2017

Investigation on Genetic Modifiers of Age at Onset of Major Depressive Disorder

Huseyin Gedik

Investigation on Genetic Modifiers of Age at Onset of Major Depressive Disorder

A dissertation submitted in partial fulfillment of the requirements for the degree of Master of Science at Virginia Commonwealth University

by
Huseyin Gedik, M.Sc.

Advisor: Silviu-Alin Bacanu, Ph.D.
Associate Professor
Department of Psychiatry
Virginia Institute for Psychiatric and Behavioral Genetics

Virginia Commonwealth University
Richmond, VA
August 2017

Acknowledgements

First, I would like to thank to Dr. Silviu-Alin Bacanu for accepting me as a graduate student.

Second, I would like to thank Dr.Hermine Maes and Dr. Timothy York for their kindness and understanding. I would like to also give credit to two post-doctoral trainees at VIPBG for their contribution to this thesis study. Principal Component analysis has been completed by Dr. Tim Bigdeli. He also helped me to have additively coded genotype post-Quality Control (QC) data. Dr. Roseann Peterson provided me with the phenotype data on CSA variable.

The thesis project depends on the post-genotyping experiment analysis cleaning and filtering the genotype data completed by Dr. Johnathan Flint`s research group at Cambridge. I give credit to their work on sequence processing and ulterior steps until variant calling.

I would like to acknowledge that I had financial support from the Turkish Government with a scholarship supporting all my educational expenses during my graduate study.

My mother and father always help me to achieve through all my academic studies so I would like to thank my parents for their understanding and support.

Lastly, but not least, I would like to thank to my wife Remziye for her support during graduate school and bearing the challenges with me all this time.

Table of Contents

List of Tables

List of Figures

.

List of Abbreviations

5HTR2A……...Serotonin 2A Receptor gene

AAO…….……...Age at Onset

CC analysis…….....Case Cohort analysis

COMT…….....Catechol-o-Methyltransferase gene

Cox PH…….....Cox Proportional Hazard

CSA…….Childhood Sexual Abuse

dbSNP…….....Single Nucleotide Polymorphism database

GATK……..Genome Analysis Tool Kit

GO……..….....Gene Ontology

g:GOSt…….....GO statistics

GWAS…….....Genome Wide Association Study

GWSA…….....Genome Wide Survival analysis

HapMap…….....Haplotype Map

KEGG……..Kyoto Encyclopedia of Genes and Genomes

LD……....Linkage Disequilibrium

LHPP……...Phospholysine Phosphohistidine Inorganic Pyrophosphate Phosphatase

MDD…….....Major Depressive Disorder

MAF……..……..Minor Allele Frequency

OMIM……..Online Mendellian Inheritance in Man

SIRT1……. .Sirtuin1

SLC6A4…….....Solute carrier family 6 member 4 (5HTT, serotonin transporter)

SNP…….....Single Nucleotide Polymorphism

SNV…….....Single Nucleotide Variation

TH…….....Tyrosine Hydroxylase gene

TPH 1…….....Tryptophan Hydroxylase 1 gene

Abstract

INVESTIGATION ON GENETIC MODIFIERS OF AGE AT ONSET OF MAJOR DEPRESSIVE DISORDER

By Huseyin Gedik, M.Sc.

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science at Virginia Commonwealth University.

Virginia Commonwealth University, 2017

Major Advisor: Silviu-Alin Bacanu, Ph.D.
Associate Professor
Department of Psychiatry
Virginia Institute for Psychiatric and Behavioral Genetics

Major Depressive Disorder (MDD) is a complex multifactorial disorder, which would lead to disability. Environmental and genetic factors are involved in MDD etiology. The aim of this project was to identify loci modifying age at onset (AAO) of MDD using survival models after adjusting for Childhood Sexual Abuse (CSA). To achieve this aim, a dataset was made available by the China Oxford and VCU Experimental Research on Genetic Epidemiology (CONVERGE) consortium. The study population had 5,220 controls and 5,282 cases with MDD. We performed two univariate association analyses using Cox Proportional Hazard (Cox PH) models. These two are Full Sample (FS), cases and controls, and only the Case Cohort (CC). No genome-wide significant associations were found in univariate analyses. Subsequent gene set enrichment analysis showed that there were significant enrichments in neurological Gene Ontology terms and some novel non-neural pathways. These findings may allow us to better understand MDD pathology.

1. Introduction

Major depressive disorder (MDD) (OMIM #608516) has lifetime prevalence in the range of 1-16.7% (Kessler & Bromet, 2013). It is also much more common in females than males and mean of age at onset (AAO) ranges from approximately 24 to 34 (Weissman et al., 1996). Psychiatric disorders, including MDD, have underlying genetic factors that modulate the risk for these disorders. A meta-analysis of twin studies showed that the heritability of major depression is 37% (Sullivan, Neale, & Kendler, 2000), which suggests that i) genetic factors are involved in MDD etiology and ii) there are also non-genetic risk factors for depression. In a twin study on Swedish twin sample, early AAO of MDD is significantly associated with increased liability to MD (Kendler, Gatz, Gardner, & Pedersen, 2005).

Early medical intervention or lifestyle changes may ameliorate the prognosis of major depression. For this reason, prognostic markers are crucial for preventing exacerbations in patients developing depression. A single highly penetrant candidate risk locus has not been reported yet for MDD so this supports the common disease common variant hypothesis. On the other hand, Genome-wide association studies (GWAS) aim to identify numerous common variants that contribute to genetic liability to disease. Finding genome wide significant genetic loci would indicate possible biological mechanisms, which could be a potential target for drug design or prognostic marker for the disorder.

The aim of this study was to uncover genetic modifiers that affect the MDD AAO in the entire sample or only within cases. Therefore, we conducted a genome wide survival analysis which tests for an association between CONVERGE genotypes and survival to onset of MDD

among 10,502 Han Chinese female participants. The association tests were performed using Cox Proportional Hazard (Cox PH) survival regressions.

2. Background Information

2.1. Depression Definition

The mood disorders, which include depression, have been known since ancient times. The classification of depression as a mental illness goes back to the description of melancholy by Hippocrates (460-377 BC). Major depression, a common chronic recurrent disorder, affects all social segments of the population. Major Depressive Disorder (MDD) (OMIM # 608516) is a common psychiatric disorder that is characterized by persistent depressed mood, reduced interests; lessened cognitive function, and/or vegetative symptoms like abnormal sleep or appetite, change in activity, lack of self-worth, suicidal thought and fatigue or loss of energy (Otte et al., 2016). To be diagnosed as MDD by DSM5 (American Psychiatric Association, 2013) out of these symptoms at least five or more should manifest in two-week period almost daily. Also, one of the symptoms has to be either depressed mood or reduced interest.

MDD is also an important public health problem due to the loss of ability to do work. The quality of life deteriorates in many ways as depression symptoms become more prominent. In particular, the worsening in social functioning can be more pronounced than in many other physical illnesses (Hirschfeld et al., 2000). The prevalence of MDD, the negative effects on both the individual and the community level, and the burden of this disorder, point to depression being an important public health problem and financial burden to the health care system.

2.2. Epidemiology

Depression is one of the common psychiatric disorders. Studies on the epidemiology of depression have resulted in different rates of prevalence in different populations. Lifetime

prevalence of MDD differs from one country to another ranging from 1% to 16.9% (Kessler & Bromet, 2013). In China, the 12 month prevalence of MDD is estimated to be 2.3% and life time prevalence is 3.3% (Gu et al., 2013). In another study of 6,694 people, whose age range 18-96 years old, the monthly prevalence of depression was 5.2% in states of California and New York in U.S. It has been reported that depression is seen at a higher rate in women and increases towards the middle of life (Ohayon, 2007).

2.3. Etiology

The etiology of MDD have been investigated for decades. Neurotransmitter systems play a crucial role in the etiology. With the start of use of monoamine oxidase inhibitor and tricyclic antidepressant (TSA) as a medical intervention to treat psychiatric disorders, "Monoamine Hypothesis" (Schildkraut, 1965) was introduced. Then, numerous studies were conducted on mood disorders, etiopathogenesis of depression, and neurotransmitters.

The monoamine hypothesis suggests that the reduction in the function and deficiency of one of the three biogenic amines (serotonin, noradrenalin, and dopamine) or the increase in the number and sensitivity of its receptors is the underlying biological mechanism of depression. This hypothesis at first explained the cause of the depression in relation with the full or partial failure of noradrenalin receptor system, especially in functionally important noradrenalin receptor sites.

The role of serotonin in depression has been extensively studied area of research. The "Serotonin/Indolamine Hypothesis" proposed that depressive disorders are a consequence of decreased serotonin levels in the brain (Racagni & Brunello, 1999; Stahl, 1998). There are also studies on symptoms of depression associated with reduced serotonin synthesis in the brain due to the lack of tryptophan (Neumeister et al., 2004).

The underlying physiological decrease in dopamine (DA) signaling may result either from a reduction in release from the anterior synaptic neurons, or from impaired signal transduction due to changes in the number of receptors, function or altered signal processing among cells (Dunlop & Nemeroff, 2007). A study measured the level of major catecholamine metabolite, produced by the monoamine oxidase and catechol-O-methyl transferase on dopamine, (HVA) in internal jugular vein. Findings pointed out that decrease in HVA concentration correlated with the increase in depression severity in the patients' with treatment-refractory depression (Lambert, Johansson, Agren, & Friberg, 2000).

As a result, there are supporting evidences of Monoamine Hypothesis such that serotonin and dopamine play a role in biology of depression.

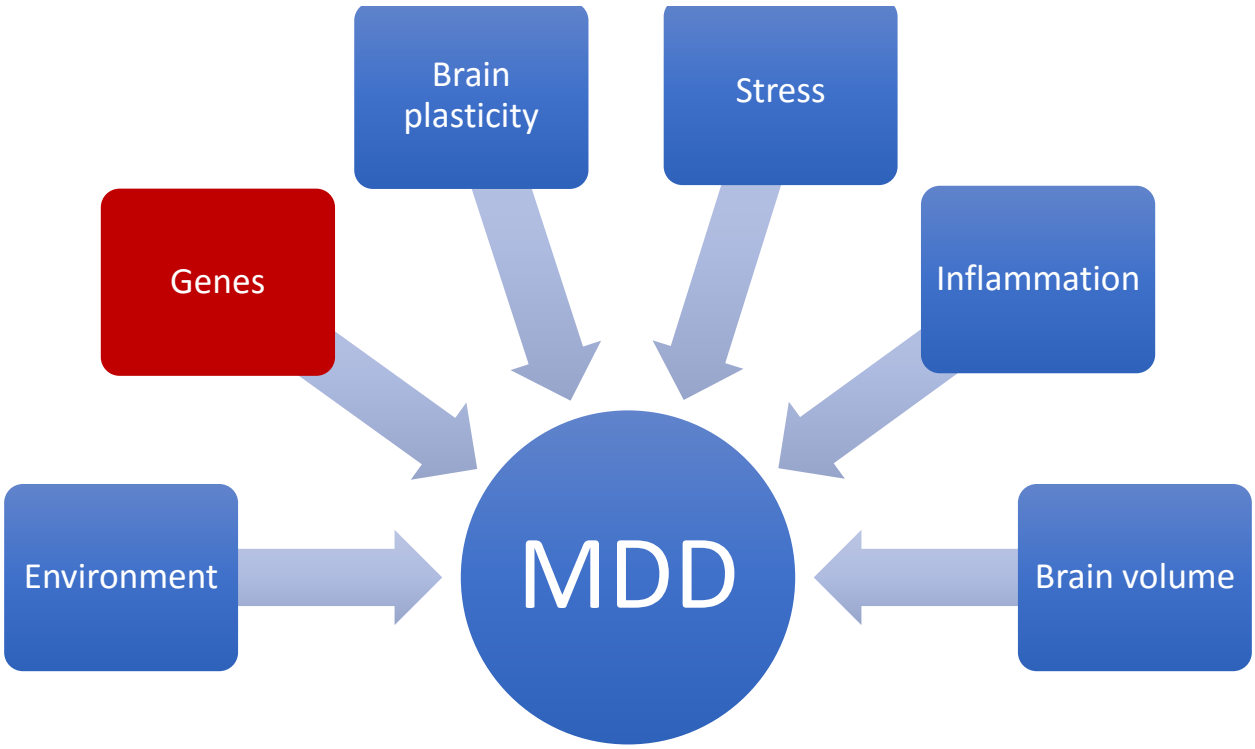


Figure.1 Etiology of MDD.

There is not any single mechanism that explaining the etiology of MDD (**Fig.1**). Stress plays a major role in etiology of MDD, specifically if encountered at earlier stages of life. A stressful event could trigger the Hypothalamic-Pituitary-Adrenal axis releasing corticosteroids (de Kloet, Joels, & Holsboer, 2005). Inflammation affects the HPA axis by increasing its activity (Besedovsky, del Rey, Sorkin, & Dinarello, 1986) The increase in activity leads to change in central nervous system hippocampal structure and function (Brown, Rush, & McEwen, 1999). Decrease in hippocampal neurogenesis could weaken the healthy response to stress (Santarelli et al., 2003). Lowered volume in hippocampus region also has a role in MDD etiology (Otte et al., 2016).

MDD is a complex disorder having both genetic and environmental components in its etiology. Two of the major environmental contributions come from the childhood adversity and stressful life events. As for the genetic component, twin studies indicated there is a heritability estimate of 37% for MDD (Sullivan et al., 2000). This suggests genetic factors are involved in etiology of MDD. Neurogenesis, HPA axis and inflammation affecting the HPA axis may also play a role due to their regulation of stress response.

2.4. Risk Factors

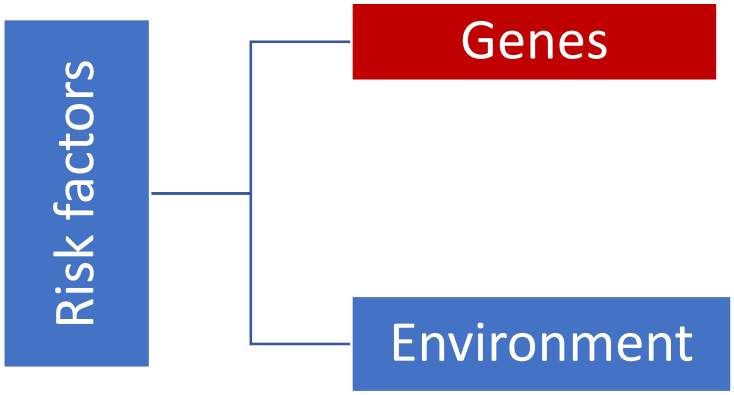


Figure.2 Risk Factors for MDD

There are various risk factors involve in MDD etiology. Among these, two broad categories are the genetic and environmental risk factors (**Fig.2**).

2.4.1. Environmental risk factors in MDD

Various studies have shown that being a woman, having a low level of education, genetic factors, the presence of a depressive personality, stressful life events, lack of close relationship, physical illness leading to loss of power and mental disorders are major risk factors for major depression (Swindle, Cronkite, & Moos, 1998). Being between the ages of 18-44, single, not working and having a low socioeconomic status are other risk factors for depression (Anthony & Petronis, 1991; Bruce, Takeuchi, & Leaf, 1991).

