

# 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 14, 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. 17–19
- DSM-5** Diagnostic and Statistical Manual of Mental Disorders, 5th Edition. 7
- GCTA** Genome-wide Complex Trait Analysis. 18–21
- GD** Gestation Day. 24
- GO** Gene Ontology. 25, 26, 29, 40
- GWAS** Genome Wide Association Study. 9, 10, 18
- LD** Linkage Disequilibrium. 10, 13–19
- LDSC** LD SCore. 18–21
- maf** Minor Allele Frequency. 19–21
- MAGMA** Multi-marker Analysis of GenoMic Annotation. 25, 29
- PC** Principle Component. 25, 27
- PGC** Psychiatric Genomics Consortium. 25, 29
- PGS** Polygenic Risk Score. 33
- RIN** RNA integrity number. 24
- RPKM** Reads Per Kilobase per Million mapped reads. 24, 26, 28
- SCZ** Schizophrenia. 7, 18, 20, 23
- SE** Standard Error. 14
- SHREK** SNP Heritability and Risk Estimation Kit. 18–21
- SNP** Single Nucleotide Polymorphism. 9–11, 13, 17–21, 25
- SVD** Singular Value Decomposition. 16
- tSVD** Truncated Singular Value Decomposition. 15–17
- WGCNA** Weighted Gene Co-expression Network Analysis. 24, 25





# Contents

<b>Declaration</b>	<b>i</b>
<b>Acknowledgments</b>	<b>iii</b>
<b>Abbreviations</b>	<b>v</b>
<b>Contents</b>	<b>vii</b>
<b>1 Introduction</b>	<b>1</b>
<b>2 Literature Review I: Identifying the Genetic Contributions</b>	<b>3</b>
2.1 Twin and Adoption Studies . . . . .	3
2.2 Family Studies . . . . .	3
2.3 Genome Wide Association Studies . . . . .	3
2.4 Next Generation Sequencings Studies . . . . .	3
2.5 Narrow Sense Heritability . . . . .	3
2.5.1 Genome-wide Complex Trait Analysis . . . . .	3
2.5.2 Liability Threshold . . . . .	3
2.6 Risk Prediction . . . . .	3
2.6.1 Polygenic Scores . . . . .	3
<b>3 Literature Review II: Schizophrenia</b>	<b>5</b>
3.1 Disease Background . . . . .	5
3.1.1 Comorbidity . . . . .	5
3.2 Disease Mechanisms . . . . .	5
3.2.1 Twin Studies . . . . .	6
3.2.2 Epidemiological Studies . . . . .	6
3.2.3 Genetic Risk Factors . . . . .	6
3.2.4 Environmental Risk Factors . . . . .	6
3.3 Current Treatments . . . . .	6
3.3.1 Antipsychotics . . . . .	6
3.3.2 Pharmacogenetics and Pharmacogenomics . . . . .	6
<b>4 Heritability Estimation</b>	<b>7</b>
4.1 Introduction . . . . .	7
4.2 Methodology . . . . .	7
4.2.1 Heritability Estimation . . . . .	7
4.2.2 Calculating the Standard Error . . . . .	11
4.2.3 Case Control Studies . . . . .	13
4.2.4 Extreme Phenotype Selections . . . . .	13
4.2.5 Calculating the Linkage Disequilibrium matrix . . . . .	13
4.2.6 Inverse of the Linkage Disequilibrium matrix . . . . .	13
4.2.7 Comparing with LD SCore . . . . .	15

4.3	Simulation . . . . .	15
4.3.1	Quantitative Trait . . . . .	16
4.3.2	Case Control Studies . . . . .	18
4.3.3	Exreme Phenotype Selections . . . . .	18
4.4	Result . . . . .	19
4.5	Discussion . . . . .	19
<b>5</b>	<b>Heritability of Schizophrenia</b>	<b>23</b>
5.1	Introduction . . . . .	23
5.2	Heritability Estimation . . . . .	23
5.2.1	Methodology . . . . .	23
5.2.2	Result . . . . .	23
5.3	Brain development and Schizophrenia . . . . .	23
5.3.1	Methodology . . . . .	23
5.3.2	Result . . . . .	25
5.4	Discussion . . . . .	29
<b>6</b>	<b>Heritability of Response to antipsychotic treatment</b>	<b>31</b>
6.1	Introduction . . . . .	31
6.2	Methodology . . . . .	31
6.3	Result . . . . .	31
6.4	Discussion . . . . .	31
<b>7</b>	<b>Risk Prediction</b>	<b>33</b>
7.1	Methodology . . . . .	33
7.1.1	Simulation . . . . .	33
7.2	Result . . . . .	33
7.3	Discussion . . . . .	33
<b>8</b>	<b>Conclusion</b>	<b>35</b>

# List of Figures

4.1	Cumulative Distribution of “gap” of the LD matrix . . . . .	15
4.2	Simulation of Quantitative Traits with 50k SNPs and 10 causal variants . . . . .	20
4.3	Simulation of Quantitative Traits with 50k SNPs and 10 causal variants(Variance) . . . . .	21
5.1	Mean Gene Expression across developmental age . . . . .	28



# List of Tables

3.1	Top 20 leading cause of years lost due to disability calculated by WHO in year 2012 . . . . .	6
5.1	Correlation of sample age with the module eigen gene . . . . .	26
5.2	GO enrichment results for the “black” network from Hippocampus . . . . .	29
5.3	GO enrichment results for the “tan” network from Amygdala . . . . .	29
S1	GO enrichment results for the “yellow” network from Amygdala . . . . .	40



# Chapter 1

## Introduction





## Chapter 2

# Literature Review I: Identifying the Genetic Contributions

The concept of Literature Review is that they are things that happened *before* the studies in this thesis. They should be the disease background and what has happened so far. Also one should identify the research gap based on the current situation and therefore leads to the current study.

### 2.1 Twin and Adoption Studies

Should briefly talk about how Twin modeling was used for finding the GE contribution. Should also mention the ACE model.

### 2.2 Family Studies

### 2.3 Genome Wide Association Studies

### 2.4 Next Generation Sequencings Studies

### 2.5 Narrow Sense Heritability

#### 2.5.1 Genome-wide Complex Trait Analysis

#### 2.5.2 Liability Threshold

### 2.6 Risk Prediction

#### 2.6.1 Polygenic Scores



## Chapter 3

# Literature Review II: 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 delusions, hallucinations, disorganized speech, 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 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 3.1), demonstrating the extent of impact from Schizophrenia to patients.

