Heritability Estimation and Risk Prediction in 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

September 29, 2015

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.....

Acknowledgements

Abbreviations

CEU Northern Europeans from Utah. 15, 17, 18

CI confidence interval. 4

DSM Diagnostic and Statistical Manual of Mental Disorders. 1, 2

GCTA Genome-wide Complex Trait Analysis. 16, 18–20, 22–24

GD Gestation Day. 26

GO Gene Ontology. 27, 28, 31, 44

GWAS Genome Wide Association Study. 5, 7, 8, 17

IQ intelligence quotient. 4

LD Linkage Disequilibrium. 8, 9, 11, 13–15, 17, 18

LDSC LD SCore. 16, 18–23

maf Minor Allele Frequency. 18–20

MAGMA Multi-marker Analysis of GenoMic Annotation. 27, 31

MSE mean squared error. 21, 23

. 3

PC Principle Component. 27, 29

PGC Psychiatric Genomics Consortium. 27, 31

PGS Polygenic Risk Score. 35

RIN RNA integrity number. 26

RPKM Reads Per Kilobase per Million mapped reads. 26, 28, 30

SCZ schizophrenia. 17, 19, 25

SE Standard Error. 13

SHREK SNP Heritability and Risk Estimation Kit. 16, 18–23

SNP Single Nucleotide Polymorphism. 8, 9, 11, 15, 17–20, 23, 27

SVD Singular Value Decomposition. 14

tSVD Truncated Singular Value Decomposition. 14, 15

WGCNA Weighted Gene Co-expression Network Analysis. 26, 27

WHO World Health Organization. 1, 2

YLD years lost due to disability. 1, 2

Contents

List of Figures

List of Tables

Chapter 1

Introduction - Heritability Estimation in Schizophrenia

1.1 Schizophrenia

Schizophrenia is a detrimental psychiatric disorder, affecting around $0.3 \sim 0.7\%$ of the population(American Psychiatric Association, 2013). It is characterized by positive symptoms including delusions, hallucinations, disorganized speech and grossly disorganized behavior, and negative symptoms such as the diminished emotional expression(American Psychiatric Association, 2013) with a typical age of onset at late adolescent or late 20s in male and late 20s or early 30s in female(Schultz, North, and C. G. Shields, 2007).

Schizophrenia not only impose long lasting health, social and financial burden not only to the patients, but also to their families(Knapp, Mangalore, and Simon, 2004). Even more so, patients with schizophrenia increased suicide rate (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 aiming to delineate the disease mechanics and be able to identify the risk factors. Arguably, the most important first step to any schizophrenia study is to have a robust and reliable disease diagnosis.

1.2 Diagnosis

Schizophrenia was first named "Dementia Praecox" by Dr. Emil Kraepelin and was later renamed as schizophrenia by Dr. Eugen Bleuler(Jablensky, 2010). Early nosological entity for schizophrenia such as that in Diagnostic and Statistical Manual of Mental Disorders (DSM)-I and DSM-II were vague and unreliable where the inter-rater agreement can be as low as 54%.(Tsuang, Stone, and Faraone, 2000; Harvey et al., 2012)

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 100,000 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

Later nosologies addressed these problem by introducing structural assessment and clear defined criteria. With these improvements, the inter-rater agreement of DSM-III raised to $\sim$ 90% (Harvey et al., 2012), suggesting the diagnosis were much more reliable.

Currently DSM is at its 5th edition(American Psychiatric Association, 2013). A patient will be diagnosed with schizophrenia(F20.9) if they 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.

1.3 Risk Factors of Schizophrenia

1.3.1 Epidemiological Studies

1.3.2 Familial Studies

Considerable efforts has been made to identify the risk factors of schizophrenia.

It was first observed that schizophrenia tends to aggregate in families where relatives of schizophrenic patients tends to have higher risk of developing schizophrenia(Gottesman, 1991). The fundamental question

to ask would be whether what is the relative contribution of genetic and environmental factors to schizophrenia? **Whether if the *difference* in phenotype is accounted by genetics or environmental factors?**

1.4 Estimation of Heritability

One key concept in quantitative genetics is *heritability*, which was defined as *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

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

One can further partition the genetic variance into variance of additive genetic effects (σ_A^2), variance of dominant genetic effects (σ_D^2) and other epistatic genetic effects (σ_I^2) such that

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

where additive genetic variance was the variance explained by the average effects of all loci involved in the determination of the trait, whereas dominant genetic effects and epistatic genetic effectss 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 from each other. Considering that dominance and non-additive genetic effects were concerning the interactive effect, which usually involve more than one copy of the alleles, thes effects are unlikely to contribute to the resemblance between relatives (Visscher, Hill, and Wray, 2008), thus it is usually more useful to consider the narrow sense heritability(h^2) which only consider the additive genetic effects:

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

In order to obtain the variance explained by the genetic effects, we can consider the following simplistic scenario:

Let there be two alleles A_1 and A_2 each has an allele frequencies of p and q, then the genotype frequencies will follow

	$A_1(p)$	$A_2(q)$
$A_1(p)$	p^2	pq
$A_2(q)$	pq	q^2

Now if we consider the effects of heterozygous genotypes to be h and d be twice the difference of effects between the two homozygous genotypes, then we have

Genotype	A_1A_1	A_1A_2	A_2A_2
Frequency	p^2	$2pq$	q^2
Genotype effect	d	h	$-d$

Based on these information, we can calculate the mean genotype effect in a population as

$$\begin{aligned}\mu_G &= p^2d + 2pqh - q^2d \\ &= (p^2 - q^2)d + 2pqh \\ &= (p-q)(p+q)d + 2pqh \\ &= (p-q)d + 2pqh \end{aligned} \tag{1.2}$$

The variance of genetic effects is therefore

$$\begin{aligned}\sigma_G^2 &= p^2d^2 + q^2d^2 + 2pqh^2 - \mu_G^2 \\ &= p^2d^2 + q^2d^2 + 2pqh^2 - [(p-q)d + 2pqh]^2 \\ &= (p^2+q^2)d^2 + 2pqh^2 - [(p-q)^2d^2 + 4p^2q^2h^2 + 4(p-q)pqdh] \\ &= [(p+q)^2 - 2pq]d^2 + 2pqh^2 - [(p^2+q^2-2pq)d^2 + 4p^2q^2h^2 + 4(p-q)pqdh] \\ &= -2pqd^2 + 2pqh^2 - [-4pqd^2 + 4p^2q^2h^2 + 4(p-q)pqdh] \\ &= 2pq[-d^2 + h^2 + 2d^2 - 2pqh^2 - 2(p-q)dh] \\ &= 2pq[d^2 + (1-2pq)h^2 - 2(p-q)dh] \\ &= 2pq[d^2 + (1-1+p^2+q^2)h^2 - 2(p-q)dh] \\ &= 2pq[d^2 + ((q+p)(q-p) + 2pq)h^2 - 2(p-q)dh] \\ &= 2pq[d^2 + (q-p)h^2 + 2(q-p)dh + 2pqh^2] \\ &= 2pq[d + (q-p)h]^2 + 4p^2q^2h^2 \end{aligned} \tag{1.3}$$

If we then consider the A_1 allele as the increasing allele, then the dosage of A_1 in the genotypes will be

Genotype	A_1A_1	A_1A_2	A_2A_2
Frequency	p^2	$2pq$	q^2
Genotype effect(y)	d	h	$-d$
Dose(x)	2	1	0

With the dosage concept, we can regress the genotype effect against the dose using linear regression. The variance due to regression will then be representative of the additive genetic effects whereas the variance due to residual will be representative of the non-additive genetic effects.

The mean of dose will be

$$\mu_x = 2p^2 + 2pq = 2p(p+q) = 2p$$

and the variance will be

$$\begin{aligned}\sigma_x^2 &= 2^2p^2 + 2pq - \mu_x^2 \\ &= 4p^2 + 2pq - 4p^2 \\ &= 2pq \end{aligned} \tag{1.4}$$

then, by considering the covariance between the dose and genetic effect, we get

$$\begin{aligned}\sigma_{x,y} &= 2p^2d + 2pqh - \mu_x\mu_G \\ &= 2p^2d + 2pqh - 2p[(p-q)d + 2pqh] \\ &= 2pqh + 2pqd - 4p^2qh \\ &= 2pqd + 2pqh(1-2p) \\ &= 2pqd + 2pqh(q-p) \\ &= 2pq[d + h(q-p)] \end{aligned} \tag{1.5}$$

with these information, we can calculate the variance due to regression as

$$\begin{aligned}\beta_{\sigma_{x,y}} &= 2pq[d + (q-p)h][d + (q-p)h] \\ &= 2pq[d + (q-p)h]^2 \\ &= \sigma_A^2 \end{aligned} \tag{1.6}$$

and the variance due to regression is represented as

$$\begin{aligned}\sigma_y^2 - \beta_{\sigma_{x,y}} &= 4p^2q^2h^2 \\ &= \sigma_D^2 \end{aligned} \tag{1.7}$$

One key feature of heritability is that it is a *ratio* of *populational* measurement at a specific time point. As a result of that, the heritability estimation might differ from one population to another and one might obtain a different heritability estimate if the method or time-point of measurement of the trait differs. An classical 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 that as they become more independent, the effect of shared environment diminishes, leading to a *increased portion* of variance in IQ explained by the variance in genetic(Bouchard, 2013).

A challenge with these calculation was that when working with discontinuous trait, the variance of phenotype is often determined by the population prevalence. Arguably, one of the key concept in study of discontinuous trait is the liability threshold model.

1.5 Liability Threshold

According 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 also affected by combination of many 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.

1.6 Twin Studies of Schizophrenia

To find out whether if genetic or environment were the main mediator in schizophrenia, researchers conducted varies twin and adoption studies. In one of the landmark study conducted by Sullivan, Kendler, and Neale, 2003, a quantitative meta-analysis was performed on 12 published schizophrenia twin studies. They 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%), suggesting that schizophrenia was largely mediated by the genetic elements.

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 a different estimation of heritability were obtained from the two studies, there is no doubt that schizophrenia is highly heritable, justify the initiative of conducting genetic research in elucidating genetic risk factor of schizophrenia.

1.7 Genetic Analysis of Schizophrenia

Although it was clear that schizophrenia is a genetic disease, family studies shown that it is unlikely for schizophrenia to be affected by genes with a large effect. Specifically, shall schizophrenia be a Mendelian disorder, then we would expect all monozygotic siblings of the proband to also suffer from schizophrenia. However, studies found that 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 J. Shields, 1967 proposed that schizophrenia follows a polygenic model where disease phenotype were determined by the additive effects from multiple genes. Thus, schizophrenia is a complex genetic disorder with complicated pattern of inheritance. Their hypothesis was supported by the calculation of Risch, 1990 by taking into account of different inheritance model and the life time morbid risk observed in relatives of affected individuals.

Another interesting conclusion from the calculation of Risch, 1990 was the effect size of individual locus. By comparing the observed life time morbid risk and the calculated risk from different models, it was observed that the model with a maximum risk of $\sim\leq 2$ was more compatible than that with risk ≥ 3. Thus, not only is schizophrenia polygenic, the effect size of each individual causal genes are likely to be small (less than 2 fold risk elevation).