For females, the rate of depression is found to be higher in separated and divorced than in married ones (Weissman et al., 1996). The effect of these risk factors differs depending on the severity of depression. For instance, biological susceptibility plays a more important role in severe cases of depression, while the role of environmental factors is more substantial in non-familial types of depression (Farmer, 1996).

One of the important environmental risk factors for MDD is CSA. Research shows that CSA is an indication of earlier AAO of depression (Gladstone et al., 2004). A report on MDD in Chinese women indicated that CSA is associated with the recurrent MDD (Chen et al., 2014).

2.4.2. Genetic risk factors in MDD

At first, psychiatric genetics tried to uncover the genetic factors affecting the liability of individuals to disease via family, twin and adoption studies. A meta-analysis of twin studies (Sullivan et al., 2000) estimated the heritability of MDD at 37%, which suggests that genetic factors are involved in etiology of MDD. It is more likely that heritable forms of depression are diagnosed especially early in life, and tend to be recurrent. It is thought that the heritability of

depression is not derived from a single locus, but from the joint effects of multiple genetic loci. Such hypothesis is supported by children of depressed patients being three to four times more likely to develop MDD (Sullivan et al., 2000; Weissman et al., 2006). Familial risk for MDD includes genetic factors, family environment, and specific risk factors of an individual (Avenevoli & Merikangas, 2006).

2.4.2.1. Candidate gene approaches

The serotonin transporter gene, *SLC6A4*, (Goldman, Glei, Lin, & Weinstein, 2010) is one of the most studied genes in major depressive disorder (Levinson, 2006). The reason for the focus of research on this gene is that there are 2 different (long / Long and short / Short) genetic polymorphisms. The short allele decreases the transporter synthesis of serotonin. This decrease may slow down the adaptation of the serotonin neurons to the stimulus (Lesch et al., 1996). An association study reported that there is a significant association between antidepressant treatment response and A allele of rs7997012 in serotonin 2A receptor (*5HTR2A*) gene (McMahon et al., 2006). Catechol-o-methyl transferase (*COMT*) (Dopamine catabolism) gene have been studied to understand the genetics of MDD. In a multicenter European cohort, results pointed out that there is an association between the COMT Val/Val genotype and MDD (Massat et al., 2005).

An association study on 300 depressed patients and 265 healthy controls showed that rs1386494 in TPH gene tryptophan hydroxylase 2 (*TPH 2*) (serotonin synthase) gene has statistically significant (P_{adj}=0.012) association with MDD after adjusting for multiple testing correction (Zill et al., 2004).

2.4.2.2. Genome-wide association studies

There were two main developments in human genome research. The Human Genome Project and the International HapMap project, which allowed us to build a reference for the human

genome sequence. They also aimed to discover the human genome by identifying the genetic locations of about 25,000 genes. Two human genome drafts were published in 2001 (Lander et al., 2001; Venter et al., 2001) and the completed human genome was published later (International Human Genome Sequencing, 2004). The International HapMap project aimed to catalog common human sequence similarities by publishing a genome-wide database of genomic sequences of individuals from different populations (International HapMap, 2005).

One of the main goals of the HapMap project is to facilitate the identification of genetic variants that are associated with diseases under the common disease common variant hypothesis (Collins, Guyer, & Charkravarti, 1997; Gershon, Alliey-Rodriguez, & Liu, 2011; Lander, 1996; Pritchard & Cox, 2002). According to this hypothesis, most genetic variants leading to complex diseases should have minor allele frequencies greater than 5%. These genetic variants are risk factors or susceptibility variants to the disease. As a result of reduce in the cost of technologies, DNA microarrays allow us to assay hundreds of thousands of genetic variants at the same time. After the HapMap project, efforts focused on identifying 'tag' SNPs as representative of haplotype blocks. These efforts opened the door for effective and cost-efficient chip-based Genome Wide Association Studies (GWAS). Eventually, this method resulted in finding common variants associated with complex disorders. After the first GWAS was successfully performed (R. J. Klein et al., 2005), a growing number of GWAS studies have been conducted.

The first large scale GWAS on 18,759 independent and unrelated European descent individuals (9,240 MDD cases and 9,519 healthy controls) did not uncover any replicable genome wide significant loci associated with MDD risk (Major Depressive Disorder Working Group of the Psychiatric et al., 2013). However, they found 15 genome wide significant SNPs in the cross-disorder analyses (MDD-Bipolar disorder). Subsequently, CONVERGE consortium reported

that two loci (*LHPP* and *SIRT1*) associated with MDD risk reached genome wide significance in a large Han Chinese cohort including only women (consortium, 2015). Another large GWAS in a European cohort found 15 independent loci associated with the risk of MDD (Hyde et al., 2016). However, they could not reproduce the CONVERGE results.

There are still gaps in the genetic architecture of MDD. For instance, gene-environment interaction analysis is still in its infancy. Similarly, the analysis of rare and structural variations would be the focus for the next MDD studies.

2.4.2.3 Genome-wide survival analysis

Genome wide studies allow us to investigate the whole genome. This approach showed bore some success when applied to psychiatric disorders, e.g. CONVERGE MDD study (consortium, 2015). However, another interesting research focus of MDD might be to uncover genetic variants that increase or decrease AAO, i.e. AAO modifiers.

The study of AAO modifiers for a complex genetic disorder is achieved by conducting a Genome-wide Survival analysis (GWSA). However, their application to complex disorders such as psychiatric disorders is rather scant. In one of them, a study on alcohol dependence uncovered three loci significantly associated with the AAO of Alcohol dependence (Kapoor et al., 2014).

2.5 Survival analysis

Survival analysis is a statistical method of research that deals with the time it takes for a subject to reach the event of interest. Today, survival analysis is an important research method for various scientific fields to investigate the time until the breakdown of equipment, the occurrence of a disease, or the time it takes until an earthquake occurs (Cox, 1972). In survival analysis, the length of interval to failure is treated as a dependent variable.

2.5.1. Censoring

The concept of censoring is what sets survival analysis apart from other statistical methods. The censored observation is an unfinished observation and provides some information about the timing of the failure to occur. This means that although a unit or an individual has been under observation for a certain period, it has not failed during this interval. In this case, the time of the event is beyond the observed censoring time, and it may or may not occur later.

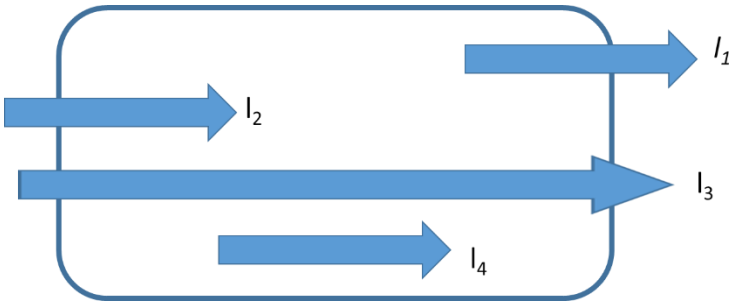


Figure.3 Possible censoring types.

For instance, some of the patients observed may still be living at the end of the study. Also, an individual under observation cannot be studied due to dropping off from the study. If the time of failure is outside the observation period due to such reasons, it is denoted as a censored observation (Lee & Wang, 2003, p. 2).

The concept of censoring is shown graphically in **Fig.3**. Here, I index (I = 1, 2, ...) refers to the time when events occur (Lee & Wang, 2003, pp. 3-4).

I_1 = the event has an end time outside the observation period (i.e. right-censored observation).

I_2 = event has an actual start time outside of the observation period (i.e. left-censored observation).

I_3 = event has started and ended at a time outside of the observation period.

I_4 = the event has known start and finish time.

In this study, there were only **right-censored** individual observations in the phenotype dataset.

2.5.2. Cox Proportional Hazards Model (Cox PH)

One of the frequently used models in survival analysis is the Cox proportional hazard (PH) model (Cox, 1972). Although the model assumes proportional hazards, there is no definite form of probability distribution for survival times. For this reason, the Cox PH model is described as a semi-parametric model. $X_1, X_2, X_3, \dots, X_n$ are explanatory variables and $x_1, x_2, x_3, \dots, x_n$ are the values of these variables. In the Cox PH model, the set of values of the explanatory variables is denoted by **x** vector, $(x_1, x_2, x_3, \dots, x_n)$. The baseline hazard function is $h_0(t)$. Cox proportional hazards model for the i^{th} individual (J. P. Klein & Moeschberger, 2003, pp. 244-245),

$$h_i(t) = h_0(t)\exp(\beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_n x_{ni}) \quad (1)$$

2.5.3. Proportional Hazards Hypothesis

The proportional hazard assumption implies that the hazard ratio is constant over time. In the survival analysis, the hazard ratio is defined as the effect of the explanatory variable on the risk of the event involved. The hazard ratio, which is the vector of explanatory variables of two groups $x = (x_1, x_2, \dots, x_n)$ and $\hat{x} = (\hat{x}_1, \hat{x}_2, \dots, \hat{x}_n)$(J. P. Klein & Moeschberger, 2003, pp. 244-245).

$$\hat{\theta} = \frac{\hat{h}_0(t)e^{\left(\sum_{j=1}^{n}\hat{\beta}_j x_j^*\right)}}{\hat{h}_0(t)e^{\left(\sum_{j=1}^{n}\hat{\beta}_j x_j\right)}} = \frac{e^{\sum_{j=1}^{n}\hat{\beta}_j x_j^*}}{e^{\sum_{j=1}^{n}\hat{\beta}_j x_j}} = e^{\left(\sum_{j=1}^{n}\hat{\beta}_j\left(x_j^* - x_j\right)\right)} \quad (2)$$

The β parameter is the natural logarithm of the hazard ratio. Equation 2 does not include the base hazard ratio $\hat{h}_0(t)$ (J. P. Klein & Moeschberger, 2003, pp. 244-245). In other words, when values for x* and x are specified, the value of the exponential term is fixed for the hazard ratio estimation, meaning that it is not time dependent. If this is denoted by constant θ, the hazard ratio can be written as given in Equation 3:

$$\hat{\theta} = \frac{\hat{h}(t,\ x^*)}{\hat{h}(t,x)} \quad (3)$$

Equation 3 is a mathematical expression of the proportional hazard assumption. Another mathematical expression of proportional hazard assumption; $\hat{\theta}\ \hat{h}$ (t, x) $= \hat{h}$ (t, x*). Here, $\hat{\theta}$ is called the proportionality constant and it is time independent (J. P. Klein & Moeschberger, 2003, p. 245). To check the proportionality assumption researchers commonly plot the survival curves using the Kaplan Meier estimates for each group (Persson, 2002, pp. 26-27). The use of the classical Cox proportional hazards model is not appropriate if the fundamental assumption of proportional hazards is grossly violated.

2. Methods

There were some major steps in an overall workflow of a GWAS as simplified in a schematic below (**Fig.4**). The scope of this study is also pointed out as thesis project on this workflow. Genome wide association genotype data was provided by the CONVERGE consortium (consortium, 2015). We are using a QC-ed subset of this dataset including genotype data and phenotype data of 10,502 individuals, e.g. age at examination, age at onset of MDD and with CSA.

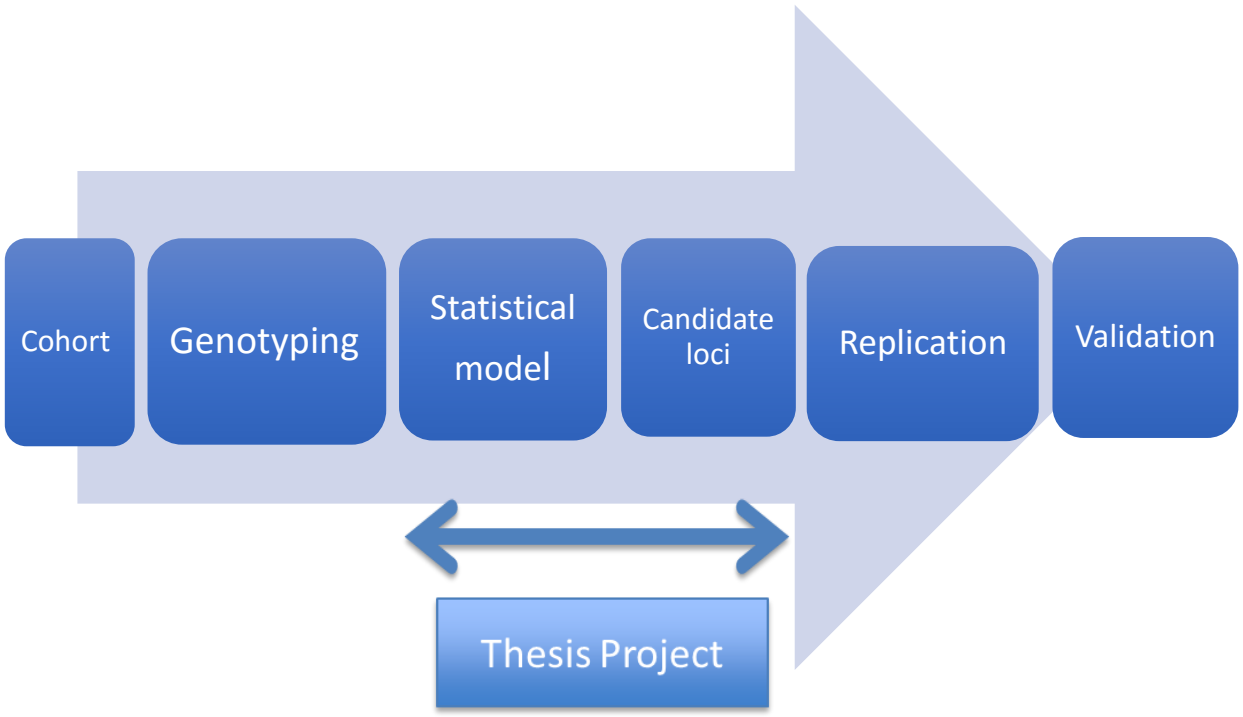


Figure.4 Work Flow of GWAS and its processing.

3.1. Study population

All participants are female and of Han Chinese descent, who recruited from 58 different mental health clinics. Controls were recruited from patients coming to the same clinics for lesser surgical operations. The case group consisted of 5,282 female participants diagnosed with MDD and the control group consisted of 5,220 females with no history of MDD. The mean age at interview for the control group was 47.66 with a standard deviation (std) of 5.61 and the age at

onset is 35.75 with std of 9.36 for the case group (**Fig.5**). There were 412 individuals in the case ohort with AAO < 18 years. These cases were excluded from the final statistical analysis.