### 3.1 Disease Background

Schizophrenia (SCZ) is a detrimental psychiatric disorder affecting around  $0.3 \sim 0.7\%$  of the population(American Psychiatric Association, 2013). SCZ was firsted name as “dementia praecox” by Dr. Emil Kraepelin and was later renamed as Schizophrenia by Dr. Eugen Bleuler(Jablensky, 2010). According to the Diagnostic and Statistical Manual of Mental Disorders, 5th Edition (DSM-5)(American Psychiatric Association, 2013), Schizophrenia is characterized by delusions, hallucinations, disorganized speech, grossly disorganized behavior and negative symptoms such as the diminished emotional expression. Delusions, hallucinations or the disorganized speech must be present for a significant portion of time during a 1-month period and there must be a continuous sign of symptoms for a 6-month period for one to be diagnosis with Schizophrenia.

#### 3.1.1 Comorbidity

### 3.2 Disease Mechanisms

Efforts have been made to identify the genetic cause of Schizophrenia.

**Table 3.1:** Top 20 leading cause of YLD calculated by WHO in year 2012. Schizophrenia is 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
<b>16</b>	<b>Schizophrenia</b>	<b>13,408</b>	<b>1.8</b>	<b>189</b>
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

### 3.2.1 Twin Studies

### 3.2.2 Epidemiological Studies

### 3.2.3 Genetic Risk Factors

Common genetic Variation

Rare Genetic Variation

### 3.2.4 Environmental Risk Factors

Maternal Immune Activation

## 3.3 Current Treatments

### 3.3.1 Antipsychotics

### 3.3.2 Pharmacogenetics and Pharmacogenomics

# Chapter 4

## Heritability Estimation

### 4.1 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 architecture. LD score regression!

### 4.2 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 Genome Wide Association Study (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 Miixin 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 Miixin 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.

#### 4.2.1 Heritability Estimation

The narrow-sense heritability is defined as

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

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

$$Y = \beta X + \epsilon \tag{4.1}$$

where  $Y$  and  $X$  are the standardized phenotype and genotype respectively.  $\epsilon$  is then the error term, accounting for the non-genetic elements contributing to the phenotype (e.g. Environment factors). Based on

eq. (4.1), one can then have

$$\begin{aligned}\text{Var}(Y) &= \text{Var}(\beta X) + \text{Var}(\epsilon) \\ \text{Var}(Y) &= \beta^2 \text{Var}(X) \\ \beta^2 \frac{\text{Var}(X)}{\text{Var}(Y)} &= 1\end{aligned}\tag{4.2}$$

$\beta^2$  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. (4.2). 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,  $\beta^2$  will be equal to the coefficient of determination ( $r^2$ ). Then, based on properties of the Pearson product-moment correlation coefficient:

$$r = \frac{t}{\sqrt{n-2+t^2}}\tag{4.3}$$

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. (4.3)

$$r^2 = \frac{t^2}{n-2+t^2}\tag{4.4}$$

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

When the effect size is small and  $n$  is big,  $r^2$  will be approximately  $\chi^2$  distributed with mean  $\sim 1$ . We can then approximate eq. (4.4) as

$$r^2 = \frac{\chi^2}{n}\tag{4.5}$$

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

$$f = \frac{\chi^2 - 1}{n}\tag{4.6}$$

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.

$$f_{observed} = f_{true} + f_{LD}\tag{4.7}$$

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

$$\begin{aligned}\mathbf{Y} &\sim f(0, 1) \\ \mathbf{X} &\sim f(0, \mathbf{R})\end{aligned}$$

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

We can then express eq. (4.1) in matrix form:

$$\mathbf{Y} = \beta^t \mathbf{X} + \epsilon\tag{4.8}$$

Definition of heritability will then become

$$\begin{aligned} \text{Heritability} &= \frac{\text{Var}(\boldsymbol{\beta}^t \mathbf{X})}{\text{Var}(\mathbf{Y})} \\ &= \text{Var}(\boldsymbol{\beta}^t \mathbf{X}) \end{aligned} \quad (4.9)$$

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

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

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

$$\begin{aligned} \text{Var}(\boldsymbol{\beta}^t \mathbf{X}) &= \text{E}_X \text{Var}_{\boldsymbol{\beta}|\mathbf{X}}(\mathbf{X}^t \boldsymbol{\beta}) + \text{Var}_X \text{E}_{(\boldsymbol{\beta}|\mathbf{X})}(\boldsymbol{\beta}^t \mathbf{X}) \\ &= \text{E}_X(\mathbf{X}^t \boldsymbol{\beta} \boldsymbol{\beta}^T \mathbf{X}) \\ &= \text{E}_X(\mathbf{X}^t \mathbf{H} \mathbf{X}) \\ &= \text{E}(\mathbf{X})^t \mathbf{H} \text{E}(\mathbf{X}) + \text{Tr}(\text{Var}(\mathbf{X} \mathbf{H})) \\ &= \text{Tr}(\text{Var}(\mathbf{X} \mathbf{H})) \\ &= \sum_i h_i^2 \end{aligned} \quad (4.10)$$

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

$$\begin{aligned} \text{Cov}(\mathbf{X}_i, \mathbf{Y}) &= \text{Cov}(\mathbf{X}_i, \boldsymbol{\beta}^t \mathbf{X} + \epsilon) \\ &= \text{Cov}(\mathbf{X}_i, \boldsymbol{\beta}^t \mathbf{X}) \\ &= \sum_j \text{Cov}(\mathbf{X}_i, \mathbf{X}_j) \boldsymbol{\beta}_j \\ &= \mathbf{R}_i \boldsymbol{\beta}_j \end{aligned} \quad (4.11)$$

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

$$\rho_i = \mathbf{R}_i \boldsymbol{\beta}_j \quad (4.12)$$

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

$$\hat{\rho}_i = \rho_i + \frac{\epsilon_i}{\sqrt{n}} \quad (4.13)$$

for some error  $\epsilon_i$ . The distribution of the correlation coefficient about the true correlation  $\rho$  is approximately

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

By making the assumption that  $\rho_i$  is close to 0 for all  $i$ , we have

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

We then define our  $z$ -statistic and  $\chi^2$ -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. (4.13) and eq. (4.12),  $\chi^2$  can then be expressed as