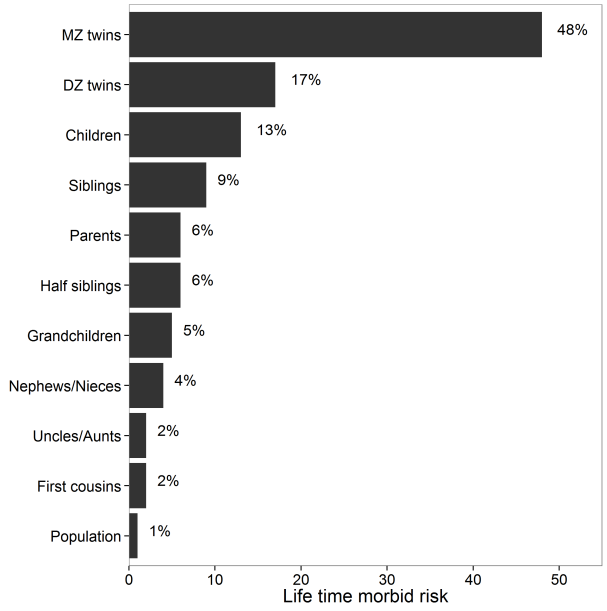


Figure 1.1: Lifetime morbid risks of schizophrenia in various classes of relatives of a proband. It was noted that the morbid risk of monozygotic (MZ) twins were only 48%, much lower than one would expect if schizophrenia follows a Mendelian pattern. Reproduced with permission from journal(Riley and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

Considerable research effort has been made in past years, aiming to identify the susceptible genes of schizophrenia. However, early linkage studies were unable to capture the small effects and largely return with inconsistent results(Harrison and Weinberger, 2005). As technology progress, large scale hypothesis free testings such as that of Genome Wide Association Study (GWAS) starts to become available.

1.8 Narrow Sense Heritability

1.8.1 Genome-wide Complex Trait Analysis

1.8.2 LD SCore

1.9 Missing heritability

1.10 Antipsychotics

1.10.1 Pharmacogenetics and Pharmacogenomics

1.10.2 Extreme Phenotype Selection

1.11 Summary

Chapter 2

Heritability Estimation

2.1 Introduction

Why we are interested in calculating the heritability? The challenges (meta analysis, only test statistics, take long time to calculate the GRM) We developed a programme SHREK Recently there is another programme published based aim to solve the same problem and we should also compare and contrast our method with them. Aim of this chapter -¿ Calculate a robust estimation of heritability for all genetic architechture. LD score regression!

2.2 Methodology

The overall aims of this study is to develop a robust algorithm for the estimation of the narrow sense heritability using only the summary statistic from a GWAS study. The work in this chapter were done in collaboration with my colleagues who have kindly provide their support and knowledges to make this piece of work possible. Dr Johnny Kwan, Dr Miaxin Li and Professor Sham have helped to laid the framework of this study. Dr Timothy Mak has derived the mathematical proof for our heritability estimation method. Miss Yiming Li, Dr Johnny Kwan, Dr Miaxin Li, Dr Timothy Mak and Professor Sham have helped with the derivation of the standard error of the heritability estimation. Dr Henry Leung has provided critical suggestions on the implementation of the algorithm.

2.2.1 Heritability Estimation

The narrow-sense heritability is defined as

$$h^2 = \frac{\mathrm{Var}(X)}{\mathrm{Var}(Y)}$$

where $\mathrm{Var}(X)$ is the variance of the genotype and $\mathrm{Var}r(Y)$ is the variance of the phenotype. In a GWAS, regression were performed between the Single Nucleotide Polymorphisms (SNPs) and the phenotypes, giving

$$Y = \beta X + \epsilon \tag{2.1}$$

where Y and X are the standardized phenotype and genotype respectively. ϵ is then the error term, accounting for the non-genetic elements contributing to the phenotype (e.g. Environment factors). Based on eq. (2.1), one can then have

$$\begin{aligned} \mathrm{Var}(Y) &= \mathrm{Var}(\beta X) + \mathrm{Var}(\epsilon) \\ \mathrm{Var}(Y) &= \beta^{\mathrm{Var}}(X) \\ \beta^2 \frac{\mathrm{Var}(X)}{\mathrm{Var}(Y)} &= 1 \end{aligned} \tag{2.2}$$

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

A challenge in calculating the heritability from GWAS data is that usually only the test-statistic or p-value were provided and one will not be able to directly calculate the heritability based on eq. (2.2). In order to estimation the heritability of a trait from the GWAS test-statistic, we first observed that when both X and Y are standardized, β^2 will be equal to the coefficient of determination (r^2). Then, based on properties of the Pearson product-moment correlation coefficient:

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

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

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

It is observed that t^2 will follow the F-distribution and when n is big, t^2 will converge into χ^2 distribution.

When the effect size is small and n is big, r^2 will be approximately χ^2 distributed with mean ~ 1. We can then approximate eq. (2.4) as

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

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

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

When there are Linkage Disequilibrium (LD) between each individual SNPs, the situation will become more complicated as each SNPs' observed effect will contains effect coming from other SNPs in LD with it.

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

To account for the LD structure, we first assume our phenotype $\boldsymbol{Y}$ and genotype $\boldsymbol{X} = (X_1, X_2, \ldots, X_m)^t$

are standardized and that

$$\boldsymbol{Y} \sim f(0,1)$$
$$\boldsymbol{X} \sim f(0,\boldsymbol{R})$$

Where $\boldsymbol{R}$ is the LD matrix between SNPs.

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

$$\boldsymbol{Y} = \boldsymbol{\beta}^t \boldsymbol{X} + \epsilon \tag{2.8}$$

Definition of heritability will then become

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

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

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

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

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

Now if we consider the covariance between SNP i (X_i) and Y, we have

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

As both X and Y are standardized, the covariance will equal to the correlation and we can define the correlation between SNP i and Y as

$$\rho_i = \boldsymbol{R}_i \boldsymbol{\beta}_j \tag{2.12}$$

In reality, the *observed* correlation usually contains error. Therefore we define the *observed* correlation to be

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

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

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

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

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

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

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

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

$$\begin{aligned} \chi^2 &= \hat{\rho}^2 n \\ &= n(\boldsymbol{R}_i\boldsymbol{\beta}_j + \frac{\epsilon_i}{\sqrt{n}})^2 \end{aligned}$$

The expectation of χ^2 is then

$$\begin{aligned} \mathrm{E}(\chi^2) &= n(\boldsymbol{R}_i\boldsymbol{\beta}\boldsymbol{\beta}^t\boldsymbol{R}_i + 2\boldsymbol{R}_i\boldsymbol{\beta}\frac{\epsilon_i}{\sqrt{n}} + \frac{\epsilon_i^2}{n}) \\ &= n\boldsymbol{R}_i\boldsymbol{H}\boldsymbol{R}_i + 1 \end{aligned}$$

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

$$\begin{aligned} \sum_i(\chi_i^2 - \mathrm{E}(\chi_i^2))^2 &= \sum_i(\chi_i^2 - (n\boldsymbol{R}_i\boldsymbol{H}\boldsymbol{R}_i + 1))^2 \\ &= \sum_i(\chi_i^2 - 1 - n\boldsymbol{R}_i\boldsymbol{H}\boldsymbol{R}_i)^2 \end{aligned}$$

If we define

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

we got

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

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

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

And the heritability is then defined as

$$\hat{Herit ability} = \mathbf{1}^t \boldsymbol{R}_{\boldsymbol{sq}}^{-1} \boldsymbol{f} \tag{2.17}$$

2.2.2 Calculating the Standard Error

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

$$\begin{aligned} \mathrm{Var}(H) &= \mathrm{E}[H^2] - \mathrm{E}[H]^2 \\ &= \mathrm{E}[(\mathbf{1}^t \boldsymbol{R}_{\boldsymbol{sq}}^{-1} \boldsymbol{f})^2] - \mathrm{E}[\mathbf{1}^t \boldsymbol{R}_{\boldsymbol{sq}}^{-1} \boldsymbol{f}](\mathrm{E}[\mathbf{1}^t \boldsymbol{R}_{\boldsymbol{sq}}^{-1} \boldsymbol{f}])^t \\ &= \mathrm{E}[\mathbf{1}^t \boldsymbol{R}_{\boldsymbol{sq}}^{-1} \boldsymbol{f} \boldsymbol{f}^t \boldsymbol{R}_{\boldsymbol{sq}}^{-1} \mathbf{1}] - \mathrm{E}[\mathbf{1}^t \boldsymbol{R}_{\boldsymbol{sq}}^{-1} \boldsymbol{f}](\mathrm{E}[\mathbf{1}^t \boldsymbol{R}_{\boldsymbol{sq}}^{-1} \boldsymbol{f}])^t \\ &= \mathbf{1}^t \boldsymbol{R}_{\boldsymbol{sq}}^{-1} \mathrm{E}[\boldsymbol{f} \boldsymbol{f}^t] \boldsymbol{R}_{\boldsymbol{sq}}^{-1} \mathbf{1} - \mathrm{E}[\mathbf{1}^t \boldsymbol{R}_{\boldsymbol{sq}}^{-1} \boldsymbol{f}](\mathrm{E}[\mathbf{1}^t \boldsymbol{R}_{\boldsymbol{sq}}^{-1} \boldsymbol{f}])^t \\ &= \mathbf{1}^t \boldsymbol{R}_{\boldsymbol{sq}}^{-1} \mathrm{Var}(\boldsymbol{f}) \boldsymbol{R}_{\boldsymbol{sq}}^{-1} \mathbf{1} + \mathrm{E}[\mathbf{1}^t \boldsymbol{R}_{\boldsymbol{sq}}^{-1} \boldsymbol{f}](\mathrm{E}[\mathbf{1}^t \boldsymbol{R}_{\boldsymbol{sq}}^{-1} \boldsymbol{f}])^t - \mathrm{E}[\mathbf{1}^t \boldsymbol{R}_{\boldsymbol{sq}}^{-1} \boldsymbol{f}](\mathrm{E}[\mathbf{1}^t \boldsymbol{R}_{\boldsymbol{sq}}^{-1} \boldsymbol{f}])^t \\ &= \mathbf{1}^t \boldsymbol{R}_{\boldsymbol{sq}}^{-1} \mathrm{Var}(\boldsymbol{f}) \boldsymbol{R}_{\boldsymbol{sq}}^{-1} \mathbf{1} \end{aligned} \tag{2.18}$$

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

We first consider the standardized genotype X_i with standard normal mean z_i and non-centrality parameter μ_i, we have

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

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

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

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

as:

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

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

$$\mathrm{Cov}(X_i^2, X_j^2) = \mathrm{E}[z_i^2z_j^2] + 4\mu_i\mu_jR_{ij} - 1$$

By definition,

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

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

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

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

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

Considering that we only have the *observed* expectation, we should re-define eq. (2.19) as

$$\mathrm{Cov}(X_i^2, X_j^2) = \frac{2R_{ij}^2 + 4R_{ij}\mu_i\mu_j}{n^2} \tag{2.20}$$

where n is the sample size.