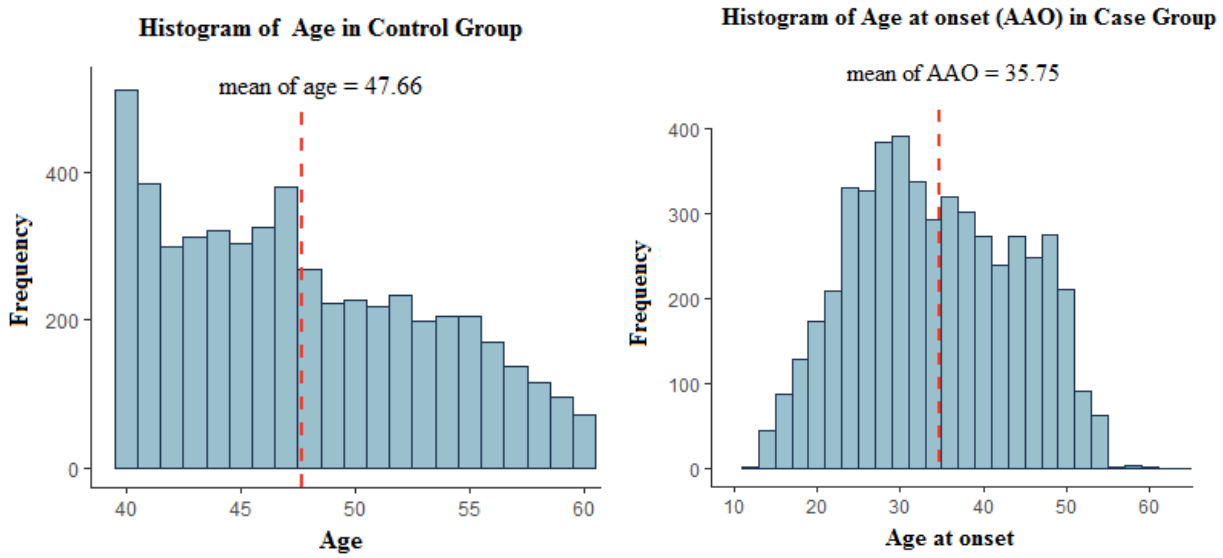


Figure.5 Histogram of age at interview distribution within the control group (left, n=5,220) and age at onset (AAO) distribution within the case group (right, n=5,282)

The clinical diagnoses of MDD were according to the Diagnostic and Statistical Manual of Mental Disorders (DSM-IV) criteria. The case group had several exclusionary criteria, such as having bipolar disorder, psychosis or mental retardation. All required permissions were granted from participating hospitals' ethics review committees. Details of the participation, the interview with individuals included into the case group and recruitment criteria were explained previously in a CONVERGE report (consortium, 2015).

3.2. Genotyping

In this study, genotyping experiment used genomic and mitochondrial DNA extracted from the saliva samples of each participant. The genetic data was low pass whole-genome sequencing data (low coverage, the average at around 1.7x) that was carried out on Illumina Hiseq. This allowed us to assay all variants either inside the protein coding or noncoding genomic regions. Alignment

of sequence reads to a reference human genome (GRCh37.p5) was completed with Stampy (v1.0.17) (Lunter & Goodson, 2011).

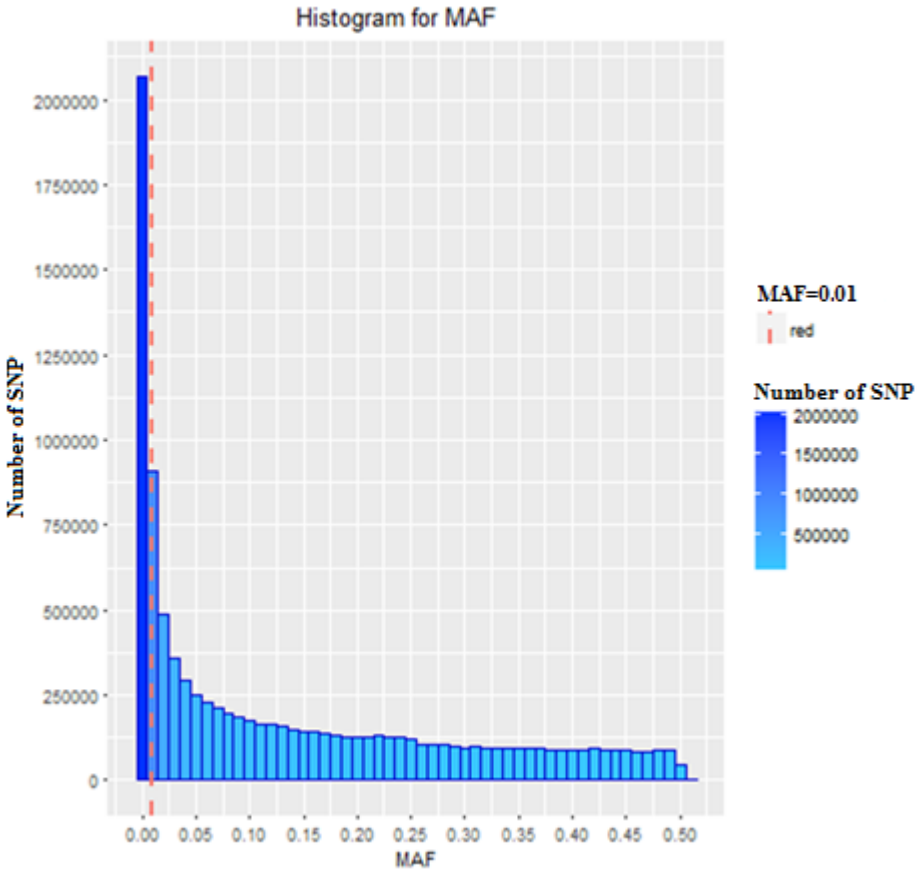


Figure.6 Distribution of Minor Allele Frequencies (MAF) of SNPs.

After filtering out SNPs with MAF < 0.01 (**Fig. 6**), out of 9,708,891 SNPs, 6,755,406 SNPs were selected for the Cox PH regression analysis.

The genotype data (single nucleotide variant, SNVs) had already been imputed and processed through stringent quality control (QC) processes. All the steps explained in this section were carried out by the CONVERGE consortium. The details of these procedures is described in the previously published research article (Cai et al., 2017).

Variant discovery and genotyping has been processed with GATK's Unified Genotyper (v2.7-2). 1000 genomes Phase 1 East Asian sample was used as the reference panel for variant calls. SNVs

were named based on the SNV database version 137 (dbSNP v137) on NCBI. For genotype imputation, BEAGLE was used (Cai et al., 2017). Genotype data included in this study has 99% called SNVs (consortium, 2015). For all called SNVs, p-value for violation of Hardy-Weinberg equilibrium (HWE) is $P > 10^{-7}$, minor allele frequencies (MAF) > 10^{-3} and information score > 0.3. For the statistical analysis, only 6,748,514 SNPs were included in the final two statistical models (FS analysis and CC analysis) described below.

3.3. Statistical model

We used a Cox proportional hazard (PH) regression model to investigate possible significant associations between AAO of MDD and genotype of SNVs (Cox, 1972). In this study, the Cox PH analyses were run on the R statistical program using the "survival" library (Therneau et al., 2000). For depression, the survival object of the Cox regression function defines the time to event, which in this case is MDD AAO. The control group was right censored, i.e. the disease phenotype has not been observed in those healthy individuals at the time of medical examination.

The ancestral PCA plot was computed using all 10,502 subjects. Principal Component 1 (PC1) and Principal Component 2 (PC2) explained the geographical distribution of the subjects (**Fig.7**). Thus, these two PCs were included in the survival regression model to account for population stratification.

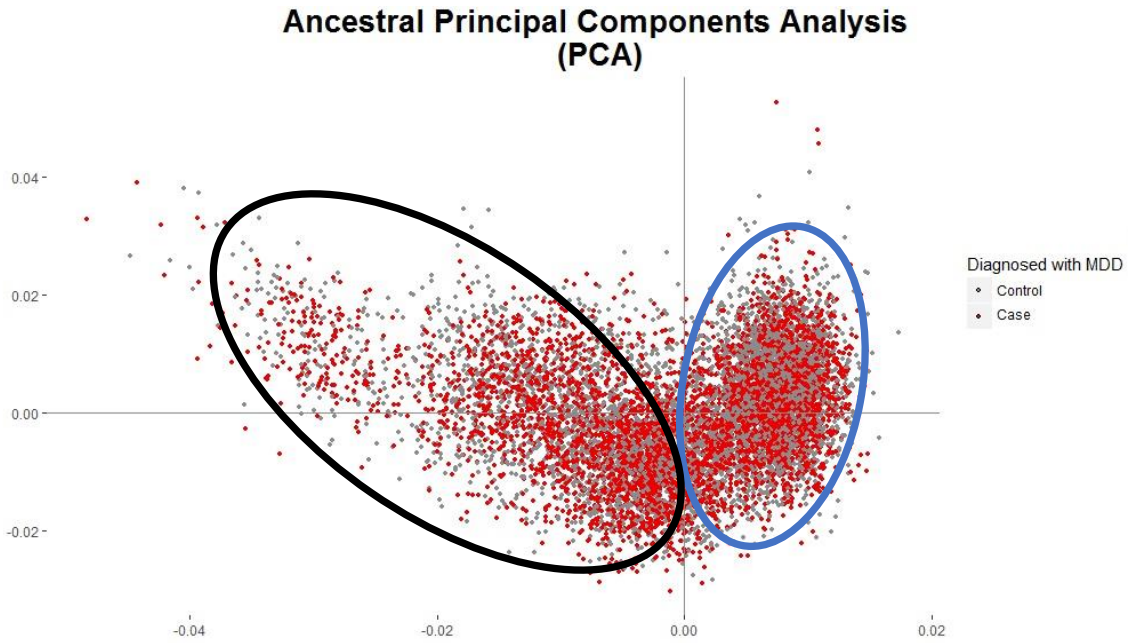


Figure.7 Ancestral Principal Component Analysis (PCA). Red and grey color filled circles represent cases and controls, respectively.

We performed two different analyses. The first one was in **Full Sample (FS)** and the second was in **Case Cohort (CC)** only. The statistical model for the analyses contains genotype as predictor and the first two principal components and CSA as covariates. I.e.

$$h_i(t, X) = h_0(t) exp(\beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4)$$

$$h(t|x) = h_0 e^{\beta_1 X_1 + X_2 \beta_2 + \beta_3 X_3 + \beta_4 X_4}$$

X_1—› Genotype vector X_2—› 1st Ancestral Principal Component

X_3—› 2nd Ancestral Principal Component X_4—› CSA vector

h_0 → Baseline hazard (no assumption).

We chose Bonferroni as the multiple testing correction method for p-values from the genetic association analyses. An example of Cox PH regression model for FS analysis in R is shown in **Fig.8**. Highlighted with yellow rectangle is the R function in the survival library. Marked in i) red is the hazard ratios and ii) blue is the significant p-value. Also highlighted in **Fig.8,** "Surv"

(survival) function in R survival library was the object specific to survival analysis. This function generates the Survival object in R function. Survival object includes individual observations of the time to event either being right censored or not. In this case, observations (AAO) in the case group are not censored whereas observations (age at interview) in control group is right-censored. The R script used for the of the Cox PH regression analysis can be found in the Appendices.

```
> G5  <- coxph(Surv(x6570$AAO, x6570$MDD) ~  x6570[,c("rs80309727_2")] + PC1 + PC2 + CSA, data = x6570)
> summary(G5)
Call:
coxph(formula = Surv(x6570$AAO, x6570$MDD) ~ x6570[, c("rs80309727_2")] +
    PC1 + PC2 + CSA, data = x6570)

  n= 9896, number of events= 4736
   (194 observations deleted due to missingness)

                             coef  exp(coef)   se(coef)      z Pr(>|z|)
x6570[, c("rs80309727_2")]  9.853e-02  1.104e+00  2.076e-02  4.746 2.08e-06 ***
PC1                        -6.053e+00  2.351e-03  1.498e+00 -4.040 5.34e-05 ***
PC2                        -9.432e+00  8.013e-05  1.489e+00 -6.336 2.36e-10 ***
CSA                         1.136e+00  3.115e+00  4.907e-02 23.151  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                           exp(coef) exp(-coef) lower .95 upper .95
x6570[, c("rs80309727_2")] 1.104e+00  9.062e-01 1.060e+00  1.149376
PC1                        2.351e-03  4.253e+02 1.248e-04  0.044305
PC2                        8.013e-05  1.248e+04 4.332e-06  0.001482
CSA                        3.115e+00  3.211e-01 2.829e+00  3.429148

Concordance= 0.571  (se = 0.005 )
Rsquare= 0.048   (max possible= 1 )
Likelihood ratio test= 489.6  on 4 df,   p=0
Wald test            = 626.2  on 4 df,   p=0
Score (logrank) test = 688.7  on 4 df,   p=0
```

Figure.8 Cox regression model in R (“survival” library).

To empirically test the significant p-values, we conducted a permutation tests on four top SNPs (**Table.5** in the Appendices), which all have MAF < 1 %. These SNPs were tested with at least $5x10^5$ permutations.

3.4. Gene set enrichment analysis

Most GWAS analyses end up with a large number of loci having numerous moderate but none/few genome wide significant association signals. For this reason, a possible approach to analyze these findings is to conduct an aggregate analysis. Gene set enrichment analysis is one of

the most well-known such methods. It allows us to aggregate information over numerous loci in biological pathways.

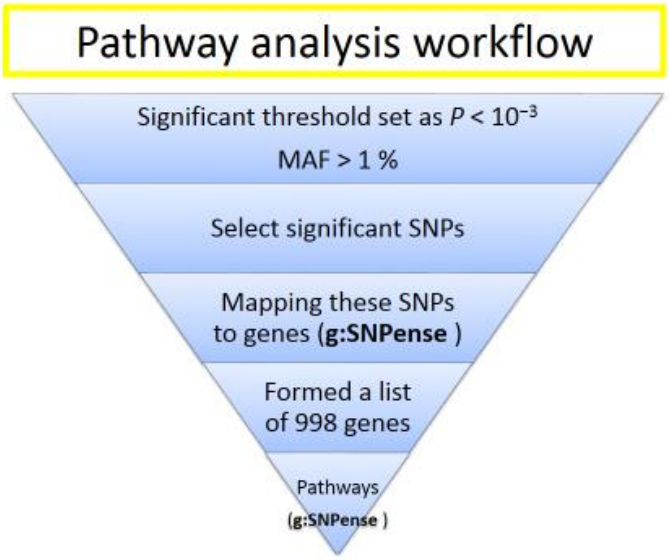


Figure.9 Gene enrichment and path analysis workflow.

To achieve this, first step was setting a SNP prioritization threshold of genetic association significance to select SNPs **(Fig.9)**. So, after the selection of the moderately significant Single Nucleotide Variations (SNVs) based on an *a priori* significance threshold ($P < 10^{-3}$). They were mapped to (sometimes multiple) genes (**Fig.10).** A variant was deemed as mapped to a gene when it was located within 50 kb of coding sequences of the gene.

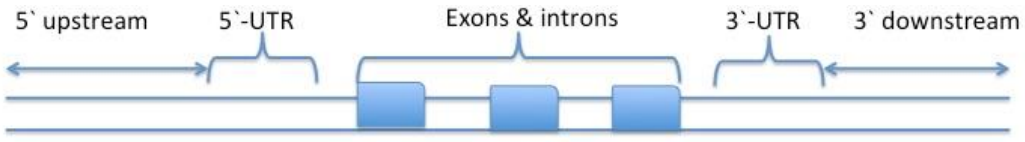


Figure.10 Schematic of gene regions.