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

The expectation of  $\chi^2$  is then

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

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

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

If we define

$$f_i = \frac{\chi_i^2 - 1}{n} \tag{4.14}$$

we got

$$\begin{aligned} \sum_i (\chi_i^2 - E(\chi_i^2))^2 &= \sum_i (f_i - \mathbf{R}_i \mathbf{H} \mathbf{R}_i)^2 \\ &= \mathbf{f} \mathbf{f}^t - 2\mathbf{f}^t \mathbf{R}_{sq} \hat{\mathbf{h}} + \hat{\mathbf{h}}^t \mathbf{R}_{sq}^t \mathbf{R}_{sq} \hat{\mathbf{h}} \end{aligned} \tag{4.15}$$

where  $\mathbf{R}_{sq} = \mathbf{R} \circ \mathbf{R}$ . By differentiating eq. (4.15) w.r.t  $\hat{\mathbf{h}}$  and set to 0, we get

$$\begin{aligned} 2\mathbf{R}_{sq}^t \mathbf{R}_{sq} \hat{\mathbf{h}} - 2\mathbf{R}_{sq} \mathbf{f} &= 0 \\ \mathbf{R}_{sq} \hat{\mathbf{h}} &= \mathbf{f} \end{aligned} \tag{4.16}$$

And the heritability is then defined as

$$Heritability = \mathbf{1}^t \mathbf{R}_{sq}^{-1} \mathbf{f} \tag{4.17}$$



### 4.2.2 Calculating the Standard Error

From eq. (4.17), we can derive the variance of heritability  $H$  as

$$\begin{aligned}
\text{Var}(H) &= E[H^2] - E[H]^2 \\
&= E[(\mathbf{1}^t \mathbf{R}_{sq}^{-1} \mathbf{f})^2] - E[\mathbf{1}^t \mathbf{R}_{sq}^{-1} \mathbf{f}] (E[\mathbf{1}^t \mathbf{R}_{sq}^{-1} \mathbf{f}])^t \\
&= E[\mathbf{1}^t \mathbf{R}_{sq}^{-1} \mathbf{f} \mathbf{f}^t \mathbf{R}_{sq}^{-1} \mathbf{1}] - E[\mathbf{1}^t \mathbf{R}_{sq}^{-1} \mathbf{f}] (E[\mathbf{1}^t \mathbf{R}_{sq}^{-1} \mathbf{f}])^t \\
&= \mathbf{1}^t \mathbf{R}_{sq}^{-1} E[\mathbf{f} \mathbf{f}^t] \mathbf{R}_{sq}^{-1} \mathbf{1} - E[\mathbf{1}^t \mathbf{R}_{sq}^{-1} \mathbf{f}] (E[\mathbf{1}^t \mathbf{R}_{sq}^{-1} \mathbf{f}])^t \\
&= \mathbf{1}^t \mathbf{R}_{sq}^{-1} \text{Var}(\mathbf{f}) \mathbf{R}_{sq}^{-1} \mathbf{1} + E[\mathbf{1}^t \mathbf{R}_{sq}^{-1} \mathbf{f}] (E[\mathbf{1}^t \mathbf{R}_{sq}^{-1} \mathbf{f}])^t - E[\mathbf{1}^t \mathbf{R}_{sq}^{-1} \mathbf{f}] (E[\mathbf{1}^t \mathbf{R}_{sq}^{-1} \mathbf{f}])^t \\
&= \mathbf{1}^t \mathbf{R}_{sq}^{-1} \text{Var}(\mathbf{f}) \mathbf{R}_{sq}^{-1} \mathbf{1}
\end{aligned} \tag{4.18}$$

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

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

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

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

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

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

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

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

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

By definition,

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

We can then calculate  $E[z_i^2 z_j^2]$  as

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

As a result, the variance covariance matrix of the  $\chi^2$  variances represented as

$$\text{Cov}(X_i^2, X_j^2) = 2R_{ij}^2 + 4R_{ij}\mu_i\mu_j \quad (4.19)$$

Considering that we only have the *observed* expectation, we should re-define eq. (4.19) as

$$\text{Cov}(X_i^2, X_j^2) = \frac{2R_{ij}^2 + 4R_{ij}\mu_i\mu_j}{n^2} \quad (4.20)$$

where  $n$  is the sample size.

By substituting eq. (4.20) into eq. (4.18), we will get

$$\text{Var}(H) = \mathbf{1}^t \mathbf{R}_{sq}^{-1} \frac{2\mathbf{R}_{sq} + 4\mathbf{R} \circ \mathbf{z} \mathbf{z}^t}{n^2} \mathbf{R}_{sq}^{-1} \mathbf{1} \quad (4.21)$$

where  $\mathbf{z} = \sqrt{\chi^2}$  from eq. (4.14), with the direction of effect as its sign and  $\circ$  is the element-wise product (Hadamard product).

Problem with eq. (4.21) 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  $\mathbf{f}$  is approximately  $\chi^2$  distributed. Therefore eq. (4.16) can be viewed as a decomposition of a vector of  $\chi^2$  distributions with degree of freedom of 1. Replacing the vector  $\mathbf{f}$  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  $\chi^2$  distribution will yield

$$\text{Var}(H) = \frac{2(e + 2H)}{n^2} \quad (4.22)$$

eq. (4.22) will gives us an heuristic estimation of the SE.

### 4.2.3 Case Control Studies

### 4.2.4 Extreme Phenotype Selections

eq. (4.17) can naturally be applied to the quantitative trait scenario.

### 4.2.5 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. (4.17), the  $\mathbf{R}_{sq}$  matrix is required. As the squared LD is used, a positive bias is induced into our  $\mathbf{R}_{sq}$  matrix.

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

$$\rho = \rho \left\{ 1 + \frac{1 - \rho^2}{2(N - 4)} \right\} \quad (4.23)$$

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

$$\rho^2 = 1 - \frac{N - 3}{N - 2} (1 - \rho^2) \left\{ 1 + \frac{2(1 - \rho^2)}{N - 3.3} \right\} \quad (4.24)$$

Therefore, we corrected the  $\mathbf{R}_{sq}$  based on eq. (4.24) such that the bias in estimation can be minimized.