By substituting eq. (2.20) into eq. (2.18), we will get

$$\mathrm{Var}(H) = \mathbf{1}^t\boldsymbol{R}_{\boldsymbol{sq}}^{-1}\frac{2\boldsymbol{R}_{\boldsymbol{sq}} + 4\boldsymbol{R}\circ\boldsymbol{z}\boldsymbol{z}^t}{n^2}\boldsymbol{R}_{\boldsymbol{sq}}^{-1}\mathbf{1} \tag{2.21}$$

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

Problem with eq. (2.21) were that not only does it requires the direction of effect, the error in the

LD matrix also tends to amplify due to its predominant role in the equation, leading to un-stable estimation of the Standard Error (SE).

Another way to get the SE is based on the fact that $\boldsymbol{f}$ is approximately χ^2 distributed. Therefore eq. (2.16) can be viewed as a decomposition of a vector of χ^2 distributions with degree of freedom of 1. Replacing the vector $\boldsymbol{f}$ with a vector of 1, we can perform the decomposition of the degree of freedom, getting the "effective number"(e) of the association(Li et al., 2011). Substituting e into the variance equation of non-central χ^2 distribution will yield

$$\mathrm{Var}(H) = \frac{2(e + 2H)}{n^2} \tag{2.22}$$

eq. (2.22) will gives us an heuristic estimation of the SE.

2.2.3 Case Control Studies

2.2.4 Extreme Phenotype Selections

eq. (2.17) can naturally be applied to the quantitative trait scenario.

2.2.5 Calculating the Linkage Disequilibrium matrix

To estimate the heritability, the population LD matrix is required. In reality, one can only obtain the LD matrix based on a subset of the population (e.g. the 1000 genome project(Project et al., 2012) or the HapMap project(Altshuler et al., 2010)). There are therefore sampling errors among the LD elements.

Now if we consider eq. (2.17), the $\boldsymbol{R_{sq}}$ matrix is required. As the squared LD is used, a positive bias is induced into our $\boldsymbol{R_{sq}}$ matrix.

Based on Shieh (2010), one can correct for bias in the Pearson correlation ρ using

$$\rho = \rho\{1 + \frac{1 - \rho^2}{2(N - 4)}\} \tag{2.23}$$

where N is the number of sample used in the calculation of ρ. Similarly, there exists a bias correction equation for ρ^2:

$$\rho^2 = 1 - \frac{N - 3}{N - 2}(1 - \rho^2)\{1 + \frac{2(1 - \rho^2)}{N - 3.3}\} \tag{2.24}$$

Therefore, we corrected the $\boldsymbol{R_{sq}}$ based on eq. (2.24) such that the bias in estimation can be minimized.

2.2.6 Inverse of the Linkage Disequilibrium matrix

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

of $\hat{\boldsymbol{h^2}}$ from eq. (2.16). This will involve solving for

$$\hat{\boldsymbol{h^2}} = \boldsymbol{R_{sq}^{-1} f} \tag{2.25}$$

where an inverse of $\boldsymbol{R_{sq}}$ is observed.

In normal circumstances (e.g. when $\boldsymbol{R_{sq}}$ is full rank and positive semi-definite), one can easily solve eq. (2.25) using the QR decomposition or LU decomposition. However, when $\boldsymbol{R_{sq}}$ is ill-conditioned, the traditional decomposition method will fail. Even if the decomposition is successfully performed, the result tends to be a meaningless approximation to the true $\hat{\boldsymbol{h^2}}$.

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

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

$$\boldsymbol{A} = \boldsymbol{U\Sigma V}^t \tag{2.26}$$

where $\boldsymbol{U}$ and $\boldsymbol{V}$ are both orthogonal matrix and $\boldsymbol{\Sigma} = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n)$ is the diagonal matrix of the *singular values*(σ_i) of matrix $\boldsymbol{A}$. Based on eq. (2.26), we can get the inverse of $\boldsymbol{A}$ as

$$\boldsymbol{A}^{-1} = \boldsymbol{V\Sigma}^{-1}\boldsymbol{U}^t \tag{2.27}$$

Where $\boldsymbol{\Sigma}^{-1} = \text{diag}(\frac{1}{\sigma_1}, \frac{1}{\sigma_2}, \ldots, \frac{1}{\sigma_n})$. Now if we consider there to be error within $\boldsymbol{B}$ such that

$$\hat{\boldsymbol{B}}_{\boldsymbol{i}} = \boldsymbol{B}_{\boldsymbol{i}} + \epsilon_i \tag{2.28}$$

we can then represent $\boldsymbol{Ax} = \boldsymbol{B}$ as

$$\begin{aligned} \boldsymbol{Ax} &= \hat{\boldsymbol{B}} \\ \boldsymbol{U\Sigma V}^t\boldsymbol{x} &= \hat{\boldsymbol{B}} \\ \boldsymbol{x} &= \boldsymbol{V\Sigma}^{-1}\boldsymbol{U}^t\hat{\boldsymbol{B}} \end{aligned} \tag{2.29}$$

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

One can also observe from eq. (2.29) that when the singular value σ_i is small, the error ϵ_i in eq. (2.28) will be drastically magnified by a factor of $\frac{1}{\sigma_i}$. Making the system of equation highly sensitive to errors in the input.

To obtain a meaningful solution from this ill-conditioned/singular matrix $\boldsymbol{A}$, we may perform the

tSVD method to obtain a pseudo inverse of $\boldsymbol{A}$. Similar to eq. (2.26), the tSVD of $\boldsymbol{A}$ can be represented as

$$\boldsymbol{A}^{+} = \boldsymbol{U}\boldsymbol{\Sigma}_k\boldsymbol{V}^t \qquad \text{and} \qquad \boldsymbol{\Sigma}_k = \text{diag}(\sigma_1, \ldots, \sigma_k, 0, \ldots, 0) \tag{2.30}$$

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

$$\sigma_i = \begin{cases} \sigma_i & \text{for} \quad \sigma_i \geq t \\ 0 & \text{for} \quad \sigma_i < t \end{cases} \tag{2.31}$$

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

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

According to Hansen (1987), matrix where its rank has meaning will have well defined rank. As LD matrix is the correlation matrix between each individual SNPs, the rank of the LD matrix is the maximum number of linear independent SNPs in the region, therefore likely to have a well-defined rank. The easiest way to test whether if the threshold t and whether if the matrix $\boldsymbol{A}$ has well-defined rank is to calculate the "gap" in the singular value:

$$gap = \sigma_k/\sigma_{k+1} \tag{2.32}$$

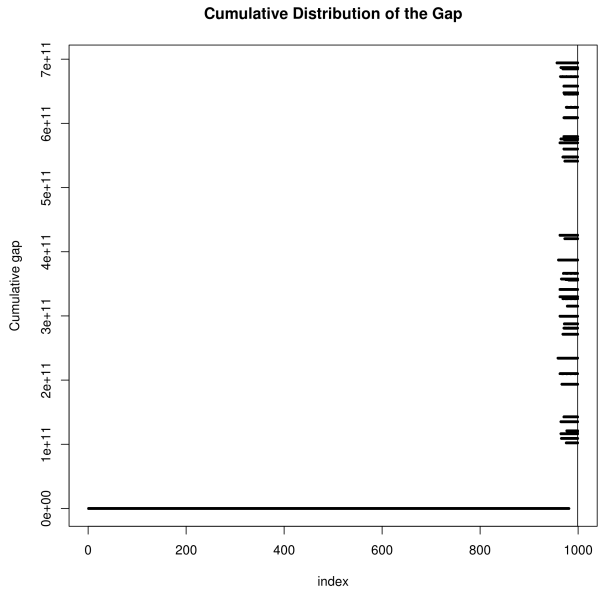
a large gap usually indicate a well-defined gap.

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

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

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

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



2.2.7 Comparing with LD SCore

2.3 Simulation

We implemented the heritability estimation in SNP Heritability and Risk Estimation Kit (SHREK) and in order to assess how well SHREK performs for heritability estimation in comparison to other current methods, we performed a series of systematic simulations. In these simulations, we compared the performance of SHREK with Genome-wide Complex Trait Analysis (GCTA)(Yang et al., 2011) and the LD SCore (LDSC)(Bulik-Sullivan et al., 2015) with and without the intercept estimation function (--no-intercept).

Through simulation, we can obtain the sample distribution of the heritability estimate under different study designs (e.g. Quantitativat traits, Case-Control studies or extreme phenotype selection). We can also evaluate the performance of different methods under varying genetic architecture (e.g. different number of Snps, different LD structures) or even with different disease models (e.g. different number of causal Snps, different heritability).

2.3.1 Quantitative Trait

One important factor to consider when carrying out a simulation is that the result of the simulation should be translatable to real life situation. Therefore, it is vital for us to consider as many different scenario as possible. When simulating a quantitative trait, there are a number of parameter for one to consider, for example, the sample size, the number of SNPs, the number of causal SNPs and the true heritability of the trait are all important parameters. However, it is also unrealistic for one to test the combination of all of these parameter as that will require a large amount of processing time. Thus, we aim to strike a balance

between comprehensive test case and a realistic simulation time.

First, although the average samples size for all current GWAS was $\sim$ 7,200 samples based on GWAS in the GWAS Catalog(Welter et al., 2014), we only used 1,000 samples in our simulation. We argue that if the tools were able to perform well when a small sample size were provided, then they should perform equally, if not better, when a larger sample size is given.

Secondly, we tried to simulate the complex LD structure in human population. Therefore, we used HAPGEN2(Su, Marchini, and Donnelly, 2011) with the 1000 genome Northern Europeans from Utah (CEU) population structure as an input to simulate samples with LD structure comparable to that in the 1000 genome CEU samples. Considering that it is unlikely for any SNPs between two chromosome to be in LD, we limit our simulation to chromosome 1 where there are a total of 670,052 SNPs information available for use in simulation. However, it was noted that as number of SNPs increase, the time required for simulating the samples and sample phenotype become prohibitive high. As a result of that, we limit the number of SNPs simulated to 50,000.

Trait complexity and trait heritability usually dictates the performance of heritability estimation. For example, if the trait is a Mendelian trait where there is a single causal SNP with large effect size, it will be relatively easy to calculate the trait's heritability. However, when a trait is polygenic with large amount of causal SNPs, each contribute a small portion of effect (e.g. schizophrenia (SCZ)), it will be challenging to estimate the heritability. We therefore varies the number of causal SNPs k with $k \in \{10, 50, 100, 1000\}$ such that different spectrum of trait complexity (e.g. oligogenic to polygenic) will be tested. One exception was Mendelian traits where we omitted from our simulation. It was because Mendelian traits usually associate with a single rare SNP with large effect size and high penetrance where its heritability can easily be estimated without the use of such complex algorithms.

