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 7, 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

GCTA Genome-wide Complex Trait Analysis. 16, 17
GD Gestation Day. 20
GWAS Genome Wide Association Study. 7, 8, 16, 17

LD Linkage Disequilibrium. 8, 11–15, 17
LDSC LD SCore. 16, 17

maf Minor Allele Frequency. 16, 17

PGS Polygenic Risk Score. 25

RIN RNA integrity number. 20
RPKM Reads Per Kilobase per Million mapped reads. 20

SCZ Schizophrenia. 19
SE Standard Error. 12
SHREK SNP Heritability and Risk Estimation Kit. 16–18
SNP Single Nucleotide Polymorphism. 7–9, 11, 15–17
SVD Singular Value Decomposition. 14

tSVD Truncated Singular Value Decomposition. 13–15

WGCNA Weighted Gene Co-expression Network Analysis. 20
WGS Whole Genome Sequencing. 16

Contents

List of Figures

Introduction

Some considerations

1. PRSice requires the phenotype to aid its selection (More information= stronger)
2. It seems like LDSC doesn't necessary perform badly in oligogenic situation. Rather, it is that when the trait is oligogenic, it is more likely for LDSC to behaviour in a strange way.
3. For each condition: extreme phenotype, quantitative trait, case control, we can have a separated review. Discuss on the benefits and challenges of each condition and the method we deal with them. So we can have two chapters (case control, quantitative trait) where extreme phenotype can be a big subsection within quantitative trait.
4. For each chapter, there will be this introduction (review on the method), our methodology (Calculation, implementation and also simulation), result (the simulation result). Then we can have the application (PGC, network)

Things that I have to include

1. Schizophrenia
   (a) Case Control (PGC)
   LDSC has basically did everything related to partitioning and genetic correlation, so we should focus on the brain network instead
   (b) Drug Response
   No one has done it before, should have a brief section here. Focus should be with the *heritability* of response, not to identify the variants. However, we can try to partition the heritability of drug response too.
   Is it really related to genetics? Or is it something related to environmental? e.g. Discouraging family members leads to not adhere to medicine
2. Heritability Estimation
   This part should be straightforward, just the algorithm and the simulation results

Chapter 1

Literature Review

1.1 Schizophrenia

1.1.1 Epidemiology

1.1.2 Psychiatric Genomics Consortium

1.2 Heritability Estimates

1.2.1 Family Studies

1.2.2 Twin and Adoption Studies

Should briefly talk about how Twin modeling was used for finding the GE contribution. Should also mention the ACE model. At the end, we can talk about the heritability estimates of SCZ and AD

1.2.3 Narrow Sense Heritability Estimates

Genome-wide Complex Trait Analysis

1.3 Risk Prediction

1.4 Summary

Chapter 2

Heritability Estimation

2.1 Introduction

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 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 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

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

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

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

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

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

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

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

And the heritability is then defined as

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

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:

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

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

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

By definition,

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

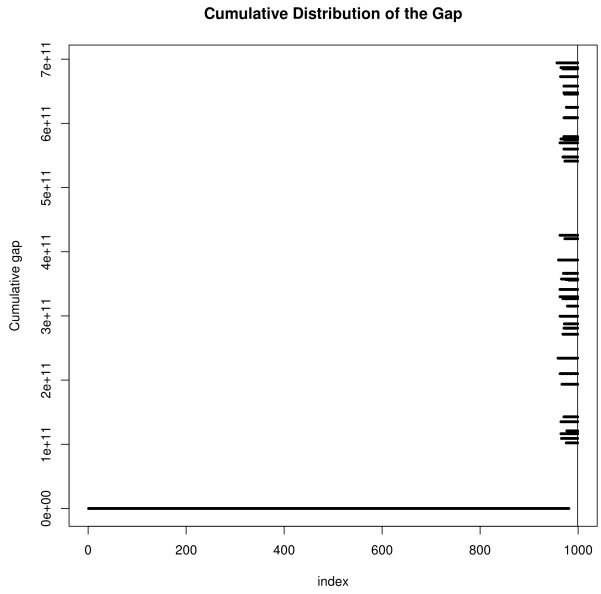
where $\boldsymbol{\Sigma}_k$ equals to replacing the smallest $n - k$ singular value replaced by 0 (Hansen, 1987). Alternatively, we can define

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

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

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



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 $\boldsymbol{A}$ 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.

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

To model a polygenic quantitative trait in GWAS and Whole Genome Sequencing (WGS) studies, we assigned per-SNP effect sizes drawn from the exponential distribution with $\lambda = 1$ to varying number of causal variants k and heritabilities H where $H \in [0, 1)$. An exponential distribution was chose based on work of Orr (1998) which suggested the exponential distribution provides a heuristic expectation about the genetic architecture of adaptation.