### 4.2.6 Inverse of the Linkage Disequilibrium matrix

In order to obtain the heritability estimation, we will require to solve eq. (4.17). If  $\mathbf{R}_{sq}$  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. (4.17) as calculating the sum of  $\hat{\mathbf{h}}^2$  from eq. (4.16). This will involve solving for

$$\hat{\mathbf{h}}^2 = \mathbf{R}_{sq}^{-1} \mathbf{f} \quad (4.25)$$

where an inverse of  $\mathbf{R}_{sq}$  is observed.

In normal circumstances (e.g. when  $\mathbf{R}_{sq}$  is full rank and positive semi-definite), one can easily solve eq. (4.25) using the QR decomposition or LU decomposition. However, when  $\mathbf{R}_{sq}$  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  $\hat{\mathbf{h}}^2$ .

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  $\mathbf{A}\mathbf{x} = \mathbf{B}$  where  $\mathbf{A}$  is ill-conditioned or singular with  $n \times n$  dimension. The SVD of  $\mathbf{A}$  can

be expressed as

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^t \quad (4.26)$$

where  $\mathbf{U}$  and  $\mathbf{V}$  are both orthogonal matrix and  $\mathbf{\Sigma} = \text{diag}(\sigma_1, \sigma_2, \dots, \sigma_n)$  is the diagonal matrix of the *singular values*( $\sigma_i$ ) of matrix  $\mathbf{A}$ . Based on eq. (4.26), we can get the inverse of  $\mathbf{A}$  as

$$\mathbf{A}^{-1} = \mathbf{V}\mathbf{\Sigma}^{-1}\mathbf{U}^t \quad (4.27)$$

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

$$\hat{\mathbf{B}}_i = \mathbf{B}_i + \epsilon_i \quad (4.28)$$

we can then represent  $\mathbf{A}\mathbf{x} = \mathbf{B}$  as

$$\begin{aligned} \mathbf{A}\mathbf{x} &= \hat{\mathbf{B}} \\ \mathbf{U}\mathbf{\Sigma}\mathbf{V}^t\mathbf{x} &= \hat{\mathbf{B}} \\ \mathbf{x} &= \mathbf{V}\mathbf{\Sigma}^{-1}\mathbf{U}^t\hat{\mathbf{B}} \end{aligned} \quad (4.29)$$

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

One can also observe from eq. (4.29) that when the singular value  $\sigma_i$  is small, the error  $\epsilon_i$  in eq. (4.28) 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  $\mathbf{A}$ , we may perform the tSVD method to obtain a pseudo inverse of  $\mathbf{A}$ . Similar to eq. (4.26), the tSVD of  $\mathbf{A}$  can be represented as

$$\mathbf{A}^+ = \mathbf{U}\mathbf{\Sigma}_k\mathbf{V}^t \quad \text{and} \quad \mathbf{\Sigma}_k = \text{diag}(\sigma_1, \dots, \sigma_k, 0, \dots, 0) \quad (4.30)$$

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

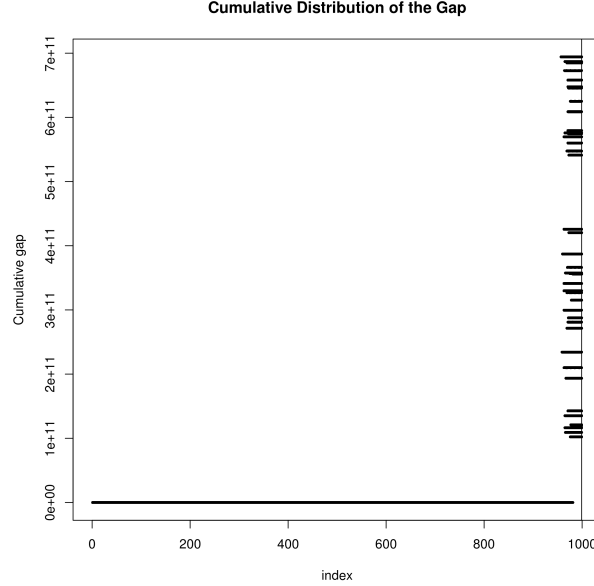
$$\sigma_i = \begin{cases} \sigma_i & \text{for } \sigma_i \geq t \\ 0 & \text{for } \sigma_i < t \end{cases} \quad (4.31)$$

where  $t$  is the tolerance threshold. Any singular value  $\sigma_i$  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  $\mathbf{x}$ . A problem with tSVD however is that it only work when matrix  $\mathbf{A}$  has a well determined numeric rank(Hansen, 1987). That is, tSVD work best when there is a large gap between  $\sigma_k$  and  $\sigma_{k+1}$ . 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  $\sigma_{k+1}$  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  $\mathbf{A}$  has well-defined rank is to calculate the

**Figure 4.1:** 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



“gap” in the singular value:

$$gap = \sigma_k / \sigma_{k+1} \quad (4.32)$$

a large gap usually indicate a well-defined gap.

In this study, we adopt the threshold as defined in MATLAB, NumPy and GNU Octave:  $t = \epsilon \times \max(m, n) \times \max(\Sigma)$  where  $\epsilon$  is the machine epsilon (the smallest number a machine can define as non-zero). And we performed 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. 4.1). 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. (4.16), and obtain the estimated heritability.

#### 4.2.7 Comparing with LD SCore

### 4.3 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 perfor-

mance 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. Quantitative 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).

### 4.3.1 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. 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

$$\begin{aligned}\beta_i &\sim \exp(1) \\ \boldsymbol{\beta} &= (\beta_1, \beta_2, \dots, \beta_k)^t \\ \boldsymbol{\gamma} &= \frac{H}{k} \boldsymbol{\beta}\end{aligned}\tag{4.33}$$

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  $\mathbf{X}$  to be the standardized genotype of  $k$  causal SNPs in  $n$  samples, one can get the phenotype of the simulated samples based on eq. (4.33) using

$$\begin{aligned}\epsilon_i &\sim N(0, \sqrt{\text{Var}(\mathbf{X}\boldsymbol{\gamma}) \frac{1 - H_{final}}{H_{final}}}) \\ \boldsymbol{\epsilon} &= (\epsilon_1, \epsilon_2, \dots, \epsilon_n)^t \\ \mathbf{y} &= \mathbf{X}\boldsymbol{\gamma} + \boldsymbol{\epsilon}\end{aligned}\tag{4.34}$$

For each batch of simulated samples, we calculate the estimated heritability using SHREK, GCTA, LDSC without intercept estimation and LDSC with intercept estimate ( $LDSC_{In}$ ) 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. (4.33) 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. (4.34)
5. Perform heritability estimation using SHREK, LDSC and GCTA
6. Repeat step 4-5 50 times
7. Repeat step 1-6 50 times

### 4.3.2 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 account 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. (4.33) 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

### 4.3.3 Extreme 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  $\frac{\text{Var}(\text{Phenotype before selection})}{\text{Var}(\text{Phenotype after selection})}$  to the estimation from GCTA and LDSC.