For this purpose, **g:SNPense** was used for SNP mapping to genes (**Fig.11**). This SNP mapping tool is an online SNP IDs converting application (http://biit.cs.ut.ee/gsnpense). It accepts

a list of at most 4,000 rs IDs as a query. This list of rs IDs were converted into a gene list. g:SNPense mapped all SNPs residing within the 50kb upstream and downstream region of the coding sequence of a gene, including 3` and 5`-UTR regions mapped to a single or multiple genes (**Fig.10**). This allowed us to form a gene list to be used as a query for gene set enrichment analyses.

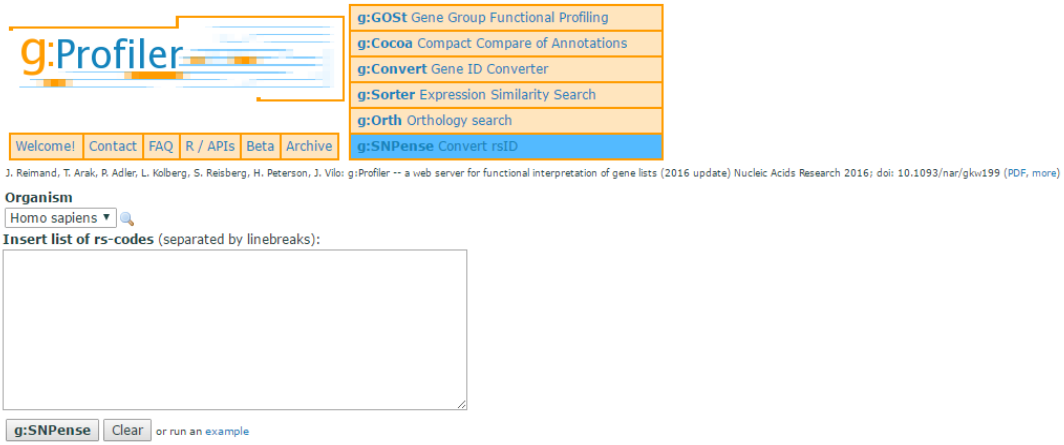


Figure.11 SNV mapping to genes (g:SNPense).

In order to test the list of genes for enrichment in any biological pathways or curated gene sets, an online tool called Gene Group Functional Profiling (**g:GOSt**) version r1730_e88_eg35 (Ensembl 88, Ensembl Genomes 35 (rev 1730, build date 2017-05-18)) was used for gene set enrichment analysis (**Fig.12)** (Reimand et al., 2016). If needed, there are older versions of g:GOSt in the archive tab on the g:Profiler website.

All g:GOSt queries had the same options selected as shown in **Fig.12**. Selected SNPs from FS analysis mapped to 998 genes. For the CC analysis, the query had 1,147 genes. The two gene lists are in the Appendices. From the options list of g:GOSt, for gene set enrichment analysis, there were only GO (Gene Ontology) terms and two biological pathways (KEGG and Reactome) selected.

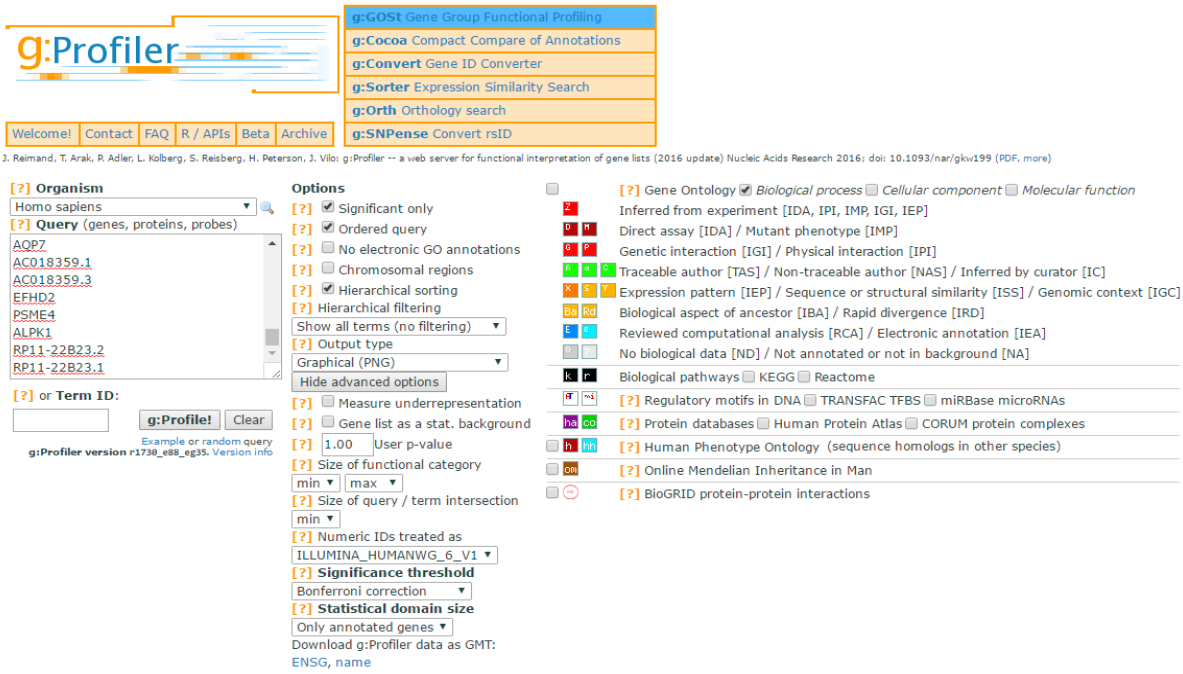


Figure.12 Gene enrichment analysis (g:GOSt).

Among other options, three can significantly affect the results of the gene enrichment analysis. These options were ordered query option, including Electronic GO annotations and multiple testing adjustment method of p-values in the gene enrichment analysis. The key point in the query was the selection of the ordered (ranked) query option. This allowed us to perform a test for enrichment based on an ordered query of the descending statistical significance of genetic associations. The selected options for the gene set enrichment analysis were the same as in **Fig.12**. The ambiguous gene identifiers (IDs) were resolved manually. By default, g:GOSt searches query gene IDs in 116 different databases. In queries, gene IDs were either HGNC (HUGO Gene Nomenclature Committee) gene symbol or Ensemble gene ID. Duplicates of gene IDs were automatically discarded by g:GOSt.

The test for significance of the pathway enrichment in the query gene list is based on a hypergeometric test. In order to show enrichment results separately for each category (source) of

GO term group and pathway gene set, we queried the gene lists once at a time selecting only one GO term category or the pathway gene set as the query option on the g:GOSt option panel.

g:GOSt can provide adjusted p-values for multiple testing based on three options, which are Bonferroni's, Benjamini–Hochberg False Discovery Rate (FDR) (Benjamini, Drai, Elmer, Kafkafi, & Golani, 2001) and a modified method (g:SCS method) developed by the g:Profiler research group (Reimand, Kull, Peterson, Hansen, & Vilo, 2007). For the gene set enrichment analyses, Bonferroni's method was method to adjust enrichment p-values for multiple testing.

For validation purposes, an additional web based tool from Max Planck Institute called ConsensusPathDB (Herwig, Hardt, Lienhard, & Kamburov, 2016) that shows the raw (unadjusted) p-values along only with their FDR adjustment. The significance for the enrichment analysis was set as q-value < 0.05. Thus, only for the purpose of this comparison, the change in options panel were deselection of the ordered query option and switching the multiple testing method from Bonferroni to FDR in g:GOSt.

4. Results

Two different Cox PH analyses, denoted **Full Sample (FS)** and **Case Cohort (CC)**, were implemented in this study to uncover modifier SNPs. Survival GWAS did not yield any genome wide significant SNPs, however, there seemed to be several suggestive SNVs (p-value < $5x10^{-5}$) associated with the AAO of the MDD.

4.1. Genome wide survival analysis results

4.1.1 All Sample analysis results

Quantile-quantile plot (Q-Q plot) showed (**Fig.13)** some departures from the null expectation (shaded area shows the 95 % confidence interval). The early deviation from the diagonal line points out that the p-values did not follow the expected null distribution. (Genomic inflation was estimated to be 1.075) This departure was either due to enrichment or to liberal tests. However, given that the permutations p-values were much larger than survival regression p-values, it is likely that the departure was due to liberal survival tests.

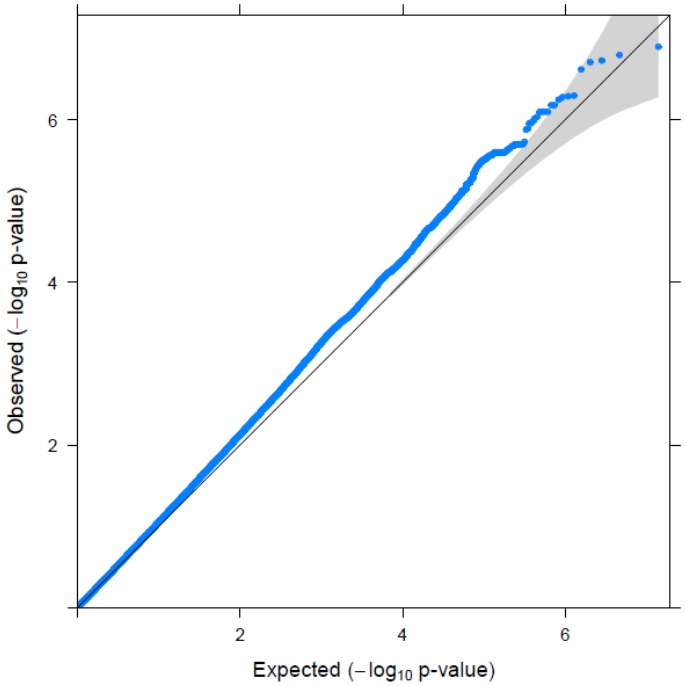


Figure.13 Quantile-quantile plot of p-values from summary statistics

Table.1 30 most significant SNVs in FS analysis.

CHR	SNV	BP	MAF	Hazard Ratio (HR)	Z score	*P* (unadjusted)
10	rs11245287	126246537	0.261	0.880	-5.286	$1.25x10^{-7}$
10	rs35841851	126236419	0.296	0.885	-5.242	$1.59x10^{-7}$
10	rs11245283	126236663	0.297	0.886	-5.214	$1.85x10^{-7}$
10	rs35936514	126244970	0.260	0.882	-5.203	$1.96x10^{-7}$
10	rs12258489	126245297	0.260	0.883	-5.167	$2.38x10^{-7}$
1	rs74359973	187629589	0.011	1.568	5.026	$5.01x10^{-7}$
6	rs55800092	4386107	0.150	0.861	-5.015	$5.30x10^{-7}$
10	rs12262706	126247468	0.292	0.890	-5.003	$5.64x10^{-7}$
6	rs17138114	4379511	0.151	0.862	-4.973	$6.61x10^{-7}$
6	rs7747061	4381445	0.151	0.862	-4.973	$6.61x10^{-7}$
6	rs56222106	4374034	0.150	0.863	-4.935	$8.02x10^{-7}$
6	rs75592374	4374582	0.150	0.863	-4.935	$8.02x10^{-7}$
6	rs1888325	4374751	0.150	0.863	-4.935	$8.02x10^{-7}$
6	rs78823440	4375064	0.150	0.863	-4.935	$8.02x10^{-7}$
6	rs1034115	4391428	0.150	0.864	-4.910	$9.12x10^{-7}$
6	rs170950	4391144	0.150	0.864	-4.903	$9.44x10^{-7}$
6	rs77648291	4375421	0.145	0.861	-4.886	$1.03x10^{-6}$
23	rs140957023	140958072	0.017	1.498	4.874	$1.10x10^{-6}$
1	rs181909501	48949127	0.020	6.235	4.842	$1.28x10^{-6}$
1	rs78146918	187610780	0.011	1.538	4.839	$1.30x10^{-6}$
13	rs116500056	107657974	0.017	1.413	4.769	$1.86x10^{-6}$
13	rs16969523	107658220	0.017	1.412	4.756	$1.97x10^{-6}$
13	rs35061615	107658285	0.017	1.412	4.756	$1.97x10^{-6}$
13	rs35932768	107658655	0.017	1.412	4.756	$1.97x10^{-6}$
13	rs61967003	107659212	0.017	1.412	4.756	$1.97x10^{-6}$
13	rs16969540	107662658	0.017	1.412	4.756	$1.97x10^{-6}$
13	rs12861527	107663978	0.017	1.412	4.756	$1.97x10^{-6}$
13	rs7990181	107667344	0.017	1.412	4.756	$1.97x10^{-6}$
10	rs80309727	107667344	0.017	1.104	4.746	$2.08x10^{-6}$
8	rs59341197	23461233	0.278	0.895	-4.742	$2.11x10^{-6}$

BP: base pair, **SNV:** Single Nucleotide Variant, ***P*:** p-value, **MAF:** Minor Allele frequency, **CHR:** Chromosome

The 30 most significant SNVs (**Table 1**) were all suggestive ($P < 5x10^{-5}$) genome wide significant associations. These SNVs are common variants and have MAF greater than 1 %, whereas SNVs on chromosome 13, 1 and 23 have MAFs that are close to 1%. The most significant SNV association was on Chromosome 10 in *LHPP* locus. This locus was also reported as having

genome wide significant association signal for SNVs inside an *LHPP* intron (consortium, 2015). Another point is that SNVs on *LHPP* with HR < 1 implicates the tested allele as being protective.

Table.2 Genes located near suggestive SNPs in FS analysis.

SNP ID	Gene	Gene region	P value	MAF	Gene	Chr	Gene Association *
rs11245287	LHPP	intron	$1.247x10^{-7}$	0.261	phospholysine phosphohistidine inorganic pyrophosphate phosphatase	10q26.13	cholesterol, anticoagulant
rs140957023	MAGEC3	intron	$1.096x10^{-6}$	0.017	MAGE family member C3	Xq27.2	NA
rs78146918	ERVMER61-1	intron	$1.305x10^{-6}$	0.011	endogenous retrovirus group MER61 member 1	1q31.1	NA
rs80309727	RNU6-523P	intron	$2.075x10^{-6}$	0.453	RNA, U6 small nuclear 523, pseudogene	10q21.3	NA
rs59341197	RNU4-71P	upstream	$2.114x10^{-6}$	0.278	RNA, U4 small nuclear 71, pseudogene	8p21.2	NA
rs192525648	CCL17	intron	$1.305x10^{-6}$	0.011	C-C motif chemokine ligand 17	16q21	NA
rs118149296	ADAM23	intron	$2.289x10^{-6}$	0.018	ADAM metallopeptidase domain 23	2q33.3	NA
rs56197202	RNU4-71P	intron	$2.327x10^{-6}$	0.273	coiled-coil domain containing 88C	14q32.11-q32.12	Dehydroepiandrosterone, Albuminuria

*NCBI Phenotype-genotype integrator NHGRI or dbGAP p-value<$5x10^{-5}$ **Chr**: Chromosome **MAF:** Minor Allele Frequency

The majority of the most significant variants that were mapped to genes were in/near genes that have no known association with any disease phenotype and have low MAF (~1%) (**Table.2**). Exceptions are rs11245287 and rs56197202, which are common variants with MAF around 26%. These two common variants are located in introns of *LHPP* and *RNU4-71P* respectively. The SNP in *LHPP* was the most significant one.