To simulate samples with genetic architecture comparable to true human population, *HAPGEN2*(Su, Marchini, and Donnelly, 2011) was used. n samples were simulated based on the genome structure of the Northern Europeans from Utah (CEU) with 20,000 SNPs where k of those are causal. In order to simulate a quantitative trait with target heritability of h, the per SNPs effect were calculated as

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

where $\boldsymbol{\gamma}$ is the vector of per SNP effect size and H is the simulated heritability. The only exception is when $k = 1$ where $\beta = H$ such that we can simulate SNPs with large effect size.

After the per SNP effect were simulated, we distribute the effect size to k randomly selected SNP(s) according to their Minor Allele Frequency (maf). Therefore for SNPs with a small maf, a larger effect size is given such that the effect size distribution in our simulation will follow that observed in the real life scenario(Manolio et al., 2009).

Assuming $\boldsymbol{X}$ to be the standardized genotype of k causal SNPs in n samples, one can get the phenotype of

the simulated samples using

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

For each batch of simulated samples, we calculate the estimated heritability using SHREK, GCTA, LDSC 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(Project et al., 2012) to construct the LD matrix and calculate the LD score respectively. The whole process were repeated 100 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.

In order to determine a realistic and reasonable sample size for all simulation condition, we manually curated the sample size of all the studies presented on the GWAS CatalogWelter et al., 2014 (version 2015-07-17) to get the distribution of sample size in existing studies. It was observed that the mean sample size of all published GWAS is $\sim$ 7,200 samples. Thus, we consider a simulation sample size of 7,200 should be comparable to general GWAS studies. Although a large sample size are generally required for GWAS, it might also worthwhile for one to test the performance when only small sample size is available. If the tools perform well with a small sample size, it will be beneficial to studies of rare complex traits where it is often difficult to collect a large amount of samples. As a result of that, we also perform simulation with 1,000 samples.

Another consideration is the number of causal SNPs as it determine the complexity of the trait which usually directly proportional with the number of causal SNPs. To capture the full spectrum of trait complexity (e.g. Mendelian to Oligogenic to Complex Traits), we selected number of causal SNPs to be $k \in \{1, 10, 50, 100, 1000\}$ such that effect size$\in (0, \frac{1}{k})$.To simplify the simulation procedure, all causal SNPs were included within the data.

To summarize,

1. Randomly select 200,000 SNPs from chromosome 1
2. Randomly generate k effect size following eq. (2.33) where $k \in \{1, 10, 50, 100, 1000\}$
3. Randomly assign the effect size to k SNPs where SNPs with small maf will get a large effect size.
4. Simulate n samples using *HAPGEN2* with $n \in \{1000, 7200\}$
5. Repeat step 4 100 times
6. Repeat step 1-5 50 times

2.3.2 Case Control Studies

Simulate of the case control studies are very much like that in the quantitative trait settings except that we will have to generate the phenotype according to a liability threshold.

2.3.3 Exreme Phenotype Selections

2.4 Result

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

2.5 Discussion

Chapter 3

Heritability of Schizophrenia

3.1 Introduction

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 from different developmental stages were provided by BrainSpan with an average of 2.2 samples per developmental stage.

Previous studies suggested that the Hippocampus is one of the brain region that are most affected in schizophrenia samples(Velakoulis et al., 2006; Nugent et al., 2007). Therefore, we focus on building the gene co-expression network of hippocampus development in this study RNA Sequencing data of the Hippocampus

region were obtained from BrainSpan and undergo a series of quality control before the construction of the network.

For each developmental stage, we select the sample with a dissection score ≥ 3 and an RNA integrity number (RIN) ≥ 8. 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.

A total of 33 samples passed the quality control with age ranged from Gestation Day (GD)8 to 23 years old.

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 developmental stage, we try to construct a co-expression network for genes across all developmental stage. The transformed RPKM values were standardized across developmental stages such that it is mean centered with standard deviation of 1.

Finally, there were 17,266 genes passing through the quality threshold and were used for the construction of co-expression network.