Besides trait complexity, the trait heritability also dictates the performance of heritability estimation. We would therefore simulate traits with heritability H where $H \in [0, 1)$. Based on the work of Orr (1998), we modeled our per-SNP effect size to follow an exponential distribution with $\lambda = 1$, which serves as a heuristic expectation of the genetic architecture of adaptation. Taken into account of the number of causal SNPs and target heritability H, we then calculate the per-SNP effect size as

$$\begin{aligned} \beta_i &\sim exp(1) \\ \boldsymbol{\beta} &= (\beta_1, \beta_2, ..., \beta_k)^t \\ \boldsymbol{\gamma} &= \frac{H}{k}\boldsymbol{\beta} \end{aligned} \tag{2.33}$$

where $\boldsymbol{\gamma}$ is the vector of per SNP effect size and H is the simulated heritability. The final heritability of the simulated trait is defined as $H_{final} = \mathbf{1}^t\boldsymbol{\gamma}$

Another consideration is the SNP Minor Allele Frequency (maf), which tends to correlate with effect size(Manolio et al., 2009) due to selection. Rare SNPs with a small maf tend to have a large effect size whereas common SNPs tend to have a small effect size. Therefore, after the per SNP effects were simulated, we distribute the effect size to k randomly selected SNP(s) according to their maf.

Finally, by assuming $\boldsymbol{X}$ to be the standardized genotype of k causal SNPs in n samples, one can

get the phenotype of the simulated samples based on eq. (2.33) using

$$\begin{aligned}\epsilon_i &\sim N(0, \sqrt{\mathrm{Var}(\boldsymbol{X}\boldsymbol{\gamma})\frac{1-H_{final}}{H_{final}}})\\ \boldsymbol{\epsilon} &= (\epsilon_1, \epsilon_2, ..., \epsilon_n)^t\\ \boldsymbol{y} &= \boldsymbol{X}\boldsymbol{\gamma} + \boldsymbol{\epsilon}\end{aligned} \tag{2.34}$$

For each batch of simulated samples, we calculate the estimated heritability using SHREK, GCTA, LDSC with intercept fixed at 1 and LDSC allowing for intercept estimation for each H. In each iteration, the sample genotype was provided to GCTA for the calculation of genetic relationship matrix (GRM) whereas for SHREK and LDSC 500 additional samples were simulated based on the 1000 genome project CEU samples(Project et al., 2012) to construct the LD matrix and calculate the LD score respectively. This is because in general situations, LDSC and SHREK will not be provide with the sample genotype, instead, these programmes were designed to work with external LD reference data. Therefore to provide a realistic simulation, an independent set of reference samples were provided for SHREK and LDSC.

The whole process were repeated 50 times to obtain the empirical variance of the estimates,. In each iteration, new set of samples were simulated with the SNPs set, the causal SNPs and the per SNP effect size remain unchanged for each H.

To summarize,

1. Randomly select 50,000 SNPs from chromosome 1
2. Randomly generate k effect size following eq. (2.33) where $k \in \{10, 50, 100, 1000\}$
3. Randomly assign the effect size to k SNPs where SNPs with small maf will get a large effect size.
4. Simulate 1,000 samples using HAPGEN2 and calculate their phenotype according to eq. (2.34)
5. Perform heritability estimation using SHREK, LDSC and GCTA
6. Repeat step 4-5 50 times
7. Repeat step 1-6 50 times

2.3.2 Case Control Studies

Similar to quantitative trait simulation, the sample size, the number of SNPs, the number of causal SNPs and the true heritability of the trait are important parameters to consider during simulation. On top of that, there are a few more parameters one must consider during the simulation of case control studies such as the population prevalence of the trait and the observed prevalence of the study.

To simulate cases and controls, we will need to simulate the liability distribution by taking into acount of the prevalence. So for example, if one like to simulate a trait with population prevalence of p and observed prevalence of q and would like to have n cases in total, one will have to simulate $\min(\frac{n+pn}{p}, \frac{n+qn}{q})$ samples. It is therefore challenging for one to simulate scenario with a small p or q values.

In this study, we fixed $q = 0.5$ and varies $p \in \{0.5, 0.1\}$. Although disease such as SCZ can have a prevalence $\approx 1\%$, the required sample numbers become infeasible for large scale simulation where a minimum of 101,000 samples will be required if we wish to obtain 1,000 cases. Despite our wish to simulate conditions with small prevalence, the limitation of computation power simply forbade us to undergo such simulation.

Once the liability distribution were simulated, cases can be drawn from samples with a liability higher than the liability threshold. The liability threshold was calculated as the value $> 1 - p$ of all values under the standard normal distribution using the *qnorm* function in R. Samples with a liability lower than the liability threshold were then considered as control samples. 1,000 cases and 1,000 controls were then randomly drawn from the corresponding population of samples.

To summarize,

1. Randomly select 50,000 SNPs from chromosome 1
2. Randomly generate k effect size following eq. (2.33) where $k \in \{10, 50, 100, 1000\}$
3. Randomly assign the effect size to k SNPs where SNPs with small maf will get a large effect size.
4. Simulate $\frac{n+qn}{q}$ samples using HAPGEN2 where $q \in \{0.5, 0.1\}$
5. Simulate sample phenotype according to the liability threshold where 1,000 cases and 1,000 controls were obtained
6. Perform heritability estimation using SHREK, LDSC and GCTA
7. Repeat step 4-6 50 times
8. Repeat step 1-7 50 times

2.3.3 Exreme Phenotype Selections

The simulation of extreme phenotype selection is very much like the combination of quantitative trait simulation. The simplest way to simulate the extreme phenotype selection is to first simulation N samples, then from this population of samples, select n samples from both end of the population and use them to perform association and heritability estimation.

Due to the similarity in nature with the simulation of quantitative traits and case control studies, we limit the number of causal SNPs simulated as 100. We also limit our simulation to select samples with phenotype on the top and bottom 10% of the population.

However, it was noted that both GCTA and LDSC did not implement heritability estimation under extreme phenotype selection. To perform a fair comparison with SHREK, we calculate the adjustment factor $\frac{\text{Var}(Phenotype\ before\ selection)}{\text{Var}(Phenotype\ after\ selection)}$ to the estimation from GCTA and LDSC.

Therefore, to summarize,

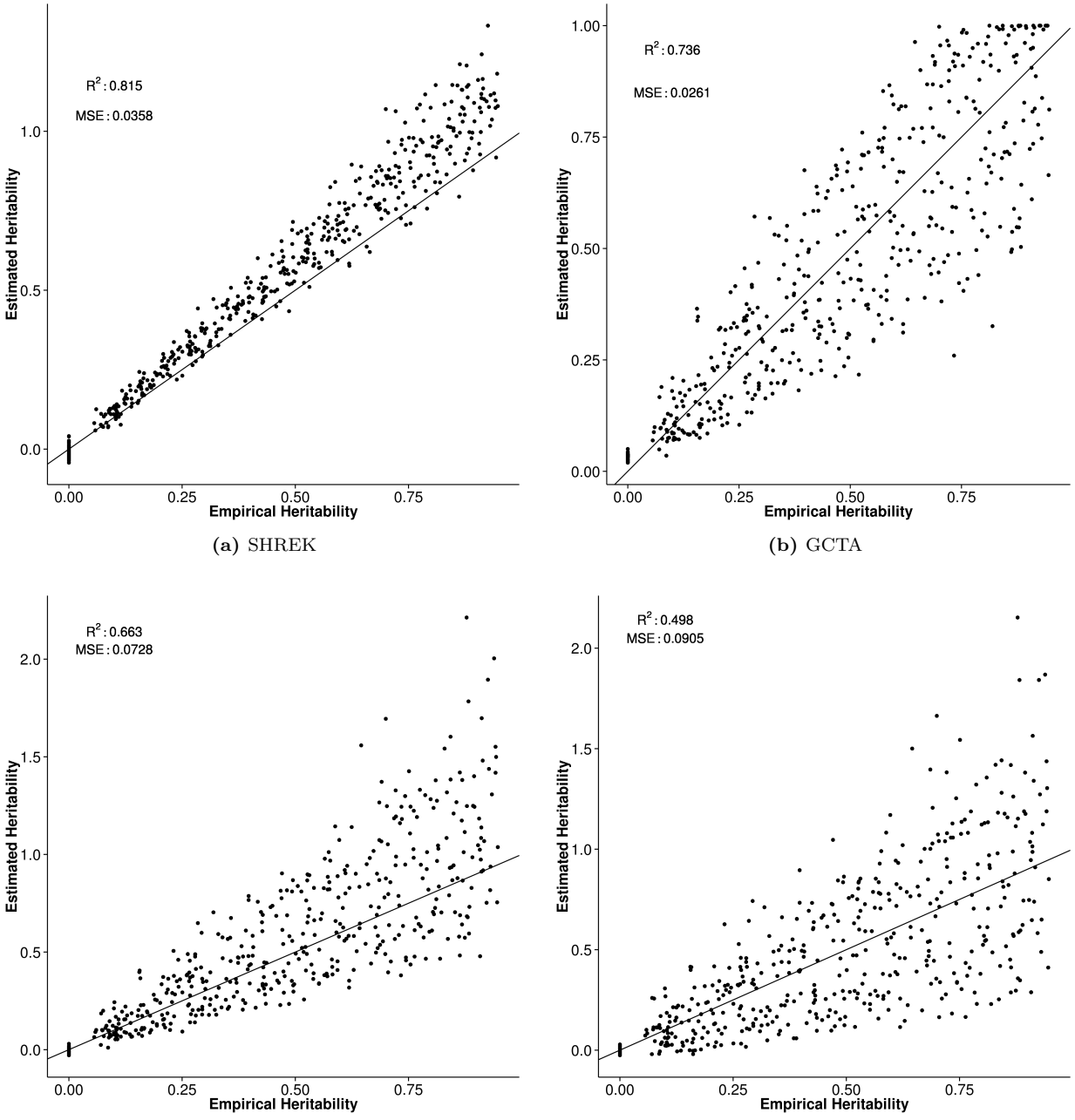
1. Randomly select 50,000 SNPs from chromosome 1

2. Randomly generate 100 effect size following eq. (2.33)
3. Randomly assign the effect size to k SNPs where SNPs with small maf will get a large effect size.
4. Simulate 10,000 samples using HAPGEN2 and calculate their phenotype according to eq. (2.34)
5. Select the top and bottom 10% of samples from the 10,000 samples.
6. Perform heritability estimation using SHREK, LDSC and GCTA
7. Manually apply the adjustment factor to estimation of GCTA and LDSC
8. Repeat step 4-7 50 times
9. Repeat step 1-8 50 times