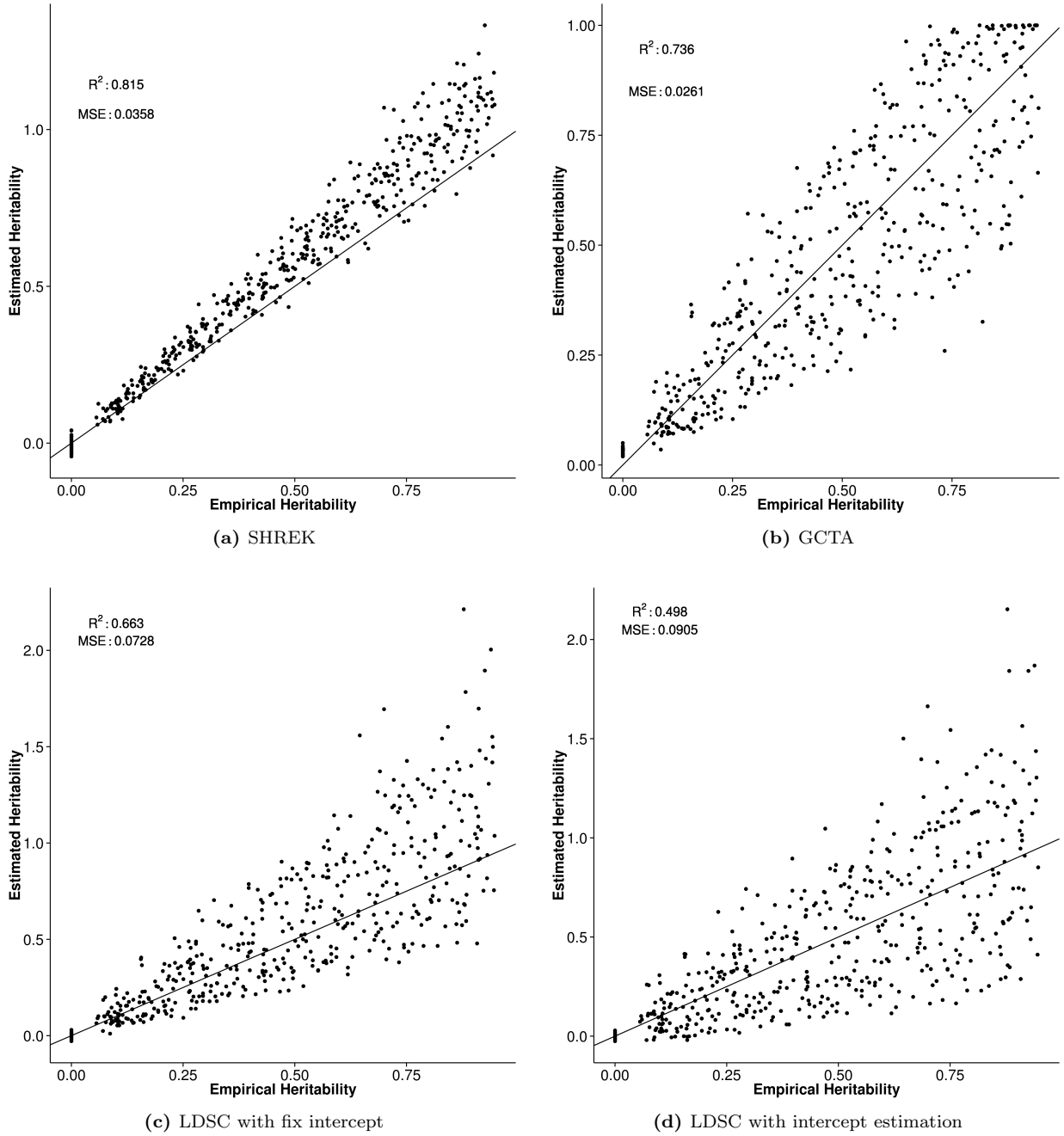
Therefore, to summarize,

1. Randomly select 50,000 SNPs from chromosome 1
2. Randomly generate 100 effect size following eq. (4.33)
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. (4.34)
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

