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 1, 2015

Declaration

Acknowledgements

Abbreviations

GWAS Genome Wide Association Study. 7, 8

LD Linkage Disequilibrium. 8

SCZ Schizophrenia. 11

SNP Single Nucleotide Polymorphism. 7, 8

Contents

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)

Chapter 1

Literature Review

1.1 Twin 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 Searching for Genetic Variants

1.2.1 Role of Common Variants

Genome Wide Association Study

Should talk about what is GWAS and how it is used. Should also talk about the current GWAS studies in SCZ and AD

1.2.2 Role of Rare Variants

Exome Sequencing

Similar to the GWAS. Talk about the Pros and Cons. Need to briefly mention the Denovo paper and Shaun's paper.

Whole Genome Sequencing

Very very brief description of WGS and the current status.

1.3 Narrow Sense Heritability

1.4 Risk Prediction

1.5 Summary

Chapter 2

Heritability Estimation

This chapter should be used in similar way as the general method section in Clara's thesis. Considering that the subsequent chapters all rely on this implementation.

2.1 Introduction

2.2 Methodology

The narrow-sense heritability is defined as

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

where $var(X)$ is the variance of the genotype and $var(Y)$ is the variance of the phenotype. In a Genome Wide Association Study (GWAS), regression were performed between the 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 equation 2.1, one can then have

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

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 equation 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 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 under the null distribution, t^2 should have mean approximately equal to 1,SNP contribution (f) is then defined as:

$$f = \frac{t^2-1}{n-2+t^2} \tag{2.5}$$

Under the condition where all SNPs were independent, the heritability of the phenotype can be simply defined as

$$h^2 = \sum_{1}^{m} f \tag{2.6}$$

where m is the number of SNP.

Considering that one of the main concept in GWAS is to be be able to "tag" the true causal variants using common SNPs based on the Linkage Disequilibrium (LD) between the SNP, it is impractical to assume the SNPs to be independent from each other. When LD exists between the SNPs, equation 2.6 will provide an over-estimation of the heritability. In order to obtain an unbiased estimation of the heritability of the phenotype, one must take into account of the linkage structure between the SNPs. To account for the LD, one can consider that the SNP contribution as a combination of the true SNP contributions and the effect from other SNPs through LD. This will require the knowledge of the correlation matrix between the f variables.

As f is a function of χ^2 following the F-distribution, we can obtain the variance covariance matrix of f by first calculating the variance covariance matrix of the χ^2 variables.

First, let that X_i be the standardized genotype with standard normal mean z_i and non-centrality parameter μ_i, we have

$$\begin{aligned}
\mathrm{E}[X_i] &= \mathrm{E}[z_i + \mu_i] \\
&= \mu_i \\
\mathrm{var}(X_i) &= \mathrm{E}[(z_i+\mu_i)^2] + \mathrm{E}[(z_i+\mu_i)]^2 \\
&= \mathrm{E}[z_i^2 + \mu_i^2 + 2z_i\mu_i] + \mu_i^2 \\
&= 1
\end{aligned}$$

Given the LD between two genotype X_i and X_j are ρ_{ij}, then

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

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

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

Given these information, we can then calculate $\text{Cov}(\chi_i^2, \chi_j^2)$ as:

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

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

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

By definition,

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

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

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

As a result, the variance covariance matrix of the X^2 can be calculated as

$$\text{cov}(X_i^2, X_j^2) = 2\rho_{ij}^2 + 4\rho_{ij}\mu_i\mu_j$$

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

3.3.2 Result

3.4 Discussion

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

5.1.1 Simulation

5.2 Result

5.3 Discussion

Chapter 6

Conclusion

Appendix