2.4 Result

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

2.5 Discussion



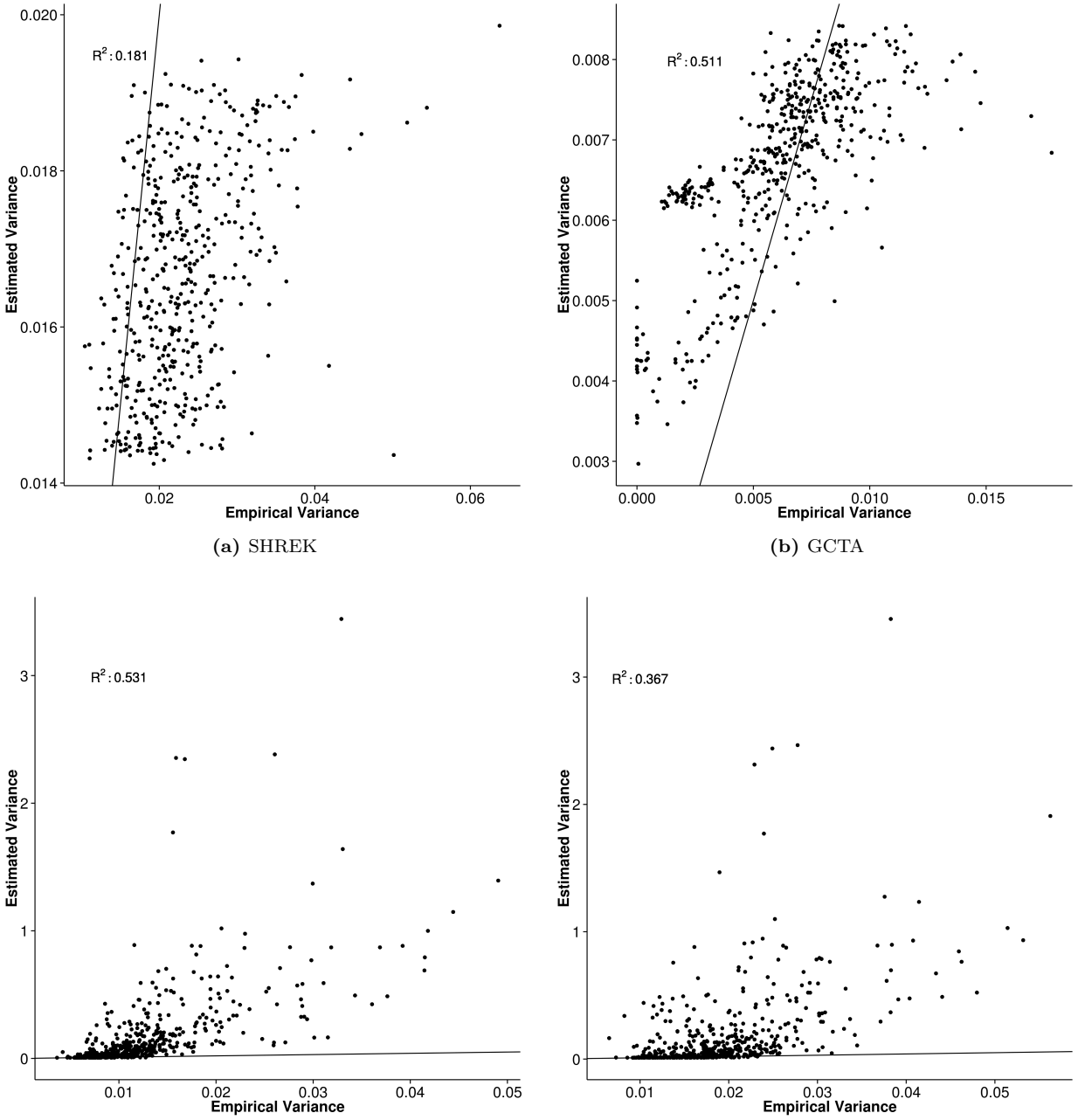
(a) SHREK

(b) GCTA

(c) LDSC with fix intercept

(d) LDSC with intercept estimation

Figure 2.2: Simulation of Quantitative Traits with 50k SNPs and 10 causal variants. The mean estimation of each round of simulation were plotted against the empirical heritability simulated. It was observed that SHREK has the highest R^2 and with mean squared error (MSE) lower than that of LDSC



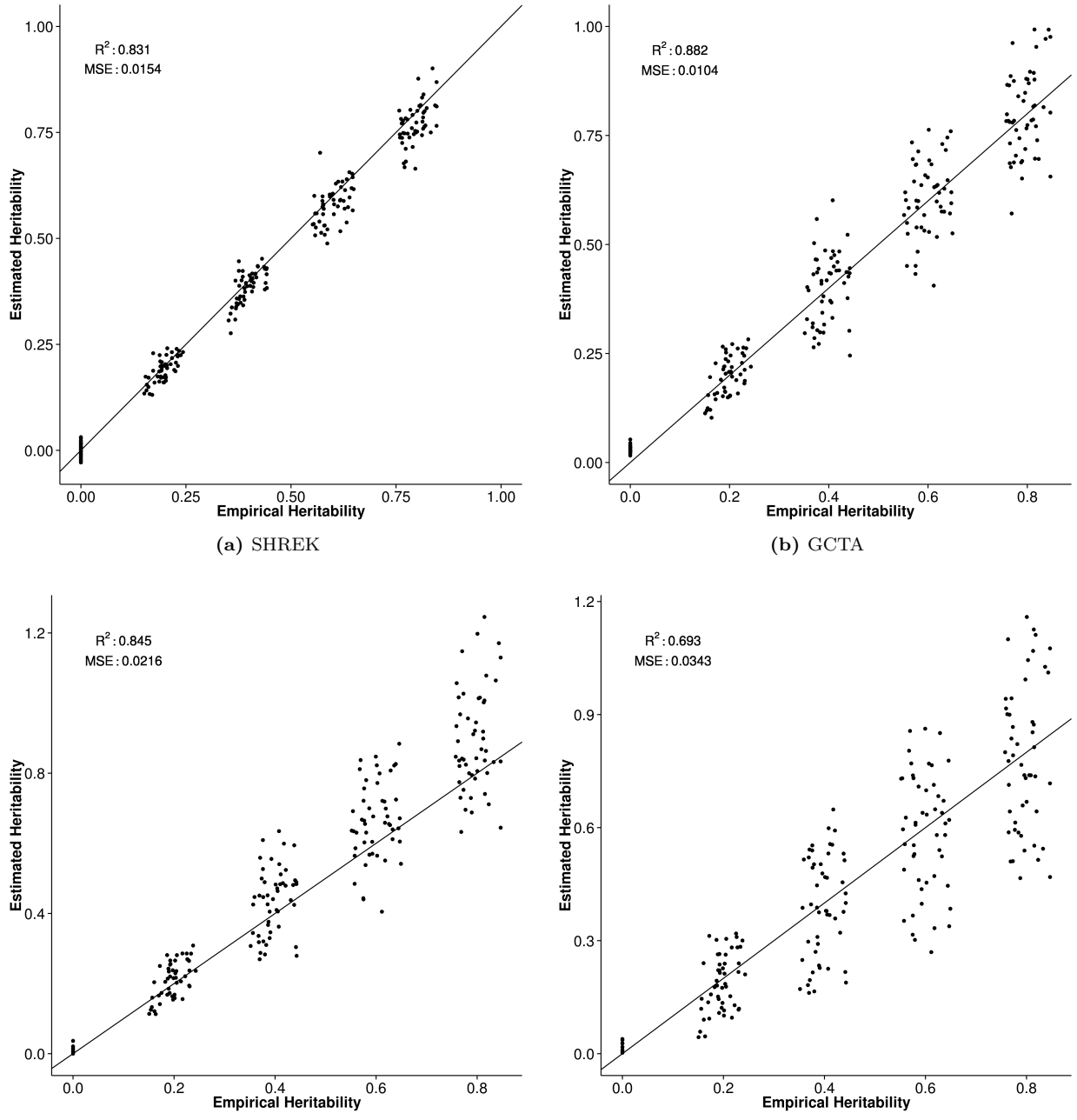
(a) SHREK

(b) GCTA

(c) LDSC with fix intercept

(d) LDSC with intercept estimation

Figure 2.3: Simulation of Quantitative Traits with 50k SNPs and 10 causal variants (Variance). When it come to the estimation of variance, it was observed that all the tools can only provide a heuristic estimation of the empirical variance, with GCTA being the most accurate and SHREK underestimating the variance. As for LDSC, the variance estimated were much higher than the empirical variance no matter if the intercept were fixed or not. It is worth noting that the empirical variance of GCTA were much lower than LDSC and SHREK whereas the empirical variance of SHREK tends to be higher than LDSC



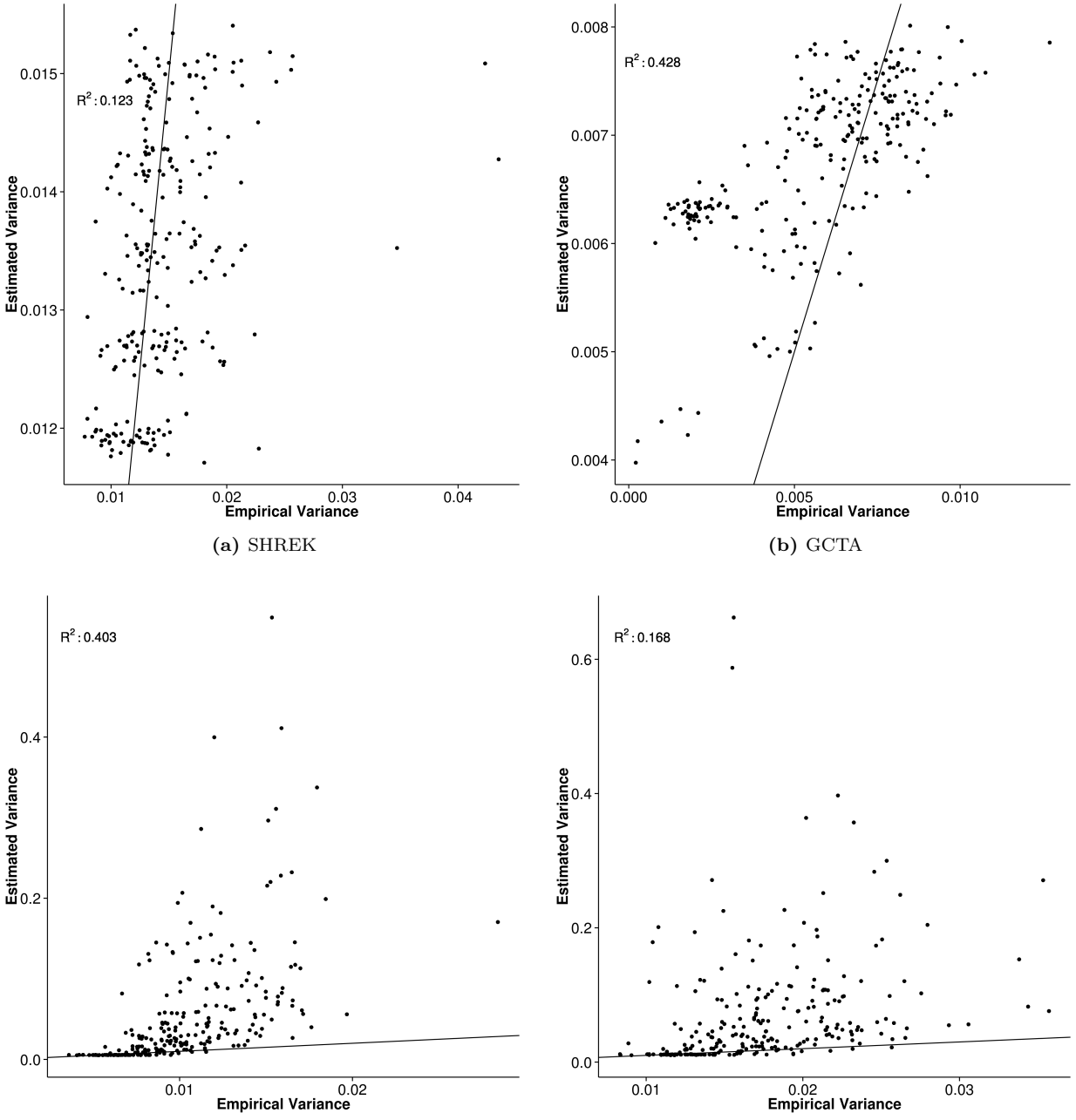
(a) SHREK

(b) GCTA

(c) LDSC with fix intercept

(d) LDSC with intercept estimation

Figure 2.4: Simulation of Quantitative Traits with 50k SNPs and 50 causal variants. The mean estimation of each round of simulation were plotted against the empirical heritability simulated. Under this condition, GCTA has the highest R^2, followed by SHREK. However, it was noted that SHREK was generally upwardly biased when compare to other tools. Also , the MSE of LDSC were substantially lower than in the condition of 10 causal SNPs.