Network Construction

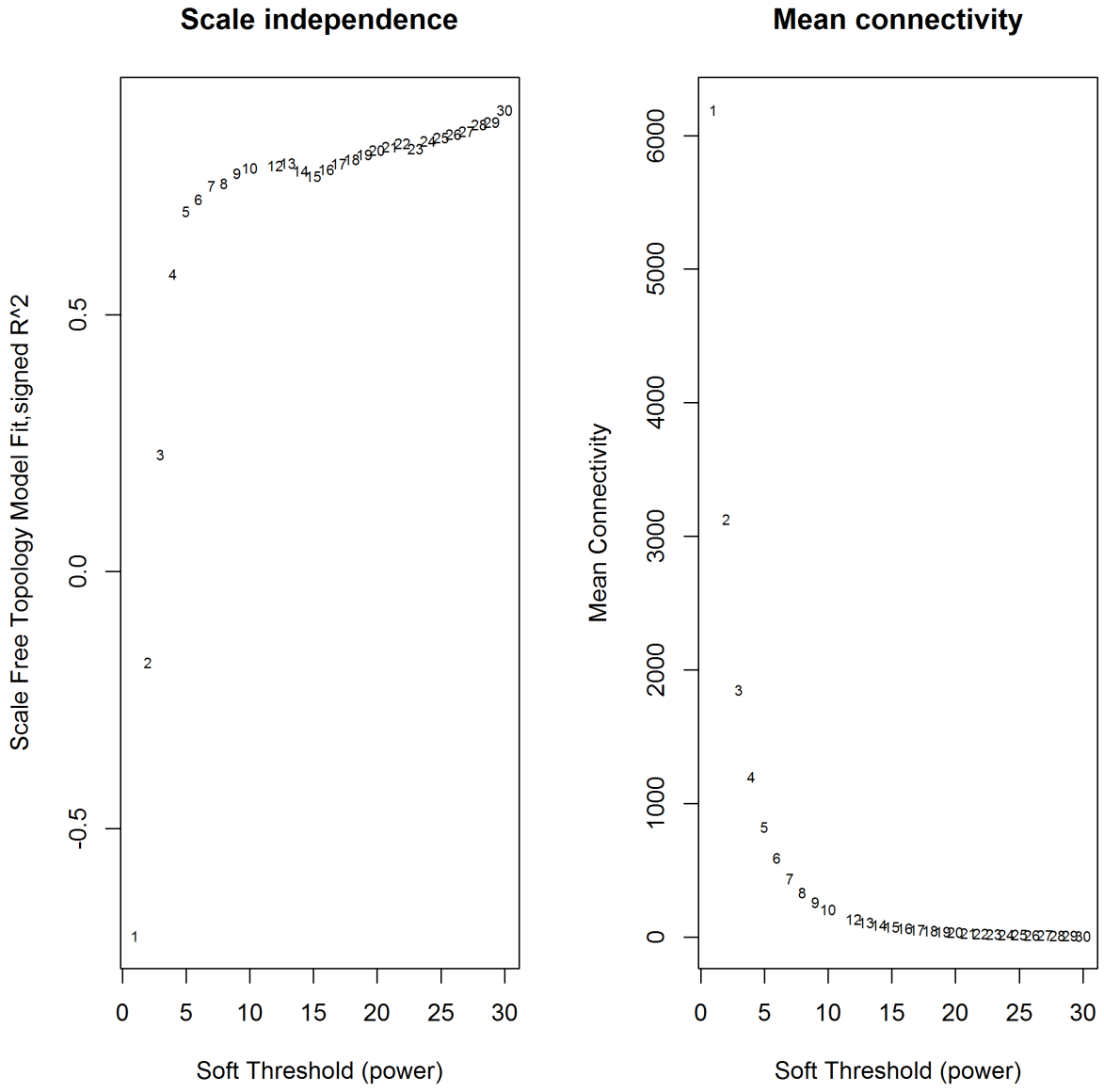
WGCNA (ver 1.47) were used for the construction of gene co-expression network(Langfelder and S Horvath, 2008). The *blockwiseModules* function, using Biweight Midcorrelation for the construction of correlation matrix and a restriction of minimum model size of 30. Soft-power threshold were set to 13 where it is the first threshold value which has $R^2 \approx 0.8$ (0.795) and the R^2 is saturated(Zhang and Steve Horvath, 2005)(fig. 3.1).

3.3.2 Result

Sample Quality Controls and Normalization of data

3.4 Discussion

Figure 3.1: Soft-power threshold selection. A soft-power of 13 were selected as it is the first threshold value having $R^2 \approx 0.8$ (0.795) and where the R^2 is saturated.



Chapter 4

Heritability of Response to antipsychotic treatment

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

[1] 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.

[2] 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:].

[3] 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.

[4] 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[pii]. URL: http://www.ncbi.nlm.nih.gov/pubmed/21139158.

[5] 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.

[6] Gaël Guennebaud, Beno\^it 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. URL: http://portal.acm.org/citation.cfm?id=891601.

[8] 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[pii]. URL: http://www.ncbi.nlm.nih.gov/pubmed/21167468.

[9] 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. URL: http://www.nature.com/doifinder/10.1038/ng.3211.

[10] 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. URL: http://www.jstor.org/stable/2411226.

[11] 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.

[12] 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. URL: http://dx.doi.org/10.1038/nature08494.

[13] 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.

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

[15] 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.

[16] 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. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3870982%5C&tool=pmcentrez%5C&rendertype=abstract.

[17] 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[pii]10.1186/1471-2105-9-559. URL: http://www.ncbi.nlm.nih.gov/pubmed/19114008.

[18] 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. URL: http://www.ncbi.nlm.nih.gov/pubmed/16646834.

Appendix