the glutamatergic system, it is possible that the n-3 PUFA rich diet can help to "rescued" some of the impaired functions in the glutamatergic system yet further studies are required. Based on our data, we were able to design a follow up study with adequate sample size and control for different confounding variables such as batch effect and cage effect.

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 (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 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. By continuing to increase the sample size of the GWAS, more associated variants will be able to be identified, thus increasing the SNP heritability. However, it is important to note that the heritability estimated from twin studies were not restricted to common SNPs. It is highly possible that other factors such as rare variants and epigenetic variants such as methylation might also contribute 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

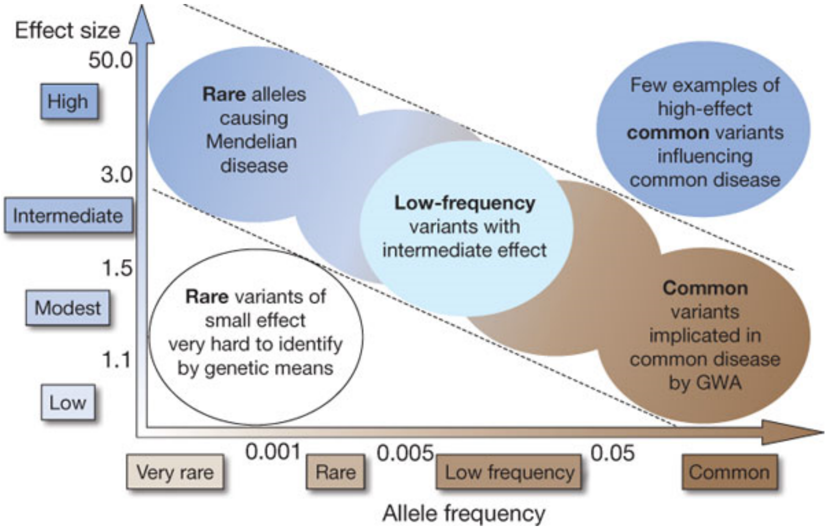

Effect size
50.0
High
3.0
Intermediate
1.5
Modest
1.1
Low
Rare alleles causing Mendelian disease
Low-frequency variants with intermediate effect
Few examples of high-effect common variants influencing common disease
Rare variants of small effect very hard to identify by genetic means
Common variants implicated in common disease by GWA
Very rare
0.001
Rare
0.005
Low frequency
0.05
Common
Allele frequency

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.

S. M. Purcell et al. (2014) was able to identify gene sets enriched by rare variants that were associated with schizophrenia using Exome Sequencing. This demonstrate the power of the sequencing technology in the identification of possible risk variants. Moreover, there was overlaps observed between genes harboring rare risk variants and those within the PGC schizophrenia GWAS (S. M. Purcell et al., 2014), suggesting that the rare variants and common variants studies are complementing each other. As more resources are devoted in to sequencing the genome of schizophrenia patients, more rare variants associated with schizophrenia are expected to be identified.

Currently, most of the focus in schizophrenia was directed to genetic variation yet it is possible that the heritability of schizophrenia is also transmitted in the form of epigenetic changes such as methylation. It was observed that the risk for individual born from a schizophrenic mother is larger than that from a schizophrenic father. This 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

American Psychiatric Association (2013). *Diagnostic and Statistical Manual of Mental Disorders*. American Psychiatric Publishing, p. 991 (cit. on pp. 1, 31).

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).

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. 30, 34, 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. 31).

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. 20–22, 149).

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. 32).

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. 28, 152).

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).

Deverman, B E and P H Patterson (2009). “Cytokines and CNS development”. eng. In: *Neuron* 64.1, pp. 61–78 (cit. on p. 17).

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).

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. 26).

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. 32).

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 pp. 33, 34).

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, 28).

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

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

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

Knapp, Martin, Roshni Mangalore, and Judit Simon (2004). “The global costs of schizophrenia.” In: *Schizophrenia bulletin* 30.2, pp. 279–293 (cit. on p. 2).

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).

Li, Miao-Xin Xin et al. (2011). “Evaluating the effective numbers of independent tests and significant p-value thresholds in commercial genotyping arrays and public imputation reference datasets”. In: *Human Genetics* 131.5, pp. 747–756 (cit. on p. 14).

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 p. 32).

Li, Q, C Cheung, R Wei, E S Hui, et al. (2009). “Prenatal immune challenge is an environmental risk factor for brain and behavior change relevant to schizophrenia: evidence from MRI in a mouse model”. eng. In: *PLoS One* 4.7, e6354 (cit. on p. 32).

Li, Q, Y O Leung, et al. (2015). “Dietary supplementation with n-3 fatty acids from weaning limits brain biochemistry and behavioural changes elicited by prenatal exposure to maternal inflammation in the mouse model.” eng. In: *Translational psychiatry* 5, e641 (cit. on p. 35).

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).

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).

Mednick (1988). “Schizophrenia Following Prenatal Exposure to an Influenza Epidemic”. In: *Arch Gen Psychiatry* 45.1 (cit. on p. 30).

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. 31, 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. 32).

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. 153).

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. 152).

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. 32).

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. 27, 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).

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 et al. (2014). "Biological insights from 108 schizophrenia-associated genetic loci". In: *Nature* 511, pp. 421–427 (cit. on pp. 15, 16, 22, 23, 150).

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. 23).

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).

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. 2, 150).

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. 153).

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. 32, 152).

Sullivan, Patrick F (2005). “The Genetics of Schizophrenia”. In: *PLoS Med* 2.7, e212 (cit. on p. 30).

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).

Szatkiewicz, J P et al. (2014). “Copy number variation in schizophrenia in Sweden”. In: *Mol Psychiatry* 19.7, pp. 762–773 (cit. on pp. 26, 27).

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 p. 28).

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).

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. 33).

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. 27).

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. 152).

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

Yang, Jian, Beben Benyamin, et al. (2010). “Common SNPs explain a large proportion of the heritability for human height.” eng. In: *Nature genetics* 42.7, pp. 565–569 (cit. on p. 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 p. 20).

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. 18, 19).

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. 17).

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. 20).