(a) SHREK

(b) GCTA

(c) LDSC with fix intercept

(d) LDSC with intercept estimation

Figure 2.5: Simulation of Quantitative Traits with 50k SNPs and 50 causal variants (Variance). Again, all the tools can only provide a heuristic estimation of the empirical variance, with GCTA being the most accurate.

Chapter 3

Heritability of Schizophrenia

3.1 Introduction

Apply Heritability estimation to the schizophrenia data. The genetic correlation and partitioning of heritability No one worked on linking schizophrenia with brain development directly?

3.2 Heritability Estimation

This will be a very simple section, focused on how to perform the heritability estimation on schizophrenia (SCZ). Should also tokenize the heritability into subcategories (e.g. immune, neuron, etc)

3.2.1 Methodology

3.2.2 Result

3.3 Brain development and Schizophrenia

Here we will perform the WGCNA and brain development network. Seeing how the whether if any brain development network were enriched with SNPs that explain the variance of phenotype

3.3.1 Methodology

Sample Quality Controls

We obtain the developmental transcriptome data from BrainSpan (`http://www.brainspan.org/`). A total of 56 samples with different age were provided by BrainSpan with an average of 2.2 samples per age.

Studies suggested Hippocampus(Velakoulis et al., 2006; Nugent et al., 2007), Amygdala and Striatum(Simpson, Kellendonk, and Kandel, 2010) are brain regions involved in the etiology of schizophrenia. Therefore, we focus on building the gene co-expression network of hippocampus, amygdala and striatum in this study It is worth noting that the Pre-frontal Cortex is also important for schizophrenia. However, as there isn't a well defined pre-frontal cortex samples from BrainSpan, we did not include the pre-frontal cortex in the current study. RNA Sequencing data of the brain regions were obtained from BrainSpan and undergo a series of quality control before the construction of the network.

For each sample age, when there are more than one samples, we select the sample with a dissection score ≥ 3 and an RNA integrity number (RIN) ≥ 7. As some developmental stage only got 1 sample passing the quality check, we limit each developmental stage to have a maximum of 1 sample such that the final network will not be driven by a particular developmental stage. If multiple samples passed through the quality check threshold, we will prefer sample with higher dissection score. Shall multiple samples have the same dissection score, we will select the one with the highest RIN. And if the samples have the same dissection score and RIN value, we will randomly select one for the network construction.

After performing the quality control, a total of 16, 18 and 15 samples were selected for hippocampus, amygdala and striatum respectively. The sample age ranged from Gestation Day (GD)8 to 23 years old representing the fetal developmental stage till the age of onset of schizophrenia.

Normalization of data

The RNA Sequencing data were represented as Reads Per Kilobase per Million mapped reads (RPKM) values. Genes with a low RPKM can usually be a result from technical or biological noise(Hart et al., 2013). To reduce noise in the final model, genes with a mean RPKM < 1 in all samples were discarded. The RPKM were then log transformed as instructed by the manual of Weighted Gene Co-expression Network Analysis (WGCNA)(Langfelder and S Horvath, 2008).

As there are insufficient samples for the construction of gene co-expression network for individual sample age, we try to construct networks with genes co-expressed through all sample stage. This is achieved by taking the standardized $\log_2$ RPKM across sample age such that all genes has a mean of 0 and standard deviation of 1.

At the end, there were 17,168 genes, 17,038 genes and 17,166 genes passing through the quality threshold and were used for the construction of co-expression network in hippocampus, amygdala and striatum respectively.

Network Construction

WGCNA (ver 1.47) were used for the construction of gene co-expression network(Langfelder and S Horvath, 2008). The *blockwiseModules* function, using Biweight Midcorrelation for the construction of correlation matrix and a restriction of minimum network size of 30. For the construction of gene co-expression networks in hippocampus, the soft-power threshold were set to 15 where it is the first threshold value which has $R^2 > 0.8$ (0.817) and the R^2 is saturated(Zhang and Steve Horvath, 2005).As for striatum, the soft-power threshold were set to 20. Again, this is the first threshold value with $R^2 > 0.8$ (0.879) and where the R^2 is

saturated.

On the other hand, for amygdala, soft-power threshold were set to 9 which is the first threshold for R^2 to reach saturation. However, with a soft-power threshold of 9, the R^2 were only 0.776, which is lower than the recommended 0.8 threshold. The reason behind this decision was that the first soft-power threshold to have $R^2 > 0.8$ is 30. Under this threshold, the mean connectivity of the resulting networks will be around 23.6 with a median connectivity of 2.51. Such level of connectivity will likely yield networks that are too small to useful. If one would like to satisfy both requirement of threshold selection, a threshold > 30 are likely required and any networks constructed will likely to be small. As a result of that, we select threshold of 9 where networks with reasonable size can be constructed.

Expression correlation with Age

The co-expression network constructed with the standardized gene expression value will contains genes that co-express in all sample age. However, this does not necessary suggest the expression of these genes are correlated with the sample age. To identify gene co-expression networks with expressions correlated with the sample age, we performed a correlation analysis between the module eigen-genes and the sample age. Network eigen-genes were calculated as the first Principle Component (PC) of expressions of the genes within individual networks using the *moduleEigengenes* function from WGCNA. Age were represented as month from conception such that 8 post-conception week will be represented as 2; 4 months will be represented as 10 and 12 years will be represented as 154 etc. Finally, correlation between age and network eigen-gene expression were calculated pearson correlation.

Functional Annotation

Gene Ontology (GO) based enrichment analysis of the significant module was performed using GOrilla(Eden et al., 2009). Genes within the networks were provide as the target gene lists and all the genes passed quality controls were used as the background gene list. As GO terms tends to be redundant and overlaps with each other, it will aid the interpretation of GO results based by clustering and reducing the GO terms based on their similarity. Thus, GO enrichment results were summerized by REViGO(Supek et al., 2011) and significant representative GO terms were obtained.

Associate Co-expression network with PGC schizophrenia data

The co-expression networks were built from normal samples and should not be representative of the brain expression pattern in schizophrenia patients. it is however interesting to see if the co-expression networks were disrupted in schizophrenia patient. To test whether if the gene co-expression networks contain genes that are jointly associated with schizophrenia, we first use Multi-marker Analysis of GenoMic Annotation (MAGMA)(Leeuw et al., 2015)(version v1.03) to compute the gene-base p-value from the SNP wise p-value obtained from Psychiatric Genomics Consortium (PGC). Gene-set enrichment analysis were then performed on networks that were significantly correlated with developmental age. As we were only interested in whether if the genes within the networks were jointly associated with schizophrenia, we only focus on the result of the self-contained gene set analysis and ignore the result from competitive analysis.

Partitioning of Heritability

3.3.2 Result

Co-Expression Network

A total of 35 networks were constructed based on the hippocampus samples with a mean network size of 421.6. On the other hand, 28 networks were constructed for amygdala with mean network size of 591.86. Finally, 25 networks with mean size of 494.52 were constructed from the striatum samples.

Of the all the networks constructed, only one network from hippocampus(table 3.1a) and three networks from amygdala(table 3.1b) were significantly correlated with sample age after bonferroni correction threshold (p-value < 0.00143 for hippocampus, p-value < 0.00179 for amygdala and p-value < 0.002 for striatum) .

By plotting the mean expression of each network against the sample age, one can inspect how the dynamic of the network changes across different developmental stage. Thus, mean expression of all the genes within the significant networks were calculated for all amygdala (n=33) and hippocampus (n=32) samples from BrainSpan. The mean RPKM values were then $\log_2$ transformed and plot against the sample age where a line of bests fit was calculated using the *stat_smooth* with the loess function from R package *ggplot2*(version 1.0.1). (fig. 3.1).

The expression pattern observed were intriguing where there both the "black"(fig. 3.1a) and "tan" (fig. 3.1b) networks have mean gene expression level increase as development progress and reaches its peak at around late adolescence ($\approx 18 - 21$), concurring with the onset age of schizophrenia. Similarly, an inverse pattern were observed with the "yellow" network where its mean expression was highest during fetal development and drop steadily to its lowest around late adolescence and increase again afterwards(fig. 3.1d).

The expression pattern of the "black" and "tan" networks are of particular interest as they follow the inverted "U" shape trajectory of the grey matter volumn observed in previous studies(Gogtay et al., 2011), suggest that they might have a role in mediating brain development.

Functional Annotation

Upon performing the GO enrichment analysis, a total of 16 GO terms were enriched in the "black" hippocampus network, 4 in the "tan" amygdala network and 45 in the "yellow" amygdala network. No GO term was enriched in the "pink" amygdala network.

The enriched GO terms of the "yellow" amygdala network were mainly related to translation and transcription and were not specific to brain function or development(table S1). On the contrary, the GO terms enriched in the "black" hippocampus network were highly relevant to brain function and development (table 3.2)(e.g. "central nervous system development" and "glutamate metabolic process") and the "tan" amygdala network were also related to ammonium ion metabolism (table 3.3) which is vita for glutamine synthesis from glutamate(Liaw, Kuo, and Eisenberg, 1995).

Together, it is highly likely that the "black" hippocampus and "tan" amygdala networks are related

Table 3.1: Correlation of sample age with the module eigen gene. Module eigen-gene was defined as the first PC of genes within the module. After correcting for multiple testing, only the black module was considered as significantly correlated with the sample age.