According to the NCBI Phenotype-genotype integrator, https://www-ncbi-nlm-nihgov.proxy.library.vcu.edu/gap/phegeni?tab=1&gene=64077, rs4315021 variant in LHPP is associated with (p-value = 3.06 x 10^{-5}) total cholesterol concentration in blood. This result particularly belongs to NHLBI Family Heart Study (dbGaP Study Accession: phs000221.v1.p1

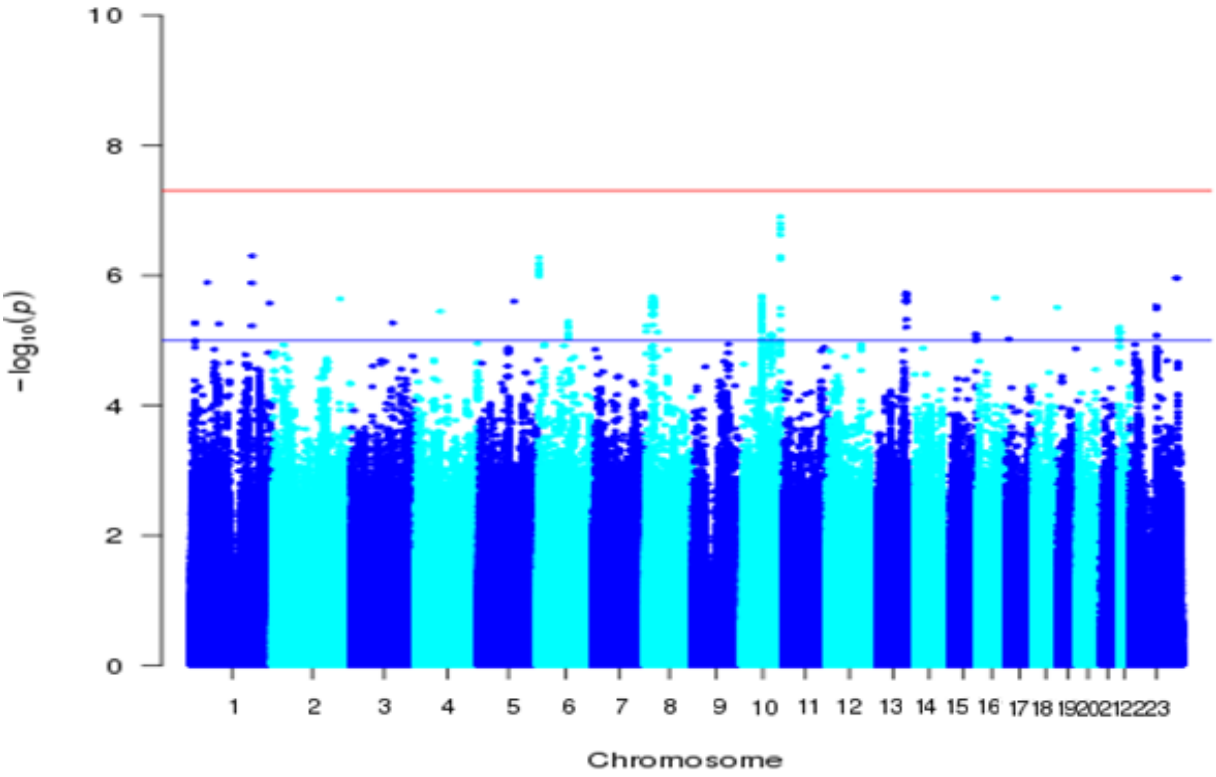


Figure.14 Manhattan plot of $-\log_{10}(P)$ for FS analysis. The red horizontal line points out the genome wide significance threshold (P = $5x10^{-8}$) and the blue horizontal line signs the suggestive significance level of genetic association (P = $5x10^{-5}$).

As mentioned before, for obtaining variant association statistics there were both case and control subjects included in the AS analysis **(Fig.14)**. LocusZoom online tool generated regional association plots allow us to explore the loci showing suggestive genome-wide significant associations. (**Fig. 15-18**). LocusZoom used all analyzed SNVs residing within the locus and p-values used as an input for regional association locus plotting (Pruim et al., 2010). To generate these plots, LD population was set to ashg19/1000 Genomes Nov 2014 ASN (East Asians).

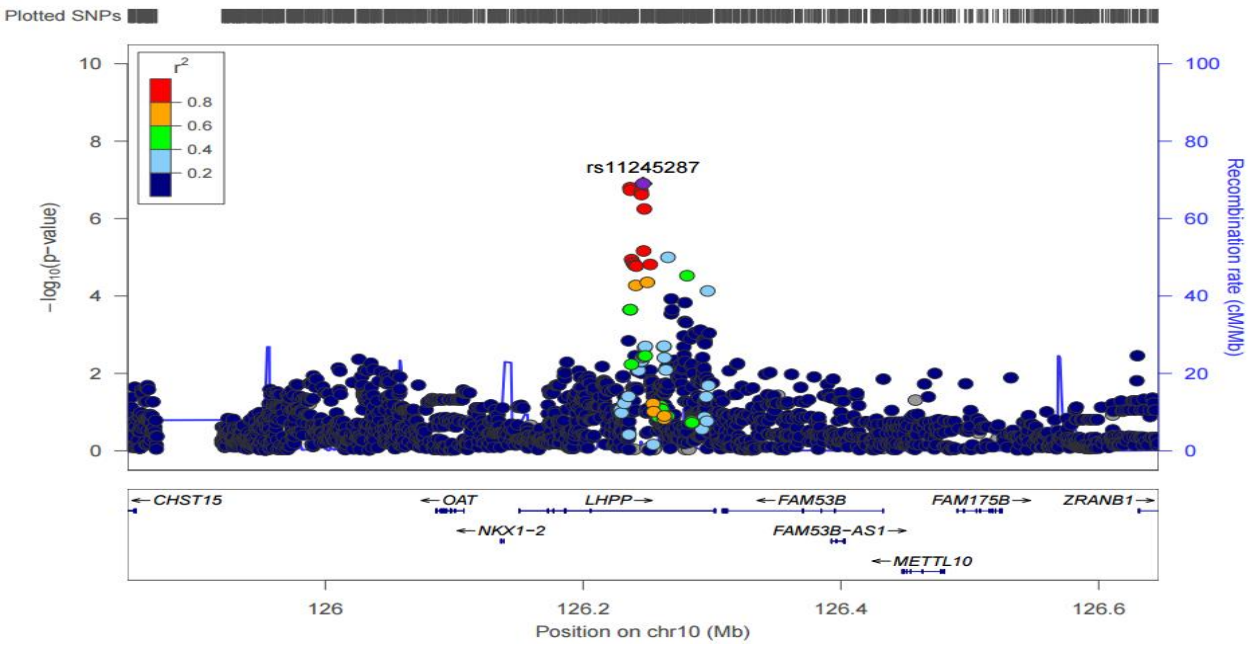


Figure.15 Locus zoom for *LHPP*.

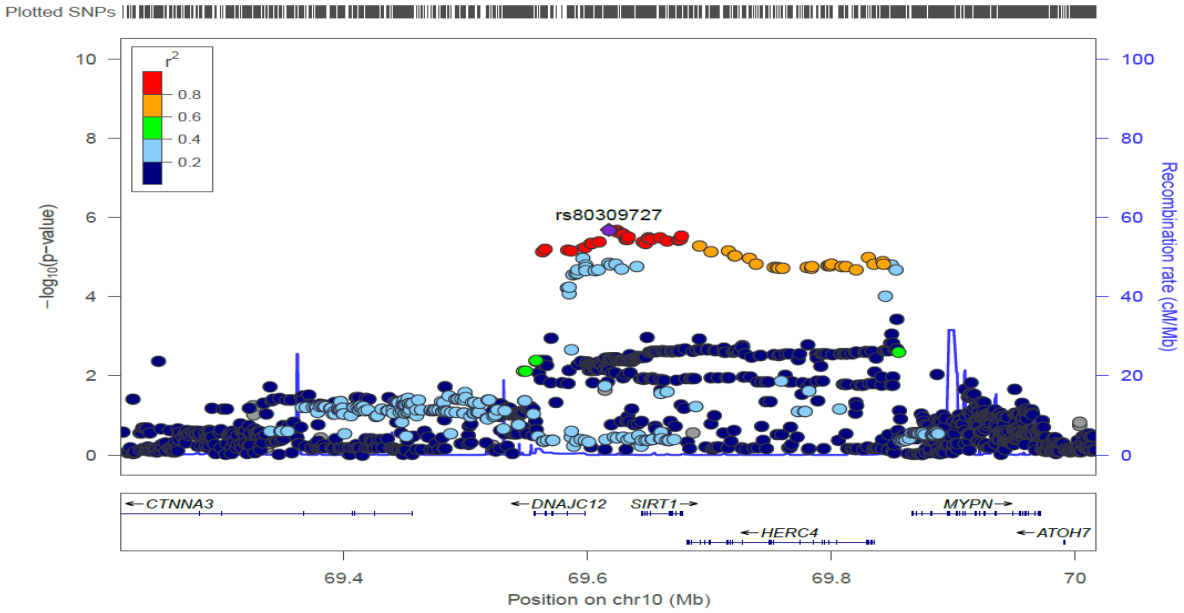


Figure.16 Locus zoom for *SIRT1*.

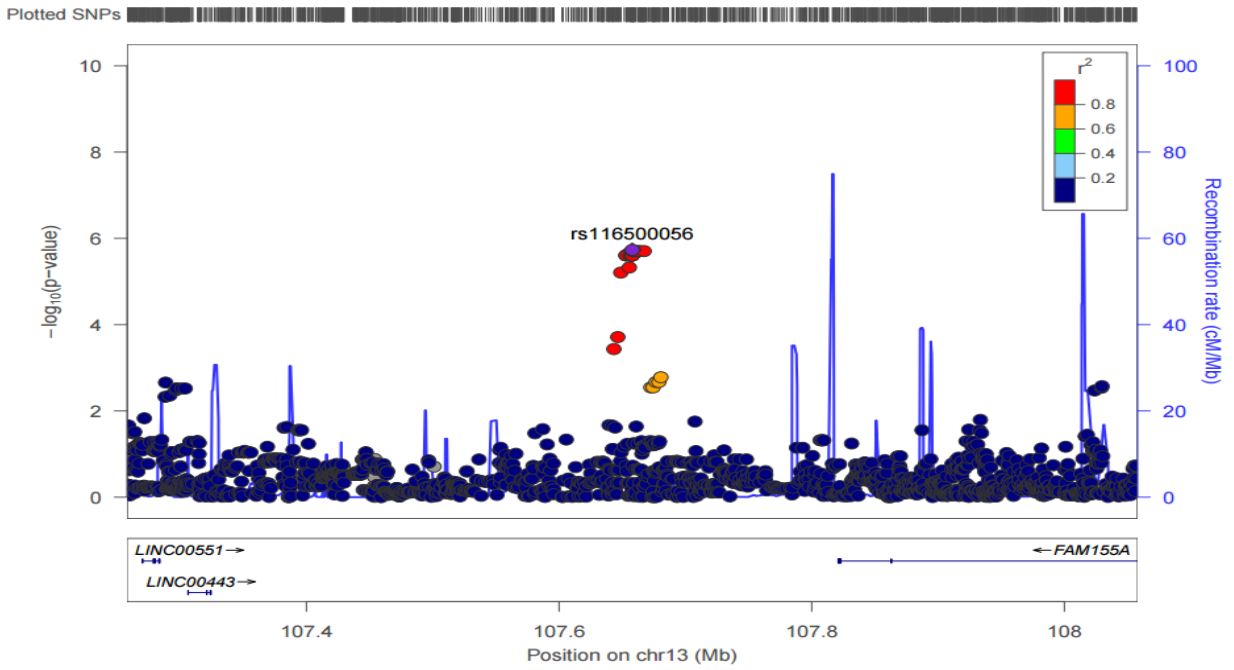


Figure.17 Locus zoom for chr13:107,257,974−108,057,974.

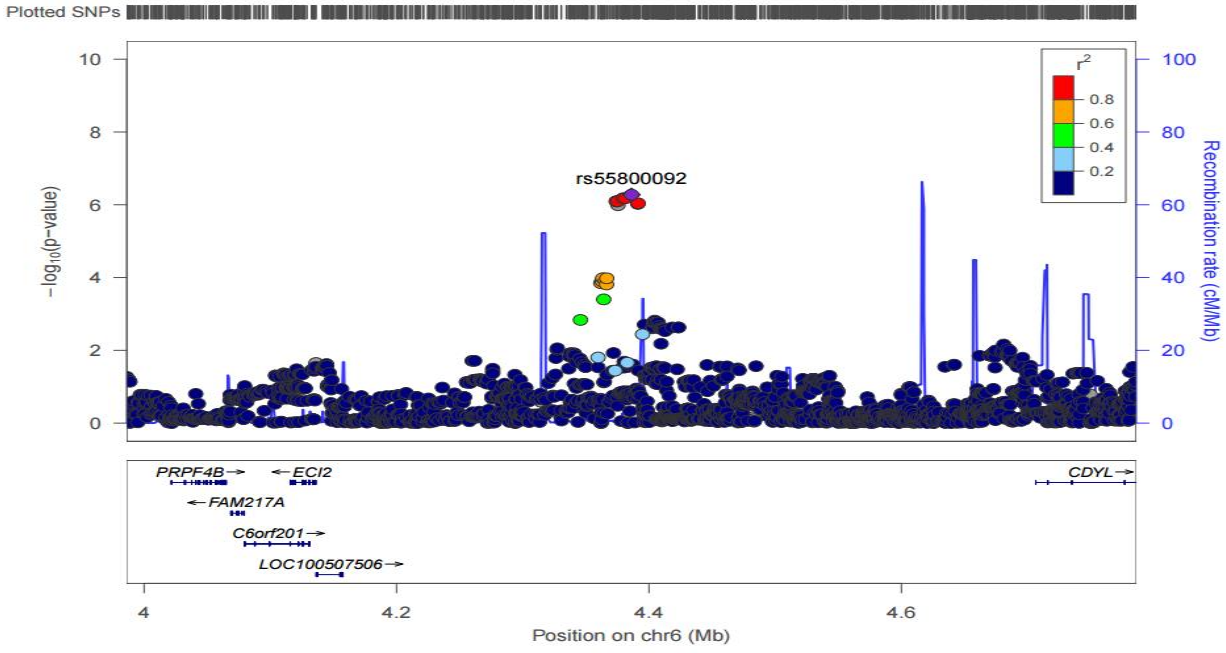


Figure.18 Locus zoom for chr6:3,986,107−4,786,107 bp.