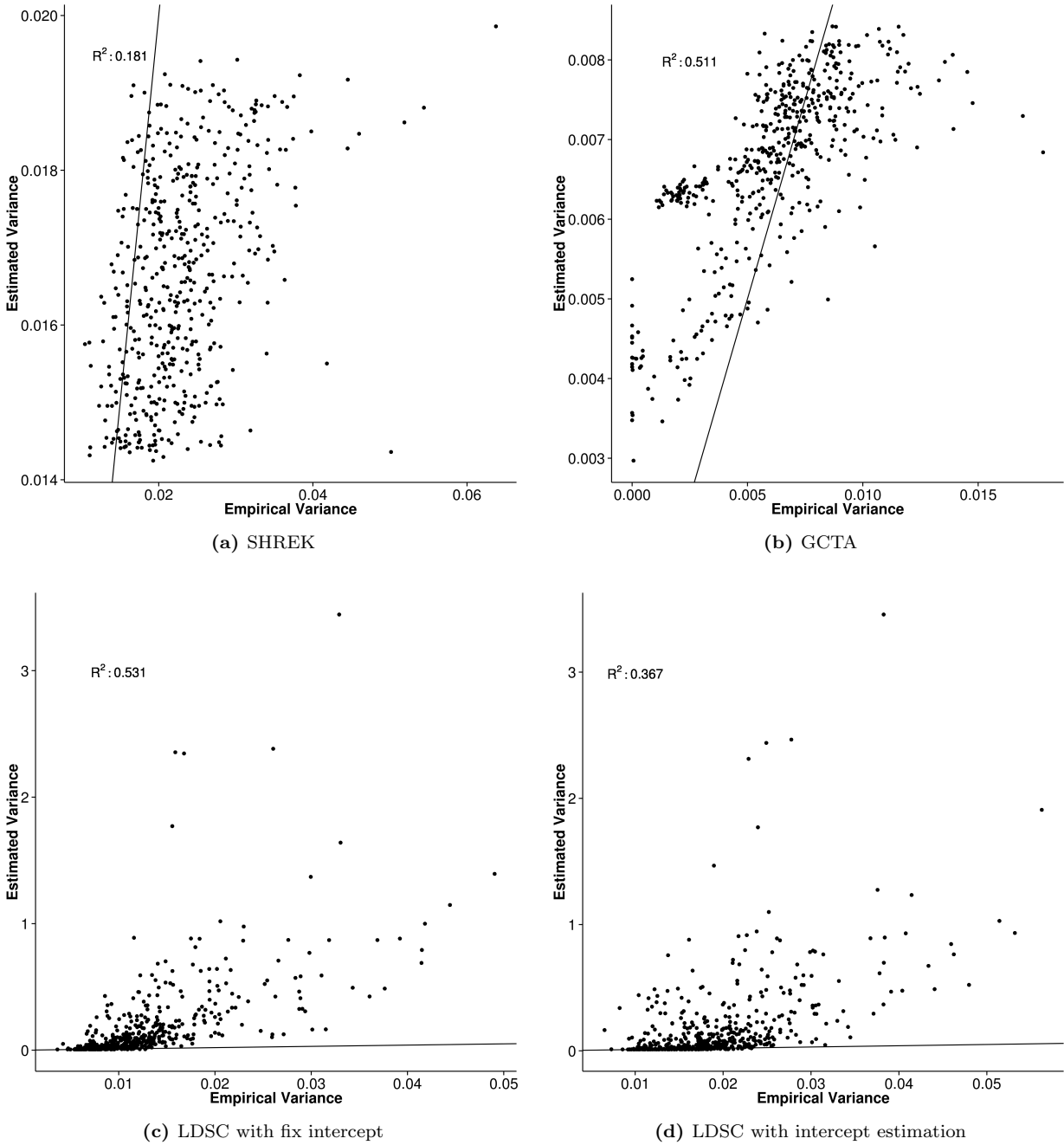
## 4.4 Result

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

## 4.5 Discussion



**Figure 4.2:** 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



**Figure 4.3:** 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



## Chapter 5

# Heritability of Schizophrenia

### 5.1 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?

### 5.2 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)

#### 5.2.1 Methodology

#### 5.2.2 Result

### 5.3 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

#### 5.3.1 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(Nugent et al., 2007; Velakoulis et al., 2006), 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  $\geq 3$  and an RNA integrity number (RIN)  $\geq 7$ . 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 be 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

## 5.3.2 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 5.1a) and three networks from amygdala(table 5.1b) were significantly correlated with sample age after bonferroni correction threshold ( $p\text{-value} < 0.00143$  for hippocampus,  $p\text{-value} < 0.00179$  for amygdala and  $p\text{-value} < 0.002$  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

**Table 5.1:** 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			(b) Amygdala		
	Correlation	Pvalue		Correlation	P-value
black	0.804653	0.000171	tan	0.849999	$7.96 \times 10^{-6}$
blue	-0.61648	0.010981	yellow	-0.757	$2.76 \times 10^{-4}$
red	-0.60207	0.013595	pink	-0.68541	$1.69 \times 10^{-3}$
darkred	-0.59137	0.015833	greenyellow	-0.67831	$1.97 \times 10^{-3}$
greenyellow	-0.56995	0.021168	red	-0.64532	$3.83 \times 10^{-3}$
yellow	0.567828	0.021763	turquoise	-0.59771	$8.80 \times 10^{-3}$
darkgrey	-0.55246	0.026474	lightyellow	-0.56347	0.0149
saddlebrown	-0.52983	0.034783	brown	0.548516	0.0184
turquoise	-0.51371	0.041809	darkgreen	-0.46366	0.0526
purple	-0.46788	0.067606	blue	-0.4604	0.0545
darkolivegreen	-0.41272	0.112122	purple	-0.44182	0.0664
sienna3	-0.39535	0.129604	darkgrey	-0.39065	0.109
darkturquoise	0.386541	0.139154	orange	-0.36966	0.131
darkorange	0.384966	0.140912	white	0.28737	0.248
darkmagenta	0.375586	0.151688	darkred	0.283247	0.255
brown	0.366095	0.163144	black	0.271383	0.276
tan	-0.36522	0.164229	salmon	-0.24203	0.333
pink	0.348979	0.18524	skyblue	0.207071	0.410
magenta	-0.32559	0.218473	cyan	0.18778	0.456
midnightblue	-0.29168	0.273014	lightgreen	0.166495	0.509
lightgreen	0.289921	0.276056	grey60	0.15156	0.548
paleturquoise	-0.28045	0.29276	midnightblue	0.136078	0.590
white	0.27727	0.29849	magenta	-0.13459	0.594
orange	0.19607	0.466754	darkturquoise	0.129954	0.607
steelblue	0.17355	0.520357	lightcyan	0.090241	0.722
skyblue	0.145869	0.589857	darkorange	-0.05166	0.839
lightyellow	-0.11665	0.667028	green	-0.04745	0.852
green	-0.09882	0.715786	royalblue	0.020456	0.936
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			



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 best fit was calculated using the *stat\_smooth* with the loess function from R package *ggplot2* (version 1.0.1). (fig. 5.1).

The expression pattern observed were intriguing where there both the “black”(fig. 5.1a) and “tan”(fig. 5.1b) 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. 5.1d).

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 volume 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 5.2) (e.g. “central nervous system development” and “glutamate metabolic process”) and the “tan” amygdala network were also related to ammonium ion metabolism (table 5.3) which is vital for glutamine synthesis from glutamate (Liaw 1995).

Together, it is highly likely that the “black” hippocampus and “tan” amygdala networks are related to brain development and function.