(a) Hippocampus

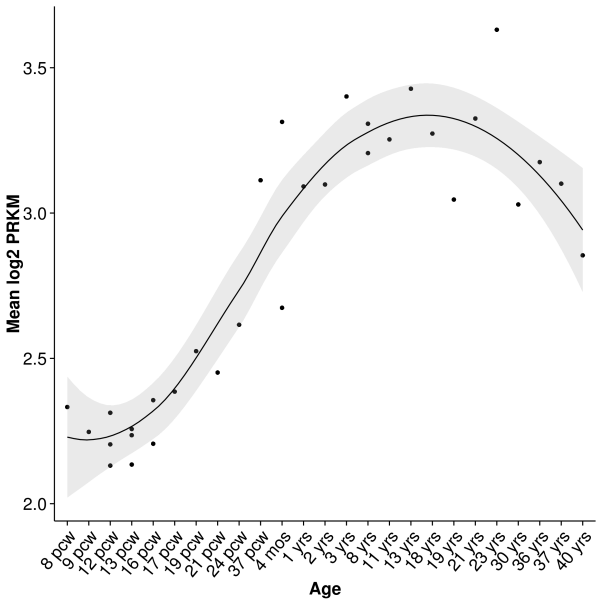
	Correlation	Pvalue
black	0.804653	0.000171
blue	-0.61648	0.010981
red	-0.60207	0.013595
darkred	-0.59137	0.015833
greenyellow	-0.56995	0.021168
yellow	0.567828	0.021763
darkgrey	-0.55246	0.026474
saddlebrown	-0.52983	0.034783
turquoise	-0.51371	0.041809
purple	-0.46788	0.067606
darkolivegreen	-0.41272	0.112122
sienna3	-0.39535	0.129604
darkturquoise	0.386541	0.139154
darkorange	0.384966	0.140912
darkmagenta	0.375586	0.151688
brown	0.366095	0.163144
tan	-0.36522	0.164229
pink	0.348979	0.18524
magenta	-0.32559	0.218473
midnightblue	-0.29168	0.273014
lightgreen	0.289921	0.276056
paleturquoise	-0.28045	0.29276
white	0.27727	0.29849
orange	0.19607	0.466754
steelblue	0.17355	0.520357
skyblue	0.145869	0.589857
lightyellow	-0.11665	0.667028
green	-0.09882	0.715786
violet	-0.08757	0.747076
lightcyan	-0.0656	0.809257
cyan	-0.06441	0.812661
darkgreen	-0.03914	0.885582
salmon	0.038727	0.886769
royalblue	-0.03785	0.889314
grey60	0.03119	0.908709

(b) Amygdala

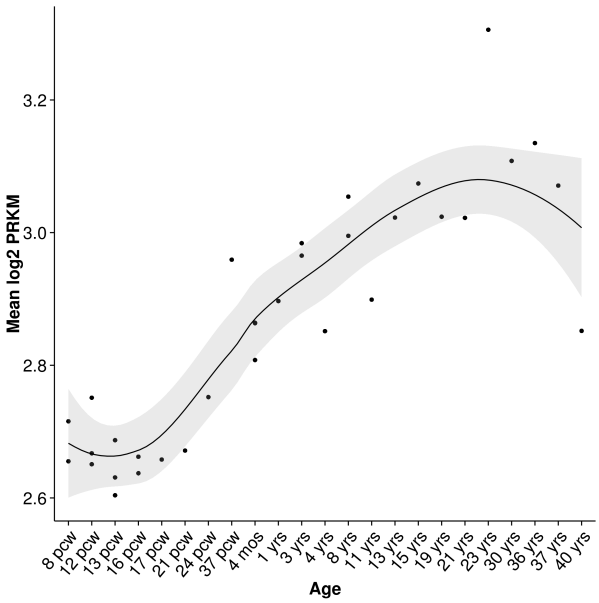
	Correlation	P-value
tan	0.849999	7.96×10^{-6}
yellow	-0.757	2.76×10^{-4}
pink	-0.68541	1.69×10^{-3}
greenyellow	-0.67831	1.97×10^{-3}
red	-0.64532	3.83×10^{-3}
turquoise	-0.59771	8.80×10^{-3}
lightyellow	-0.56347	0.0149
brown	0.548516	0.0184
darkgreen	-0.46366	0.0526
blue	-0.4604	0.0545
purple	-0.44182	0.0664
darkgrey	-0.39065	0.109
orange	-0.36966	0.131
white	0.28737	0.248
darkred	0.283247	0.255
black	0.271383	0.276
salmon	-0.24203	0.333
skyblue	0.207071	0.410
cyan	0.18778	0.456
lightgreen	0.166495	0.509
grey60	0.15156	0.548
midnightblue	0.136078	0.590
magenta	-0.13459	0.594
darkturquoise	0.129954	0.607
lightcyan	0.090241	0.722
darkorange	-0.05166	0.839
green	-0.04745	0.852
royalblue	0.020456	0.936

Figure 3.1: Mean Gene Experssion across developmental age. Mean RPKM values of genes in the significant modules were plot with respect to the sample age. A loess smoothing curve was also plotted.

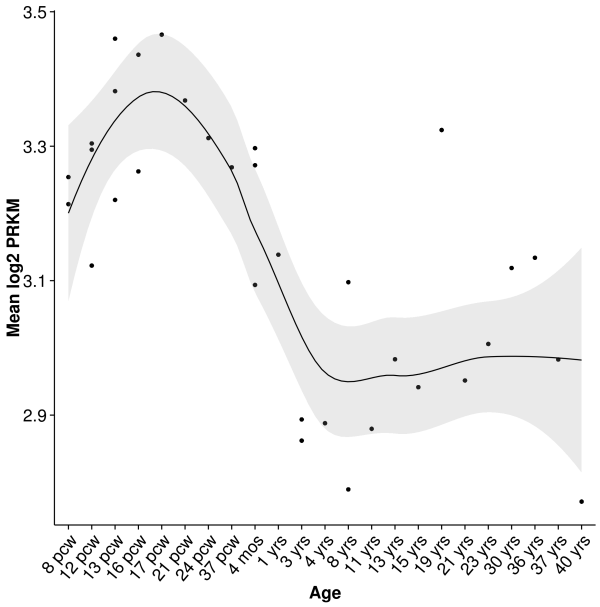
(a) "Black" Network from Hippocampus



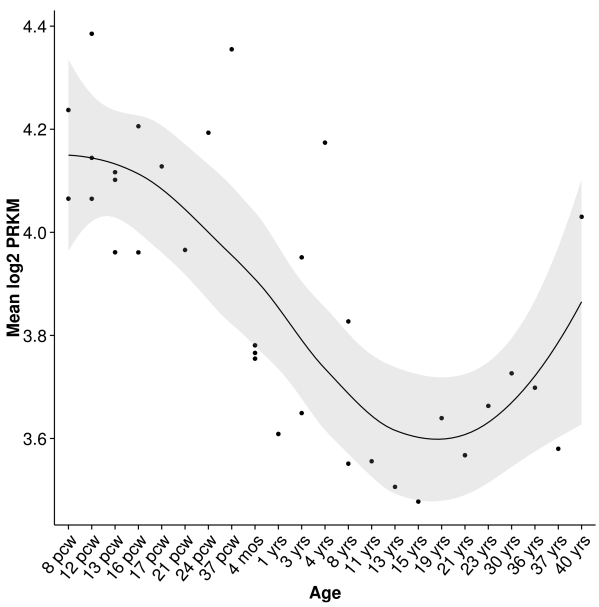
(b) "Tan" Network from Amygdala



(c) "Pink" Network from Amygdala



(d) "Yellow" Network from Amygdala



to brain development and function.

Table 3.2: GO enrichment results for the "black" network from Hippocampus. Among the enriched GO terms, it was most interesting to identify a number of brain developmental related GO terms such as "central nervous system development", "axon ensheathment in central nervous system", "glutamate metabolic process" and "positive regulation of gliogenesis". Surprisingly, GO related to immune systems were also observed "positive regulation of production of molecular mediator of immune response".

term_ID	description	p-value
GO:0019752	carboxylic acid metabolic process	4.92×10^{-6}
GO:0007417	central nervous system development	5.94×10^{-5}
GO:0002821	positive regulation of adaptive immune response	6.12×10^{-5}
GO:0006082	organic acid metabolic process	1.03×10^{-3}
GO:0032291	axon ensheathment in central nervous system	1.86×10^{-3}
GO:1901565	organonitrogen compound catabolic process	1.99×10^{-3}
GO:0006536	glutamate metabolic process	3.54×10^{-3}
GO:0021762	substantia nigra development	3.73×10^{-3}
GO:0044281	small molecule metabolic process	4.34×10^{-3}
GO:0030194	positive regulation of blood coagulation	4.59×10^{-3}
GO:0009607	response to biotic stimulus	6.14×10^{-3}
GO:0002702	positive regulation of production of molecular mediator of immune response	6.21×10^{-3}
GO:0034103	regulation of tissue remodeling	6.21×10^{-3}
GO:0014015	positive regulation of gliogenesis	7.47×10^{-3}
GO:0098542	defense response to other organism	7.95×10^{-3}
GO:0019835	cytolysis	8.72×10^{-3}

Table 3.3: GO enrichment results for the "tan" network from Amygdala. Unlike the "black" network, only a small number of GO terms were enriched. However, these GO terms are relatively specific to amine/ammonium ion metabolism. Interestingly, ammonium ion are essential to the synthesis of glutamine from glutamate, suggesting that this network might be relate to the glutamate system.

term_ID	description	p-value
GO:0097164	ammonium ion metabolic process	1.37×10^{-3}
GO:0044106	cellular amine metabolic process	4.2×10^{-3}
GO:0009308	amine metabolic process	5.41×10^{-3}
GO:0046519	sphingoid metabolic process	6.01×10^{-3}

Associate Co-expression network with PGC schizophrenia data

Although the co-expression network were extremely interesting for their expression pattern and functional enrichment in brain development and function related GO terms, there were no evidence of their involvement nor importance in schizophrenia. Therefore it is of particular interest for us to test whether if genes within these co-expression networks were associated withs schizophrenia.

First, gene base p-value of 18,622 genes were calculated using p-values from the PGC schizophrenia working group(Ripke et al., 2014). Gene set enrichment analysis were then performed using MAGMA(Leeuw et al., 2015) to test whether if there genes within the "black" hippocampus and "tan" amygdala networks were significantly associated with schizophrenia.

Based on the self-contained gene set enrichment analysis, genes within both networks were significantly associated with schizophrenia with p-value of 1.38×10^{-41} for the "tan" amygdala network and $2.70 \times 10{-}74$ for the "black" hippocampus network. These suggest that these networks might be disrupted in schizophrenia patients.

Partitioning of Heritability

3.4 Discussion

Chapter 4

Heritability of Response to antipsychotic treatment

Important to schizophrenia research

4.1 Introduction

Here we try to use Beatrice's data and estimate the heritability explained in drug response. Should also repeat the region-wise heritability

4.2 Methodology

4.3 Result

4.4 Discussion

Chapter 5

Risk Prediction

5.1 Methodology

We can define the traditional Polygenic Risk Score (PGS) as

$$\hat{Y} = diag(\beta)X \tag{5.1}$$

where X is the standardized genotype, β is the test-statistic calculated from other studies.

5.1.1 Simulation

5.2 Result

5.3 Discussion

Chapter 6

Conclusion

Bibliography

Altshuler, David M et al. (2010). "Integrating common and rare genetic variation in diverse human populations." In: *Nature* 467.7311, pp. 52–58. DOI: 10.1038/nature09298 (cit. on pp. 15, 17).

American Psychiatric Association (2013). *Diagnostic and Statistical Manual of Mental Disorders*. American Psychiatric Publishing, p. 991. DOI: 10.1176/appi.books.9780890425596.744053 (cit. on pp. 1, 2).

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. DOI: 10.1017/thg.2013.54 (cit. on p. 5).