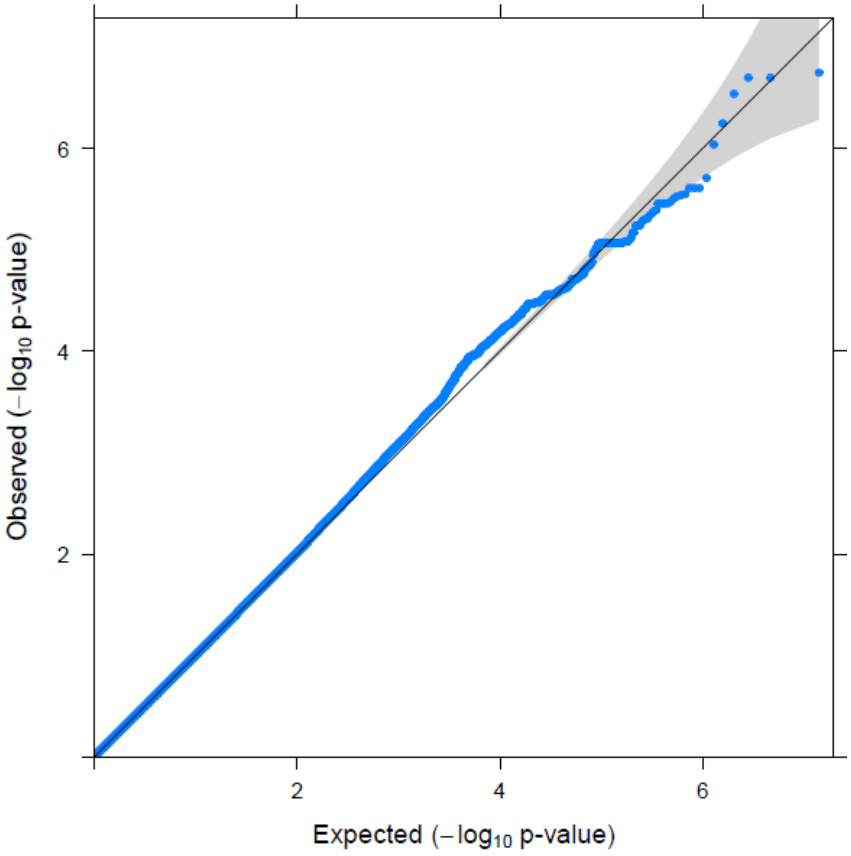


Figure.19 Quantile-quantile plot for CC analysis.

4.1.2 Case Cohort analysis results

CC analysis also showed some departure from null expectation **(Fig.19)**. Genomic inflation was $\lambda = 1.014$. As shown in **Fig.19**, there was a deflation in the association p-values towards the right tail of the Q-Q plot, however, the top six significant SNV associations, were within the 95% confidence interval of the expected normal distribution.

Table.3 30 most significant SNVs from the CC analysis.

CHR	SNP	BP	MAF	Hazard Ratio (HR)	Z score	*P* (unadjusted)
7	rs690911	49939147	0.0562	1.264	5.219	$1.80x10^{-7}$
7	rs2366025	50080818	0.0563	1.264	5.200	$1.99x10^{-7}$
7	rs6944334	50018441	0.0560	1.264	5.200	1.9910^{-7}
7	rs692270	49892044	0.0576	1.257	5.132	$2.86x10^{-7}$
7	rs691417	49887533	0.0562	1.254	5.002	$5.69x10^{-7}$
3	rs151020965	11904509	0.0222	1.427	4.908	$9.21x10^{-7}$
21	rs74591900	32762165	0.0265	1.439	4.760	$1.93x10^{-6}$
7	rs72590997	52362582	0.1270	0.864	-4.712	$2.45x10^{-6}$
7	rs72590998	52362961	0.1270	0.864	-4.712	$2.45x10^{-6}$
7	rs72591000	52379298	0.1269	0.864	-4.712	$2.45x10^{-6}$
7	rs9642460	52384825	0.1271	0.864	-4.684	$2.82x10^{-6}$
16	rs77700579	81490144	0.0952	1.181	4.679	$2.88x10^{-6}$
7	rs7778784	50135200	0.0577	1.231	4.675	$2.94x10^{-6}$
12	rs11173011	40046408	0.2019	1.132	4.664	$3.10x10^{-6}$
6	rs36142524	101946461	0.0218	1.446	4.648	$3.34x10^{-6}$
7	rs2221656	50099404	0.0579	1.229	4.640	$3.48x10^{-6}$
14	rs1285804	91819984	0.0161	1.436	4.639	$3.49x10^{-6}$
14	rs1285806	91822382	0.0161	1.436	4.639	$3.49x10^{-6}$
14	rs1285807	91823106	0.0161	1.436	4.639	$3.49x10^{-6}$
14	rs1285811	91830078	0.0161	1.436	4.639	$3.49x10^{-6}$
12	rs7295237	71016832	0.1204	1.176	4.609	$4.05x10^{-6}$
2	rs4666290	30149295	0.2274	1.133	4.601	$4.21x10^{-6}$
7	rs79417164	50170934	0.0282	1.345	4.592	$4.40x10^{-6}$
12	rs75833894	86726432	0.0461	1.276	4.582	$4.60x10^{-6}$
2	rs13414785	30149187	0.2323	1.132	4.570	$4.87x10^{-6}$
14	rs1285768	91797825	0.0123	1.515	4.569	$4.90x10^{-6}$
5	rs2135026	155810433	0.1743	1.141	4.563	$5.05x10^{-6}$
12	rs7978510	40038294	0.2018	1.130	4.558	$5.16x10^{-6}$
9	rs1819343	117328001	0.3978	0.908	-4.543	$5.55x10^{-6}$
7	rs1483080	50045825	0.0641	1.208	4.533	$5.82x10^{-6}$

BP: Base Pair, **SNV:** Single Nucleotide Variant, ***P*:** p-value, **MAF:** Minor Allele Frequency, **CHR:** Chromosome

The 30 most significant SNVs in CC analysis (**Table.3**) were all suggestive ($P < 5x10^{-5}$) genome wide significant associations. The most significant SNV association was Chromosome 7 rs690911 ($P = 1.80x10^{-7}$, HR=1.264). This locus have two overlapping genes that are *ZPBP* and *VWC2.*

Table.4 SNVs mapped to genes in the CC analysis.

SNP ID	Gene	Gene region	MAF	p-value	Gene	Chr	Gene Association*
rs690911	*ZPBP/ VWC2*	intron	0.056	1.796×10^{-7}	zona pellucida binding protein / von Willebrand factor C domain containing 2	7p12.2	NA/cholesterol, triglycerides, sleep, calcium, platelet aggregation
rs151020965	*FANCD2P2*	intron	0.022	9.209×10^{-7}	Fanconi anemia complementation group D2 pseudogene 2	3p25.2	NA
rs74591900	*TIAM1*	intron	0.027	1.935×10^{-6}	T-cell lymphoma invasion and metastasis 1	21q22.11	ALS, Hip, Neuroblastoma, coronary disease, lipids
rs77700579	*CMIP*	intron	0.095	2.881×10^{-6}	c-Maf inducing protein (plays a role in T-cell signaling pathway)	16q23.2-q23.3	Adiponectin, cholesterol, HDL, Type 2 Diabetes, Body height
rs7778784	*C7ORF72*	upstream	0.058	2.938×10^{-6}	chromosome 7 open reading frame 72	7p12.2	NA / SLE, Chrone disease, Hippocampus
rs11173011	*C12ORF40*	intron	0.202	3.103×10^{-6}	chromosome 12 open reading frame 40	12q12	NA
rs36142524	*GRIK2*	intron	0.022	3.344×10^{-6}	kainate type subunit 2ily tyrosine kinase	6q16.3	Body weight, Cholestrol, LDL, Lipoprotein, FSH, BMI
rs1285804	*CCDC88C*	intron	0.016	3.493×10^{-6}	coiled-coil domain containing 88C	14q32.11 -q32.12	Insulin, insulin resistance, BMI

***NCBI Phenotype-genotype integrator NHGRI or dbGAP p-value < 5×10^{-5} **Chr**: Chromosome **MAF:** Minor Allele Frequency

Most of the significant variants in CC analysis are known to exhibit significant association with other disease phenotypes (**Table.4**). Most of these SNPs are rather rare (MAF < 10%) with the exception of rs11173011, which is a common variant with MAF = 26%. This common SNP resides in *C12ORF40* (chromosome 7 open reading frame 40). NCBI Phenotype-genotype integrator indicates that there are significant associations between variants in *C12ORF40* and Systemic Lupus Erythematosus (SLE), Crohn's disease.

The most significant association signal was located in one of the introns of von Willebrand factor C domain containing 2 (*VWC2*). According to the NCBI Phenotype-genotype integrator, there is a significant gene association between variants in *VWC2* loci and cholesterol, triglycerides, sleep, calcium and platelet aggregation.

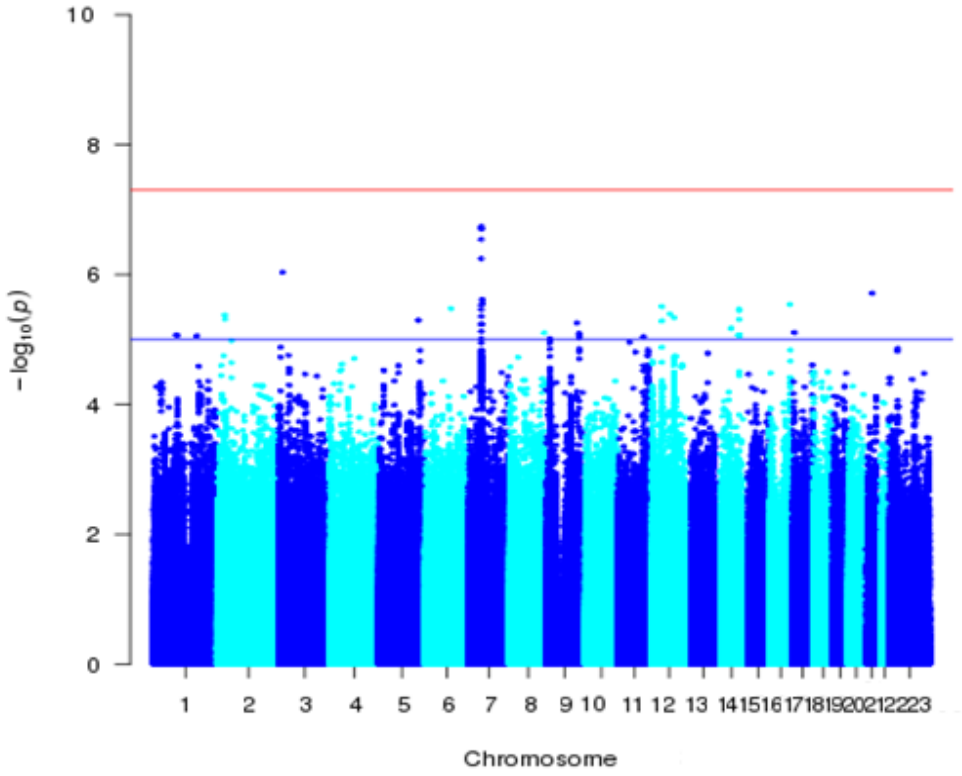


Figure.20 Manhattan plot of $-\log_{10}(P)$ from CC analysis.

CC analysis yielded the best signal on chromosome 7, which is, however, only suggestive (**Fig.20**). Locus zoom plot (**Fig.21**) explored the SNVs, which are in Linkage Disequilibrium with the most significant SNV (rs690911) in the *VWC2* loci, nearby the *ZPBP* and *C7ORF72 loci*.

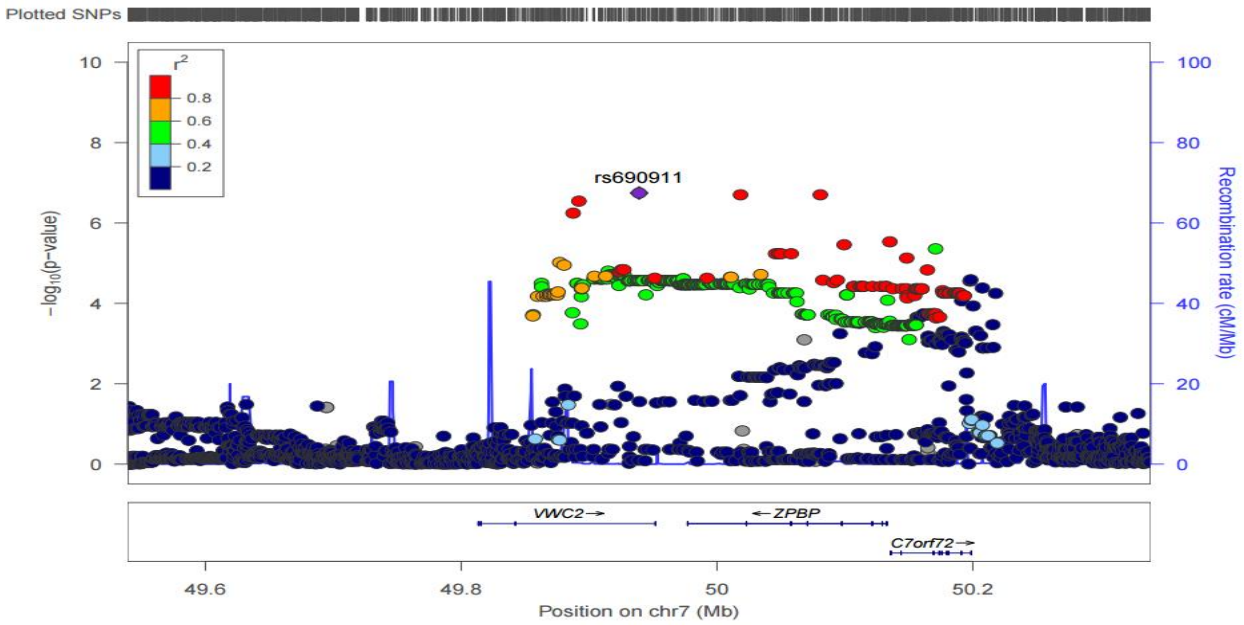


Figure.21 Locus zoom for *VWC2*, *ZPBP* and *C7ORF72*.

4.2 Gene Set Enrichment Results

Gene set enrichment analysis results are shown in summary (**Table.5)** and in **Fig.23** pertaining to the graphical outputs obtained from the g:GOSt (**Fig.12**). The remaining results are in Appendices. [The gene lists associated with the two analyses are in Appendices (**Table.7** and **Table.8**).]

In summary, **Table.5** shows the top significant gene set enrichments of GO term categories and Pathway (KEGG and Reactome) gene sets in each gene list of two models. Most of the significant enrichments were found in *Cellular Component* and *Biological Process* GO categories. The most significant enrichment for FS analysis was cell periphery (GO:0071944, $q = 3.23 \times 10^{-5}$). The most significant enrichment for CC analysis was synapse (GO:0045202, $q = 1.33 \times 10^{-5}$). Neuronal system (REAC:112316) was significantly enriched in both analyses.

Table.5 Summary table of gene set enrichment results.