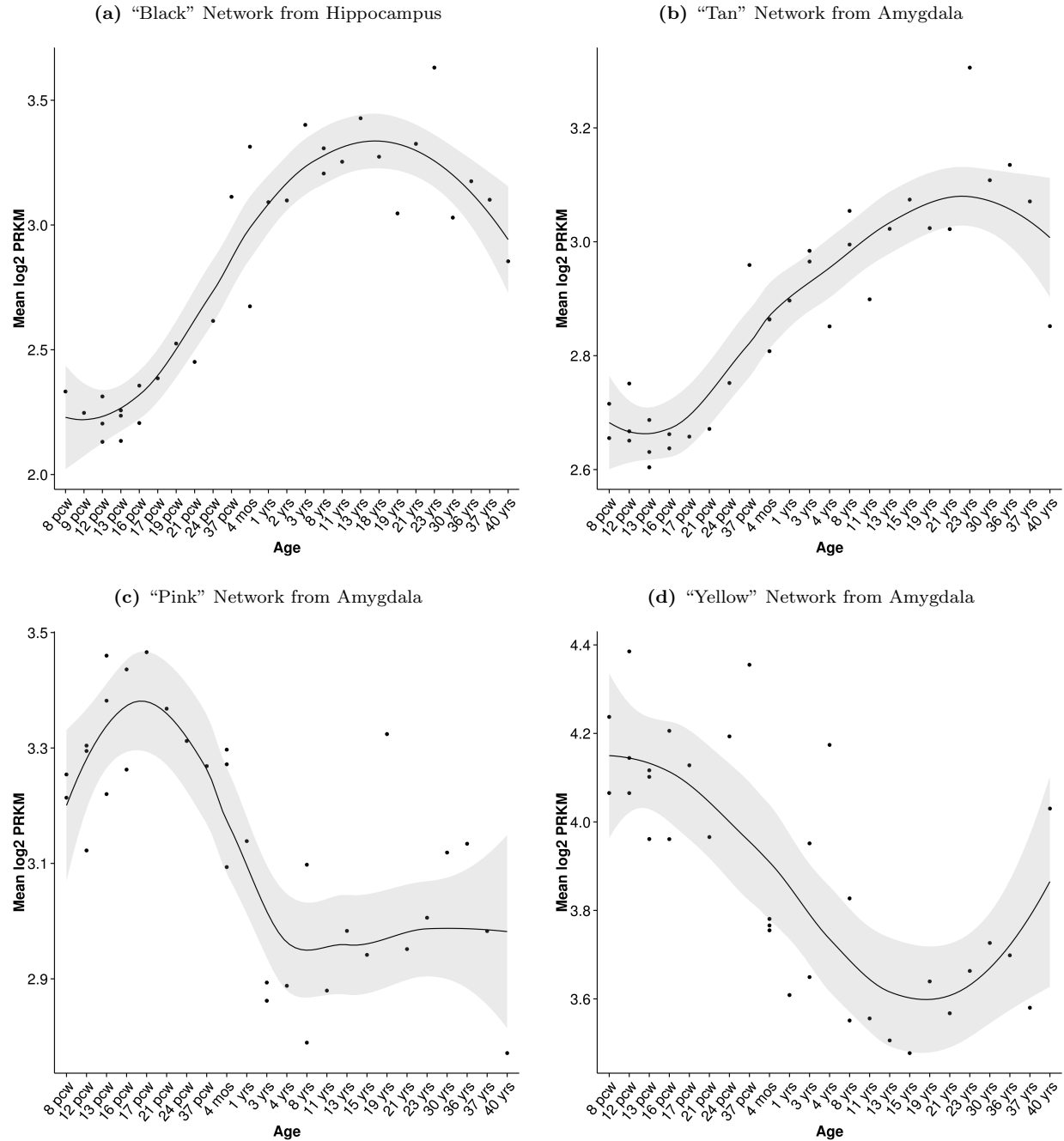
### Associate Co-expression network with PGC schizophrenia data

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 with 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  $1.38 \times 10^{-41}$  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.

**Figure 5.1:** Mean Gene Expression 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.



**Table 5.2:** 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	$4.92 \times 10^{-6}$
GO:0007417	central nervous system development	$5.94 \times 10^{-5}$
GO:0002821	positive regulation of adaptive immune response	$6.12 \times 10^{-5}$
GO:0006082	organic acid metabolic process	$1.03 \times 10^{-3}$
GO:0032291	axon ensheathment in central nervous system	$1.86 \times 10^{-3}$
GO:1901565	organonitrogen compound catabolic process	$1.99 \times 10^{-3}$
GO:0006536	glutamate metabolic process	$3.54 \times 10^{-3}$
GO:0021762	substantia nigra development	$3.73 \times 10^{-3}$
GO:0044281	small molecule metabolic process	$4.34 \times 10^{-3}$
GO:0030194	positive regulation of blood coagulation	$4.59 \times 10^{-3}$
GO:0009607	response to biotic stimulus	$6.14 \times 10^{-3}$
GO:0002702	positive regulation of production of molecular mediator of immune response	$6.21 \times 10^{-3}$
GO:0034103	regulation of tissue remodeling	$6.21 \times 10^{-3}$
GO:0014015	positive regulation of gliogenesis	$7.47 \times 10^{-3}$
GO:0098542	defense response to other organism	$7.95 \times 10^{-3}$
GO:0019835	cytolysis	$8.72 \times 10^{-3}$

**Table 5.3:** 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	$1.37 \times 10^{-3}$
GO:0044106	cellular amine metabolic process	$4.2 \times 10^{-3}$
GO:0009308	amine metabolic process	$5.41 \times 10^{-3}$
GO:0046519	sphingoid metabolic process	$6.01 \times 10^{-3}$

## Partitioning of Heritability

## 5.4 Discussion



## Chapter 6

# Heritability of Response to antipsychotic treatment

Important to schizophrenia research

### 6.1 Introduction

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

### 6.2 Methodology

### 6.3 Result

### 6.4 Discussion



## Chapter 7

# Risk Prediction

### 7.1 Methodology

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

$$\hat{Y} = \text{diag}(\beta)X \tag{7.1}$$

where  $X$  is the standardized genotype,  $\beta$  is the test-statistic calculated from other studies.