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. DOI: 10.1038/ng.3211 (cit. on p. 18).

Eden, E et al. (2009). "GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists". eng. In: *BMC Bioinformatics* 10, p. 48. DOI: 10.1186/1471-2105-10-481471-2105-10-48[pii] (cit. on p. 29).

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. DOI: 10.1111/j.1469-1809.1965.tb00500.x (cit. on p. 6).

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

Gogtay, Nitin et al. (2011). "Age of onset of schizophrenia: Perspectives from structural neuroimaging studies". In: *Schizophrenia Bulletin* 37.3, pp. 504–513. DOI: 10.1093/schbul/sbr030 (cit. on p. 30).

Gottesman, II (1991). *Schizophrenia genesis: The origins of madness*. WH Freeman/Times Books/Henry Holt & Co (cit. on pp. 2, 6).

Gottesman, II and J Shields (1967). "A polygenic theory of schizophrenia". In: *Proceedings of the National Academy of Sciences* 58.1, pp. 199–205 (cit. on p. 7).

Guennebaud, Gaël, Beno\^it Jacob, et al. (2010). *Eigen v3*. http://eigen.tuxfamily.org (cit. on p. 16).

Hansen, Per Christian (1987). "The truncated SVD as a method for regularization". In: *Bit* 27.4, pp. 534–553. DOI: 10.1007/BF01937276 (cit. on p. 17).

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. DOI: 10.1038/sj.mp.4001686 (cit. on p. 7).

Hart, Traver et al. (2013). "Finding the active genes in deep RNA-seq gene expression studies." In: *BMC genomics* 14, p. 778. DOI: 10.1186/1471-2164-14-778 (cit. on p. 28).

Harvey, Philip D. et al. (2012). "Diagnosis of schizophrenia: Consistency across information sources and stability of the condition". In: *Schizophrenia Research* 140.1-3, pp. 9–14. DOI: 10.1016/j.schres.2012.03.026 (cit. on pp. 1, 2).

Jablensky, Assen (2010). "The diagnostic concept of schizophrenia: its history, evolution, and future prospects." In: *Dialogues in clinical neuroscience* 12.3, pp. 271–87 (cit. on p. 1).

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

Langfelder, P and S Horvath (2008). "WGCNA: an R package for weighted correlation network analysis". eng. In: *BMC Bioinformatics* 9, p. 559. DOI: 1471-2105-9-559[pii]10.1186/1471-2105-9-559 (cit. on p. 28).

Leeuw, Christiaan a. de et al. (2015). "MAGMA: Generalized Gene-Set Analysis of GWAS Data". In: *PLOS Computational Biology* 11.4, e1004219. DOI: 10.1371/journal.pcbi.1004219 (cit. on pp. 29, 33).

Li, Miao-Xin Xin et al. (2011). "Evaluating the effective numbers of independent tests and significant p-value thresholds in commercial genotyping arrays and public imputation reference datasets". In: *Human Genetics* 131.5, pp. 747–756. DOI: 10.1007/s00439-011-1118-2 (cit. on p. 15).

Liaw, S H, I Kuo, and D Eisenberg (1995). "Discovery of the ammonium substrate site on glutamine synthetase, a third cation binding site." In: *Protein science : a publication of the Protein Society* 4.11, pp. 2358–2365. DOI: 10.1002/pro.5560041114 (cit. on p. 30).

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. DOI: 10.1016/S0140-6736(09)60072-6 (cit. on p. 6).

Manolio, Teri a et al. (2009). "Finding the missing heritability of complex diseases." In: *Nature* 461.7265, pp. 747–753. DOI: 10.1038/nature08494 (cit. on p. 19).

Neumaier, Arnold (1998). "Solving Ill-Conditioned and Singular Linear Systems: A Tutorial on Regularization". In: *SIAM Review* 40.3, pp. 636–666. DOI: 10.1137/S0036144597321909 (cit. on p. 16).

Nugent, Tom F. et al. (2007). "Dynamic mapping of hippocampal development in childhood onset schizophrenia". In: *Schizophrenia Research* 90.1-3, pp. 62–70. DOI: 10.1016/j.schres.2006.10.014 (cit. on p. 28).

Orr, H Allen (1998). "The Population Genetics of Adaptation: The Distribution of Factors Fixed during Adaptive Evolution". In: *Evolution* 52.4, pp. 935–949 (cit. on p. 19).

Project, Genomes et al. (2012). "An integrated map of genetic variation from 1,092 human genomes". In: *Nature* 491.7422, pp. 56–65. DOI: http://www.nature.com/nature/journal/v491/n7422/abs/nature11632.html\#supplementary-information (cit. on pp. 15, 20).

Riley, Brien and Kenneth S Kendler (2006). "Molecular genetic studies of schizophrenia." In: *European journal of human genetics : EJHG* 14.6, pp. 669–680. DOI: 10.1038/sj.ejhg.5201571 (cit. on p. 7).

Ripke, Stephan et al. (2014). "Biological insights from 108 schizophrenia-associated genetic loci". In: *Nature* 511, pp. 421–427. DOI: 10.1038/nature13595 (cit. on p. 33).

Risch, N (1990). "Linkage strategies for genetically complex traits. I. Multilocus models." In: *American Journal of Human Genetics* 46.2, pp. 222–228 (cit. on p. 7).

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. DOI: 10.1001/archpsyc.64.10.1123 (cit. on p. 1).

Schultz, Stephen H., Stephen W. North, and Cleveland G. Shields (2007). "Schizophrenia: A review". In: *American Family Physician* 75.12, pp. 1821–1829 (cit. on p. 1).

Shieh, G (2010). "Estimation of the simple correlation coefficient". eng. In: *Behav Res Methods* 42.4, pp. 906–917. DOI: 10.3758/BRM.42.4.90642/4/906[pii] (cit. on p. 15).

Simpson, Eleanor H., Christoph Kellendonk, and Eric Kandel (2010). "A Possible Role for the Striatum in the Pathogenesis of the Cognitive Symptoms of Schizophrenia". In: *Neuron* 65.5, pp. 585–596. DOI: 10.1016/j.neuron.2010.02.014 (cit. on p. 28).

Su, Zhan, Jonathan Marchini, and Peter Donnelly (2011). "HAPGEN2: Simulation of multiple disease SNPs". In: *Bioinformatics* 27.16, pp. 2304–2305. DOI: 10.1093/bioinformatics/btr341 (cit. on p. 19).

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. DOI: 10.1001/archpsyc.60.12.1187 (cit. on p. 6).

Supek, F et al. (2011). "REVIGO summarizes and visualizes long lists of gene ontology terms". eng. In: *PLoS One* 6.7, e21800. DOI: 10.1371/journal.pone.0021800PONE-D-11-04111[pii] (cit. on p. 29).

Tsuang, Ming T., William S. Stone, and Stephen V. Faraone (2000). "Toward reformulating the diagnosis of schizophrenia". In: *American Journal of Psychiatry* 157.7, pp. 1041–1050. DOI: 10.1176/appi.ajp.157.7.1041 (cit. on p. 1).

Velakoulis, Dennis et al. (2006). "Hippocampal and amygdala volumes according to psychosis stage and diagnosis". In: *Archives of general psychiatry* 63, pp. 139–149 (cit. on p. 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. 3, 6).

Welter, Danielle et al. (2014). "The NHGRI GWAS Catalog, a curated resource of SNP-trait associations". In: *Nucleic Acids Research* 42.D1, pp. 1001–1006. DOI: 10.1093/nar/gkt1229 (cit. on p. 19).

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

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. DOI: 10.1016/j.ajhg.2010.11.011S0002-9297(10)00598-7[pii] (cit. on p. 18).

Zhang, Bin and Steve Horvath (2005). "A general framework for weighted gene co-expression network analysis." eng. In: *Statistical applications in genetics and molecular biology* 4.1, Article17. DOI: 10.2202/1544-6115.1128 (cit. on p. 28).

Supplementary Materials

Table S1: GO enrichment results for the "yellow" network from Amygdala. Most of the terms are related to translation transcription and are not specific enough to understand the true function of the network

term_ID	description	p-value
GO:0006614	SRP-dependent cotranslational protein targeting to membrane	1.63×10^{-47}
GO:0006412	translation	5.76×10^{-27}
GO:0006171	cAMP biosynthetic process	1.5×10^{-3}
GO:0042274	ribosomal small subunit biogenesis	3.13×10^{-8}
GO:0006413	translational initiation	2.12×10^{-28}
GO:0006414	translational elongation	7.63×10^{-31}
GO:0006415	translational termination	3.92×10^{-32}
GO:0071702	organic substance transport	4.16×10^{-10}
GO:0002181	cytoplasmic translation	1.19×10^{-7}
GO:1901566	organonitrogen compound biosynthetic process	6×10^{-17}
GO:0070972	protein localization to endoplasmic reticulum	5.36×10^{-44}
GO:0071822	protein complex subunit organization	2.21×10^{-6}
GO:0071826	ribonucleoprotein complex subunit organization	3.67×10^{-3}
GO:0022411	cellular component disassembly	1.55×10^{-17}
GO:1902578	single-organism localization	4.3×10^{-8}
GO:1902580	single-organism cellular localization	1.3×10^{-15}
GO:0000028	ribosomal small subunit assembly	1.91×10^{-7}
GO:0033036	macromolecule localization	1.74×10^{-11}
GO:0070727	cellular macromolecule localization	1.09×10^{-16}
GO:0006605	protein targeting	1.64×10^{-28}
GO:0022613	ribonucleoprotein complex biogenesis	3.32×10^{-9}
GO:0044765	single-organism transport	4.19×10^{-8}
GO:0044085	cellular component biogenesis	4.81×10^{-7}
GO:0061024	membrane organization	5.29×10^{-15}
GO:0051641	cellular localization	1.65×10^{-14}
GO:0044267	cellular protein metabolic process	6.57×10^{-5}
GO:0019083	viral transcription	1.64×10^{-44}
GO:1901564	organonitrogen compound metabolic process	1.15×10^{-14}
GO:0019538	protein metabolic process	4.17×10^{-5}
GO:0006401	RNA catabolic process	1.34×10^{-29}
GO:0010467	gene expression	1.4×10^{-11}
GO:0044419	interspecies interaction between organisms	4.74×10^{-15}
GO:0043604	amide biosynthetic process	4.35×10^{-26}
GO:0043603	cellular amide metabolic process	4.94×10^{-21}
GO:0044764	multi-organism cellular process	1.11×10^{-15}
GO:0016072	rRNA metabolic process	2.03×10^{-3}
GO:0016071	mRNA metabolic process	1.58×10^{-15}
GO:0000184	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	4.81×10^{-42}
GO:0034655	nucleobase-containing compound catabolic process	1.29×10^{-24}
GO:0044699	single-organism process	2.4×10^{-5}
GO:0051179	localization	7.96×10^{-8}
GO:0051704	multi-organism process	6.02×10^{-12}
GO:0071840	cellular component organization or biogenesis	3.83×10^{-3}
GO:0048871	multicellular organismal homeostasis	6.56×10^{-3}
GO:0009056	catabolic process	1.8×10^{-11}

Appendix