Model	GO Term / Pathway ID	GO Term category	Term or Pathway name	Gene set	Gene list	Genes Shared	q_B	q_{FDR}
FS	**GO:0071944**	Cellular Component	cell periphery	5283	795	180	$3.23x10^{-5}$	$6.66x10^{-8}$
	GO:0044700	Biological Process	single organism signaling	6263	396	110	$9.47x10^{-4}$	$9.63x10^{-7}$
	GO:0065008	Biological Process	regulation of biological quality	3428	380	69	$1.04x10^{-3}$	$1.06x10^{-6}$
	GO:0097458	Cellular Component	neuron part	5283	795	180	$1.10x10^{-3}$	$1.12x10^{-6}$
	GO:0045202	Cellular Component	synapse	809	944	50	$1.63x10^{-3}$	$1.65x10^{-6}$
	GO:0055085	Biological Process	transmembrane transport	1395	526	48	$3.77x10^{-3}$	$3.84x10^{-6}$
	REAC:112316	Reactome	neuronal system	359	886	26	$4.76x10^{-3}$	$8.74x10^{-4}$
	KEGG: 05231	KEGG	choline metabolism in cancer	101	664	10	$2.36x10^{-2}$	$1.35x10^{-3}$
CC	**GO:0045202**	Cellular Component	synapse	809	1049	60	$1.33x10^{-5}$	$1.98x10^{-8}$
	GO:0097458	Cellular Component	neuron part	1322	1130	91	$2.02x10^{-5}$	$3.01x10^{-9}$
	GO:0120036	Cellular Component	plasma membrane bounded cell projection organization	1323	1112	86	$4.59x10^{-5}$	$1.35x10^{-7}$
	GO:0007156	Biological Process	homophilic cell adhesion via plasma membrane adhesion molecules	156	509	16	$5.67x10^{-5}$	$8.44x10^{-8}$
	GO:0016358	Biological Process	dendrite development	203	1039	25	$7.08x10^{-5}$	$1.05x10^{-7}$
	KEGG:04724	KEGG	glutamergic synapse	114	855	14	$2.82x10^{-4}$	$8.46x10^{-6}$
	KEGG:04713	KEGG	circardian entrainment	96	855	13	$2.33x10^{-4}$	$6.98x10^{-6}$
	GO:0005509	Molecular Function	calcium ion binding	701	477	30	$5.38x10^{-3}$	$8x10^{-6}$
	REAC:392154	Reactome	Nitricoxide stimulates guanylate cyclase	25	188	4	$3.21x10^{-2}$	$6.66x10^{-4}$
	REAC:112316	Reactome	neuronal system	359	1078	27	$5x10^{-2}$	$1.04x10^{-3}$

GO: Gene Ontology **KEGG:** Kyoto Encyclopedia of Genes and Genomes q_B**:** Bonferroni adjusted p-value q_{FDR}: Benjamini–Hochberg False Discovery Rate adjusted p-value
Effective domain size / FDR significance: KEGG: 7,168 / $P = 3.36x10^{-3}$, REAC: 10,635 / $P = 3.68x10^{-4}$, GO:18,971 / $P = 6.28x10^{-3}$

All queries had the same panel of options for the gene set enrichment analysis as shown in **Fig.12**. The graphical output is in a matrix form with color coded cells designating the kind of annotation of the gene.

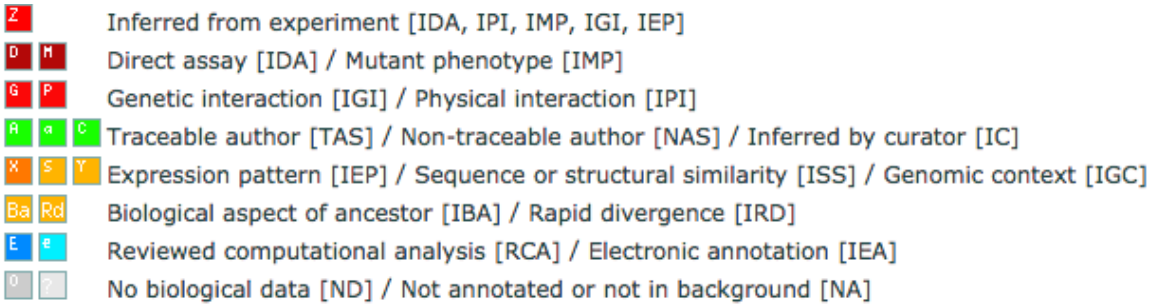


Figure.22 Functional gene annotation used in "Evidence code".

There are seventeen different annotations (**Fig.22**). These codes are called "Evidence code", which are helpful to interpret the gene enrichment analysis results in terms of reliability. In brief, the experimentally verified red color-coded annotations are more reliable than the blue color-coded ones, electronically annotated by *in silico* analysis.

The graphical outputs of g:GOSt FS enrichment analyses reveal significant GO terms and pathways (**Fig. 23, and Fig.25-28 in Appendices**). For instance, **Fig.23** shows a portion of the graphical output of g:GOSt from the ranked query of the gene list FS analysis. This output was the result of the query gene list of FS gene enrichment analysis when only the *Biological Process* GO term was selected on the options panel. The rest of the query genes are not on the g:GOSt graphical output figures due to large horizontal dimension of the graphical output from g:GOSt.

The most significant GO term among the *Biological Process* category (**Fig. 23)** was single organism signaling (GO:0044700, $q = 9.47 \times 10^{-4}$).

source	term name Gene Ontology (Biological process)	term ID	n. of term genes	n. of query genes	n. of common genes	corrected p-value
BP	taxis	GO:0042330	551	515	25	1.75e-02
BP	chemotaxis	GO:0006935	550	515	25	1.70e-02
BP	cell communication	GO:0007154	6287	396	109	2.42e-03
BP	signaling	GO:0023052	6275	396	110	1.06e-03
BP	single organism signaling	GO:0044700	6263	396	110	9.47e-04
BP	nervous system development	GO:0007399	2216	956	99	3.62e-02
BP	transmembrane transport	GO:0055085	1395	526	48	3.77e-03
BP	ion transmembrane transport	GO:0034220	1045	385	31	6.18e-03
BP	synapse organization	GO:0050808	249	401	14	1.29e-02
BP	localization	GO:0051179	6175	829	201	1.92e-02
BP	regulation of biological quality	GO:0065008	3428	380	69	1.04e-03
BP	regulation of membrane potential	GO:0042391	386	855	28	7.45e-03
BP	regulation of postsynaptic membrane potential	GO:0060078	135	577	12	3.58e-02
BP	cellular response to purine-containing compound	GO:0071415	9	337	4	9.45e-03
BP	cellular response to caffeine	GO:0071313	8	337	4	5.29e-03

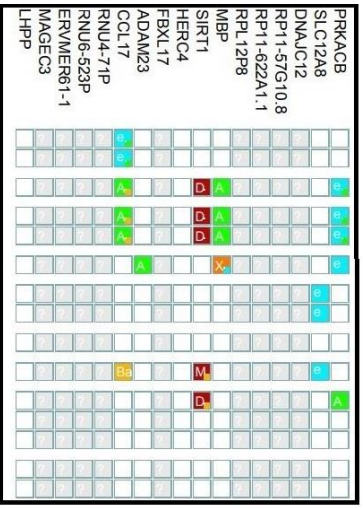


Figure.23 g:GOSt results for *Biological Process* GO terms in FS analysis.

The following two top two significantly enriched GO terms were regulation of biological quality (GO:0065008, $q = 1.04x10^{-3}$) and the transmembrane transport (GO:0055085, $q = 3.77x10^{-3}$). There were 1,395 genes in transmembrane transport GO term. The gene list query for this GO term consisted of 526 genes. Number of shared genes between the gene query list and the term gene set was 48 genes.

4.3 Cross platform comparison of gene enrichment analysis results between ConsensuspathDB and g:Profiler.

In order to assess the reliability of the g:Profiler platform, we used the same query gene list in another online gene set enrichment tool called Consensus Path DataBase (CPDB) from Max Planck Institute for Molecular Genetics. The slight difference between the results of the two platforms (**Table.6)** were mainly due to the number of genes recognized as the query genes and the number of genes by default in each GO term.

Table.6 A brief comparison of FS analysis enrichment results of the same queries between Consensus Path Database and g:Profiler.

GO term ID	GO term	g:Profiler/CPDB			
		gene set size	overlapping genes	p-value (unadjusted)	q-value
GO:0048648	Cell Development	1907 / 1859	87 / 87	NA / 5.79×10^{-6}	2.78×10^{-3}/2.41×10^{-3}
GO:0044700	single organism signaling	6263 / 6201	237 / 227	NA / 1.47×10^{-5}	3.53×10^{-4}/7.18×10^{-4}
GO:0071944	cell periphery	5283 / 5176	217 / 205	NA / 1.86×10^{-7}	1.96×10^{-6}/6.84×10^{-6}
GO:0097060	synaptic membrane	289 / 267	24 / 21	NA / 3.83×10^{-5}	4.95×10^{-4}/4.85×10^{-4}

GO: Gene Ontology **q-value:** Benjamini-Hochberg method for False discovery rate **NA:** Not applicable **CPDB:** Consensus Path DataBase

5. Discussion

This study aimed to use whole genome survival regression to uncover genetic modifiers of MDD AAO. To achieve the aim, we employed two Cox PH analyses Full Sample (FS) and Case Cohort (CC) only. These analyses used ancestry principal components (PCs) and child sexual abuse (CSA) as covariates. While there were no statistically significant univariate findings, aggregate analyses show enrichment in some of the expected places, e.g. neuronal and synaptic pathways. However, CC analysis revealed some interesting non-neuronal pathways that might provide avenues for future treatments.

The most significant genetic association in all univariate analyses occurred in the intronic region of *LHPP*. Previously, in the CONVERGE consortium GWAS, this locus has been reported as having genome wide significant SNP (consortium, 2015). The reason for getting similar genetic associations in the FS analysis would be that the FS analysis used the same genotype data and almost the same cohort. Maybe not unexpectedly, because using only cases, Cox PH CC analysis yielded rather different results from FS analysis.

The reason for not finding any genome wide significant loci in both analyses could be due to lack of power for this analysis. Similar to most initial studies of psychiatric disorders, an increase in the study sample size would uncover loci that have fallen just below the genome wide significant threshold in current study. (For our two analyses (FS and CC) we had 80% power to detect variants with Hazard Ratio > 1.4 and MAFs > 0.01.)

Unlike univariate results, the gene enrichment results from the two analyses were much more similar when we look at the gene set enrichments in GO term gene sets. The gene enrichment results for FS analysis pointed to cell periphery GO term (GO:0071944, $q = 3.23 \times 10^{-5}$), synapse (GO:0045202, $q = 1.63 \times 10^{-3}$), neuron part (GO:0097458, $q = 1.10 \times 10^{-3}$), transmembrane transport

(GO:0055085, $q = 3.77x10^{-3}$), regulation of biological quality (GO:0065008, $q = 1.04x10^{-3}$), single organism signaling (GO:0044700, $q = 9.47x10^{-4}$), Choline metabolism in cancer (KEGG:05231, $q = 2.36x10^{-2}$) and Neuronal system (REAC:112316, $q = 4.76x10^{-3}$).

CC analysis showed significant enrichment of Plasma membrane bounded cell projection organization (GO:0120036, $q = 4.59x10^{-5}$), Dendrite development (GO:0016358, $q = 7.08x10^{-5}$), Homophilic cell adhesion via plasma membrane adhesion molecules (GO:0007156, $q = 5.67x10^{-5}$), Neuron part (GO:0097458, $q = 2.02x10^{-5}$), Synapse (GO:0045202, $q = 1.33x10^{-5}$), Calcium ion binding (GO:0005509, $q = 5.38x10^{-3}$), Glutamergic synapse (KEGG:04724, $q = 2.82x10^{-4}$), Circadian entrainment (KEGG:04713, $q = 2.33x10^{-4}$), nitric oxide stimulates guanylate cyclase (REAC:392154, $q = 5x10^{-2}$) and Neuronal system (REAC:112316, $q = 3.21x10^{-2}$) the gene sets in query gene list. For pathway enrichment (KEGG and Reactome) Glutamergic synapse (KEGG:04724, $q = 2.82x10^{-4}$) and Circadian entrainment (KEGG:04713, $q = 2.33x10^{-4}$) showed significant signals. Given that CC analysis did not assume that controls would develop the diseases eventually, we believe that this analysis provided somewhat more reliable results for validation (and, possibly, treatment).

CC pathway enrichment analysis gave us two significantly enriched terms that are biologically relevant to the MDD pathology. These two terms were Glutamergic synapse (KEGG:04724, $q = 2.82x10^{-4}$) and dendrite development (GO:0016358, $q = 7.08x10^{-5}$). A copy number variation (CNV) study (Glessner et al., 2010), supports Glutamergic synapse. In this research, the duplication of chromosomal region (5q35.1) spanning the *SLIT3* gene, which codes for a protein responsible for cell migration and axon guidance, was observed in 5 unrelated individuals with MDD. A pathway analysis on genome-wide association datasets reported that Neuroticism association results indicates a significant enrichment in axon guidance Reactome gene

set (Kim et al., 2015). A genome wide gene expression study indicated that the postmortem brain samples from individuals who committed suicide with or without MDD shows a significant change in gene expression of genes involved in GABAergic (inhibitory) and glutamatergic (excitatory) neurotransmission (Sequeira et al., 2009).

Functionally similar genes could be on chromosomal regions close to each other. These genes with high LD may inflate the findings in gene set enrichment analysis. We have not accounted for the LD structure of SNVs in the gene set enrichment analysis. Another drawback of the gene set enrichment analysis is that we could not map all the SNVs to genes. Unmapped intergenic SNVs were eliminated in the SNP mapping to gene step (g:SNPense). Also, the SNPs which coincide within the overlapping or neighboring genes at extremely close proximity to each other may or may not be mapped to all genes in that chromosomal region.

In gene enrichment results, the same group of genes intersected multiple different GO terms neighboring function or structures. This could be observed in FS analysis in which *Biological Process* term had significant enrichment in Cell communication (GO:0007154) and Signaling (GO:0023052) GO terms. These two GO terms obviously have overlapping genes in their gene sets. Because of this, they both came out as significantly enriched GO terms. This same idea also applies to the result of gene set enrichment for CC analysis. The *Cellular component* terms Synapse (GO:0045202, $q = 1.33 \times 10^{-5}$) and Neuron part (GO:0097458, $q = 2.02 \times 10^{-5}$) were the most significantly enriched gene sets in the query gene list for CC analysis. It is highly probable that these two terms have overlapping gene sets, and so both of them are highly enriched.

The strength of the current study relies on a well-characterized and recruited case group. The Han Chinese population is relatively more homogenous than most of the other ethnic groups in terms of genetic makeup. We claim that the more homogenous the study population means the

more reliable the genetic epidemiology study. However, a significant association finding in one ethnic group may not replicate in another ethnic group. So, the findings might not generalize to non-Han Chinese population.