#### 7.1.1 Simulation

### 7.2 Result

### 7.3 Discussion





## Chapter 8

## Conclusion



# Bibliography

- [1] David M Altshuler et al. “Integrating common and rare genetic variation in diverse human populations.” In: *Nature* 467.7311 (2010), pp. 52–58. ISSN: 0028-0836. DOI: 10.1038/nature09298.
- [2] American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders*. American Psychiatric Publishing, 2013, p. 991. ISBN: 089042554X. DOI: 10.1176/appi.books.9780890425596.744053.
- [3] Brendan K Bulik-Sullivan et al. “LD Score regression distinguishes confounding from polygenicity in genome-wide association studies”. In: *Nature Genetics* 47.3 (2015), pp. 291–295. ISSN: 1061-4036. DOI: 10.1038/ng.3211.
- [4] E Eden et al. “GORilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists”. eng. In: *BMC Bioinformatics* 10 (2009), p. 48. DOI: 10.1186/1471-2105-10-481471-2105-10-48[pil].
- [5] Nitin Gogtay et al. “Age of onset of schizophrenia: Perspectives from structural neuroimaging studies”. In: *Schizophrenia Bulletin* 37.3 (2011), pp. 504–513. ISSN: 05867614. DOI: 10.1093/schbul/sbr030.
- [6] Gaël Guennebaud, Benoît Jacob, et al. *Eigen v3*. <http://eigen.tuxfamily.org>. 2010.
- [7] Per Christian Hansen. “The truncated SVD as a method for regularization”. In: *Bit* 27.4 (1987), pp. 534–553. ISSN: 00063835. DOI: 10.1007/BF01937276.
- [8] Traver Hart et al. “Finding the active genes in deep RNA-seq gene expression studies.” In: *BMC genomics* 14 (2013), p. 778. ISSN: 1471-2164. DOI: 10.1186/1471-2164-14-778.
- [9] Assen Jablensky. “The diagnostic concept of schizophrenia: its history, evolution, and future prospects.” In: *Dialogues in clinical neuroscience* 12.3 (Jan. 2010), pp. 271–87. ISSN: 1294-8322.
- [10] Martin Knapp, Roshni Mangalore, and Judit Simon. “The global costs of schizophrenia.” In: *Schizophrenia bulletin* 30.2 (2004), pp. 279–293. ISSN: 0586-7614.
- [11] P Langfelder and S Horvath. “WGCNA: an R package for weighted correlation network analysis”. eng. In: *BMC Bioinformatics* 9 (2008), p. 559. DOI: 1471-2105-9-559[pil]10.1186/1471-2105-9-559.
- [12] Christiaan a. de Leeuw et al. “MAGMA: Generalized Gene-Set Analysis of GWAS Data”. In: *PLOS Computational Biology* 11.4 (2015), e1004219. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1004219.
- [13] Miao-Xin Li et al. “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 (2011), pp. 747–756. DOI: 10.1007/s00439-011-1118-2.
- [14] Teri a Manolio et al. “Finding the missing heritability of complex diseases.” In: *Nature* 461.7265 (2009), pp. 747–753. ISSN: 0028-0836. DOI: 10.1038/nature08494.

- [15] Arnold Neumaier. “Solving Ill-Conditioned and Singular Linear Systems: A Tutorial on Regularization”. In: *SIAM Review* 40.3 (1998), pp. 636–666. ISSN: 0036-1445. DOI: 10.1137/S0036144597321909.
- [16] Tom F. Nugent et al. “Dynamic mapping of hippocampal development in childhood onset schizophrenia”. In: *Schizophrenia Research* 90.1-3 (2007), pp. 62–70. ISSN: 09209964. DOI: 10.1016/j.schres.2006.10.014.
- [17] H Allen Orr. “The Population Genetics of Adaptation: The Distribution of Factors Fixed during Adaptive Evolution”. In: *Evolution* 52.4 (1998), pp. 935–949. ISSN: 00143820.
- [18] Genomes Project et al. “An integrated map of genetic variation from 1,092 human genomes”. In: *Nature* 135.V (2012), pp. –9. ISSN: 00280836. DOI: 10.1038/nature11632. arXiv: /www.pubmedcentral.nih.gov/articlerender.fc [Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., & Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic acids research*, 29(1), 308–11. Retrieved from http:].
- [19] Stephan Ripke et al. “Biological insights from 108 schizophrenia-associated genetic loci”. In: *Nature* 511 (2014), pp. 421–427. ISSN: 0028-0836. DOI: 10.1038/nature13595.
- [20] Sukanta Saha, David Chant, and John McGrath. “A Systematic Review of Mortality in Schizophrenia”. In: *Archives of general psychiatry* 64.10 (2007), pp. 1123–1131. ISSN: 0003-990X. DOI: 10.1001/archpsyc.64.10.1123.
- [21] Stephen H. Schultz, Stephen W. North, and Cleveland G. Shields. “Schizophrenia: A review”. In: *American Family Physician* 75.12 (2007), pp. 1821–1829. ISSN: 0002838X.
- [22] G Shieh. “Estimation of the simple correlation coefficient”. eng. In: *Behav Res Methods* 42.4 (2010), pp. 906–917. DOI: 10.3758/BRM.42.4.90642/4/906[pil].
- [23] Eleanor H. Simpson, Christoph Kellendonk, and Eric Kandel. “A Possible Role for the Striatum in the Pathogenesis of the Cognitive Symptoms of Schizophrenia”. In: *Neuron* 65.5 (2010), pp. 585–596. ISSN: 08966273. DOI: 10.1016/j.neuron.2010.02.014.
- [24] Zhan Su, Jonathan Marchini, and Peter Donnelly. “HAPGEN2: Simulation of multiple disease SNPs”. In: *Bioinformatics* 27.16 (2011), pp. 2304–2305. ISSN: 13674803. DOI: 10.1093/bioinformatics/btr341.
- [25] F Supek et al. “REVIGO summarizes and visualizes long lists of gene ontology terms”. eng. In: *PLoS One* 6.7 (2011), e21800. DOI: 10.1371/journal.pone.0021800PONE-D-11-04111[pil].
- [26] Dennis Velakoulis et al. “Hippocampal and amygdala volumes according to psychosis stage and diagnosis”. In: *Archives of general psychiatry* 63 (2006), pp. 139–149.
- [27] Danielle Welter et al. “The NHGRI GWAS Catalog, a curated resource of SNP-trait associations”. In: *Nucleic Acids Research* 42.D1 (2014), pp. 1001–1006. ISSN: 03051048. DOI: 10.1093/nar/gkt1229.
- [28] World Health Organization. *WHO methods and data sources for global burden of disease estimates*. Tech. rep. Geneva, 2013.
- [29] J Yang et al. “GCTA: a tool for genome-wide complex trait analysis”. eng. In: *Am J Hum Genet* 88.1 (2011), pp. 76–82. DOI: 10.1016/j.ajhg.2010.11.011S0002-9297(10)00598-7[pil].
- [30] Bin Zhang and Steve Horvath. “A general framework for weighted gene co-expression network analysis”. eng. In: *Stat Appl Genet Mol Biol* 4.1 (2005), Article17. ISSN: 1544-6115. DOI: 10.2202/1544-6115.1128.

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

# Appendix