6. Conclusion

Sequencing, rather than array based genotyping methods, allows us to study practically all non-coding and coding variants in the genome. Two statistical analyses yielded different, albeit non-significant, univariate results were due to differences in cohorts analyzed. The pathway enrichment results indicated that the suggestive univariate signals showed significant enrichment in nervous system cell biology and function. However, suggestive variants sometimes were mapped to multiple genes and adjustment for LD structure was not performed. Thus, an additional replication cohort may be needed to validate the significant enrichment findings. Nonetheless, aggregate results mostly mirror MDD research literature explaining MDD biology as a disorder of synaptic transmission systems, ion channels, receptors and dendrite development. However, some of the non-neural pathways (e.g. nitric oxide) from the case cohort analysis might provide good targets for new treatments.

References

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Arlington, VA: Author.

Anthony, J. C., & Petronis, K. R. (1991). Suspected risk factors for depression among adults 18-44 years old. *Epidemiology, 2*(2), 123-132.

Avenevoli, S., & Merikangas, K. R. (2006). Implications of high-risk family studies for prevention of depression. *Am J Prev Med, 31*(6 Suppl 1), S126-135. doi:10.1016/j.amepre.2006.07.003

Benjamini, Y., Drai, D., Elmer, G., Kafkafi, N., & Golani, I. (2001). Controlling the false discovery rate in behavior genetics research. *Behav Brain Res, 125*(1-2), 279-284.

Besedovsky, H., del Rey, A., Sorkin, E., & Dinarello, C. A. (1986). Immunoregulatory feedback between interleukin-1 and glucocorticoid hormones. *Science, 233*(4764), 652-654.

Brown, E. S., Rush, A. J., & McEwen, B. S. (1999). Hippocampal remodeling and damage by corticosteroids: implications for mood disorders. *Neuropsychopharmacology, 21*(4), 474-484. doi:10.1016/S0893-133X(99)00054-8

Bruce, M. L., Takeuchi, D. T., & Leaf, P. J. (1991). Poverty and psychiatric status. Longitudinal evidence from the New Haven Epidemiologic Catchment Area study. *Arch Gen Psychiatry, 48*(5), 470-474.

Cai, N., Bigdeli, T. B., Kretzschmar, W. W., Li, Y., Liang, J., Hu, J., . . . Flint, J. (2017). 11,670 whole-genome sequences representative of the Han Chinese population from the CONVERGE project. *Sci Data, 4*, 170011. doi:10.1038/sdata.2017.11

Chen, J., Cai, Y., Cong, E., Liu, Y., Gao, J., Li, Y., . . . Flint, J. (2014). Childhood sexual abuse and the development of recurrent major depression in Chinese women. *PLoS One, 9*(1), e87569. doi:10.1371/journal.pone.0087569

Collins, F. S., Guyer, M. S., & Charkravarti, A. (1997). Variations on a theme: cataloging human DNA sequence variation. *Science, 278*(5343), 1580-1581.

consortium, C. (2015). Sparse whole-genome sequencing identifies two loci for major depressive disorder. *Nature, 523*(7562), 588-591. doi:10.1038/nature14659

http://www.nature.com/nature/journal/v523/n7562/abs/nature14659.html#supplementary-information

Cox, D. R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological), 34*(2), 187-220.

de Kloet, E. R., Joels, M., & Holsboer, F. (2005). Stress and the brain: from adaptation to disease. *Nat Rev Neurosci, 6*(6), 463-475.

Dunlop, B. W., & Nemeroff, C. B. (2007). The role of dopamine in the pathophysiology of depression. *Arch Gen Psychiatry, 64*(3), 327-337. doi:10.1001/archpsyc.64.3.327

Farmer, A. E. (1996). The genetics of depressive disorders. *International Review of Psychiatry, 8*(4), 369-372. doi:10.3109/09540269609051551

Gershon, E. S., Alliey-Rodriguez, N., & Liu, C. (2011). After GWAS: searching for genetic risk for schizophrenia and bipolar disorder. *Am J Psychiatry, 168*(3), 253-256. doi:10.1176/appi.ajp.2010.10091340

Gladstone, G. L., Parker, G. B., Mitchell, P. B., Malhi, G. S., Wilhelm, K., & Austin, M. P. (2004). Implications of childhood trauma for depressed women: an analysis of pathways from childhood sexual abuse to deliberate self-harm and revictimization. *Am J Psychiatry, 161*(8), 1417-1425. doi:10.1176/appi.ajp.161.8.1417

Glessner, J. T., Wang, K., Sleiman, P. M., Zhang, H., Kim, C. E., Flory, J. H., . . . Hakonarson, H. (2010). Duplication of the SLIT3 locus on 5q35.1 predisposes to major depressive disorder. *PLoS One, 5*(12), e15463. doi:10.1371/journal.pone.0015463

Goldman, N., Glei, D. A., Lin, Y. H., & Weinstein, M. (2010). The serotonin transporter polymorphism (5-HTTLPR): allelic variation and links with depressive symptoms. *Depress Anxiety, 27*(3), 260-269. doi:10.1002/da.20660

Gu, L., Xie, J., Long, J., Chen, Q., Chen, Q., Pan, R., . . . Su, L. (2013). Epidemiology of major depressive disorder in mainland china: a systematic review. *PLoS One, 8*(6), e65356. doi:10.1371/journal.pone.0065356

Herwig, R., Hardt, C., Lienhard, M., & Kamburov, A. (2016). Analyzing and interpreting genome data at the network level with ConsensusPathDB. *Nat. Protocols, 11*(10), 1889-1907. doi:10.1038/nprot.2016.117

http://www.nature.com/nprot/journal/v11/n10/abs/nprot.2016.117.html#supplementary-information

Hirschfeld, R. M., Montgomery, S. A., Keller, M. B., Kasper, S., Schatzberg, A. F., Moller, H. J., . . . Bourgeois, M. (2000). Social functioning in depression: a review. *J Clin Psychiatry, 61*(4), 268-275.

Hyde, C. L., Nagle, M. W., Tian, C., Chen, X., Paciga, S. A., Wendland, J. R., . . . Winslow, A. R. (2016). Identification of 15 genetic loci associated with risk of major depression in individuals of European descent. *Nat Genet, 48*(9), 1031-1036. doi:10.1038/ng.3623

International HapMap, C. (2005). A haplotype map of the human genome. *Nature, 437*(7063), 1299-1320. doi:10.1038/nature04226

International Human Genome Sequencing, C. (2004). Finishing the euchromatic sequence of the human genome. *Nature, 431*(7011), 931-945. doi:10.1038/nature03001

Kapoor, M., Wang, J. C., Wetherill, L., Le, N., Bertelsen, S., Hinrichs, A. L., . . . Goate, A. (2014). Genome-wide survival analysis of age at onset of alcohol dependence in extended high-risk COGA families. *Drug Alcohol Depend, 142*, 56-62. doi:10.1016/j.drugalcdep.2014.05.023

Kendler, K. S., Gatz, M., Gardner, C. O., & Pedersen, N. L. (2005). Age at onset and familial risk for major depression in a Swedish national twin sample. *Psychol Med, 35*(11), 1573-1579. doi:10.1017/S0033291705005714

Kessler, R. C., & Bromet, E. J. (2013). The epidemiology of depression across cultures. *Annu Rev Public Health, 34*, 119-138. doi:10.1146/annurev-publhealth-031912-114409

Kim, H. N., Kim, B. H., Cho, J., Ryu, S., Shin, H., Sung, J., . . . Kim, H. L. (2015). Pathway analysis of genome-wide association datasets of personality traits. *Genes Brain Behav, 14*(4), 345-356. doi:10.1111/gbb.12212

Klein, J. P., & Moeschberger, M. L. (2003). *Survival analysis : Techniques for censored and truncated data*. New York, NY: Springer-Verlag New York Inc.

Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J. Y., Sackler, R. S., Haynes, C., . . . Hoh, J. (2005). Complement factor H polymorphism in age-related macular degeneration. *Science, 308*(5720), 385-389. doi:10.1126/science.1109557

Lambert, G., Johansson, M., Agren, H., & Friberg, P. (2000). Reduced brain norepinephrine and dopamine release in treatment-refractory depressive illness: evidence in support of the catecholamine hypothesis of mood disorders. *Arch Gen Psychiatry, 57*(8), 787-793.

Lander, E. S. (1996). The new genomics: global views of biology. *Science, 274*(5287), 536-539.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., . . . International Human Genome Sequencing, C. (2001). Initial sequencing and analysis of the human genome. *Nature, 409*(6822), 860-921. doi:10.1038/35057062

Lee, E. T., & Wang, J. W. (2003). *Statistical methods for survival data analysis* (3rd ed.). New York: J. Wiley.

Lesch, K. P., Bengel, D., Heils, A., Sabol, S. Z., Greenberg, B. D., Petri, S., . . . Murphy, D. L. (1996). Association of anxiety-related traits with a polymorphism in the serotonin transporter gene regulatory region. *Science, 274*(5292), 1527-1531.

Levinson, D. F. (2006). The genetics of depression: a review. *Biol Psychiatry, 60*(2), 84-92. doi:10.1016/j.biopsych.2005.08.024

Lunter, G., & Goodson, M. (2011). Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res, 21*(6), 936-939. doi:10.1101/gr.111120.110

Major Depressive Disorder Working Group of the Psychiatric, G. C., Ripke, S., Wray, N. R., Lewis, C. M., Hamilton, S. P., Weissman, M. M., . . . Sullivan, P. F. (2013). A mega-analysis of genome-wide association studies for major depressive disorder. *Mol Psychiatry, 18*(4), 497-511. doi:10.1038/mp.2012.21

Massat, I., Souery, D., Del-Favero, J., Nothen, M., Blackwood, D., Muir, W., . . . Mendlewicz, J. (2005). Association between COMT (Val158Met) functional polymorphism and early onset in patients with major depressive disorder in a European multicenter genetic association study. *Mol Psychiatry, 10*(6), 598-605. doi:10.1038/sj.mp.4001615

McMahon, F. J., Buervenich, S., Charney, D., Lipsky, R., Rush, A. J., Wilson, A. F., . . . Manji, H. (2006). Variation in the gene encoding the serotonin 2A receptor is associated with outcome of antidepressant treatment. *Am J Hum Genet, 78*(5), 804-814. doi:10.1086/503820

Neumeister, A., Nugent, A. C., Waldeck, T., Geraci, M., Schwarz, M., Bonne, O., . . . Drevets, W. C. (2004). Neural and behavioral responses to tryptophan depletion in unmedicated patients with remitted major depressive disorder and controls. *Arch Gen Psychiatry, 61*(8), 765-773. doi:10.1001/archpsyc.61.8.765

Ohayon, M. M. (2007). Epidemiology of depression and its treatment in the general population. *J Psychiatr Res, 41*(3-4), 207-213. doi:10.1016/j.jpsychires.2006.10.006

Otte, C., Gold, S. M., Penninx, B. W., Pariante, C. M., Etkin, A., Fava, M., . . . Schatzberg, A. F. (2016). Major depressive disorder. *Nat Rev Dis Primers, 2*, 16065. doi:10.1038/nrdp.2016.65

Owzar, K., Li, Z., Cox, N., & Jung, S. H. (2012). Power and sample size calculations for SNP association studies with censored time-to-event outcomes. *Genet Epidemiol, 36*(6), 538-548. doi:10.1002/gepi.21645

Persson, I. (2002). *Essays on the Assumption of Proportional Hazards in Cox Regression*. Uppsala: Acta Universitatis Upsaliensis.

Pritchard, J. K., & Cox, N. J. (2002). The allelic architecture of human disease genes: common disease-common variant...or not? *Hum Mol Genet, 11*(20), 2417-2423.

Pruim, R. J., Welch, R. P., Sanna, S., Teslovich, T. M., Chines, P. S., Gliedt, T. P., . . . Willer, C. J. (2010). LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics, 26*(18), 2336-2337. doi:10.1093/bioinformatics/btq419

Racagni, G., & Brunello, N. (1999). Physiology to functionality: the brain and neurotransmitter activity. *Int Clin Psychopharmacol, 14 Suppl 1*, S3-7.

Reimand, J., Arak, T., Adler, P., Kolberg, L., Reisberg, S., Peterson, H., & Vilo, J. (2016). g:Profiler-a web server for functional interpretation of gene lists (2016 update). *Nucleic Acids Res, 44*(W1), W83-89. doi:10.1093/nar/gkw199

Reimand, J., Kull, M., Peterson, H., Hansen, J., & Vilo, J. (2007). g:Profiler--a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res, 35*(Web Server issue), W193-200. doi:10.1093/nar/gkm226

Santarelli, L., Saxe, M., Gross, C., Surget, A., Battaglia, F., Dulawa, S., . . . Hen, R. (2003). Requirement of hippocampal neurogenesis for the behavioral effects of antidepressants. *Science, 301*(5634), 805-809. doi:10.1126/science.1083328

Schildkraut, J. J. (1965). The catecholamine hypothesis of affective disorders: a review of supporting evidence. *Am J Psychiatry, 122*(5), 509-522. doi:10.1176/ajp.122.5.509

Sequeira, A., Mamdani, F., Ernst, C., Vawter, M. P., Bunney, W. E., Lebel, V., . . . Turecki, G. (2009). Global brain gene expression analysis links glutamatergic and GABAergic alterations to suicide and major depression. *PLoS One, 4*(8), e6585. doi:10.1371/journal.pone.0006585

Stahl, S. M. (1998). Basic psychopharmacology of antidepressants, part 1: Antidepressants have seven distinct mechanisms of action. *J Clin Psychiatry, 59 Suppl 4*, 5-14.

Sullivan, P. F., Neale, M. C., & Kendler, K. S. (2000). Genetic epidemiology of major depression: review and meta-analysis. *Am J Psychiatry, 157*(10), 1552-1562. doi:10.1176/appi.ajp.157.10.1552

Swindle, R. W., Jr., Cronkite, R. C., & Moos, R. H. (1998). Risk factors for sustained nonremission of depressive symptoms: a 4-year follow-up. *J Nerv Ment Dis, 186*(8), 462-469.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., . . . Zhu, X. (2001). The sequence of the human genome. *Science, 291*(5507), 1304-1351. doi:10.1126/science.1058040

Weissman, M. M., Bland, R. C., Canino, G. J., Faravelli, C., Greenwald, S., Hwu, H. G., . . . Yeh, E. K. (1996). Cross-national epidemiology of major depression and bipolar disorder. *JAMA, 276*(4), 293-299.

Weissman, M. M., Wickramaratne, P., Nomura, Y., Warner, V., Pilowsky, D., & Verdeli, H. (2006). Offspring of depressed parents: 20 years later. *Am J Psychiatry, 163*(6), 1001-1008. doi:10.1176/ajp.2006.163.6.1001

Zill, P., Baghai, T. C., Zwanzger, P., Schule, C., Eser, D., Rupprecht, R., . . . Ackenheil, M. (2004). SNP and haplotype analysis of a novel tryptophan hydroxylase isoform (TPH2) gene provide evidence for association with major depression. *Mol Psychiatry, 9*(11), 1030-1036. doi:10.1038/sj.mp.4001525

Appendices

I. R script for the Cox Proportional hazard model run on the VCU VIPBG Light cluster

```
library(survival)
library(gtools)
library(dplyr)


data2 <-read.table("/home/projects/CONVERGE/Genos/raw/converge.n11443.dec2014.sample.wrk.txt",sep = "",header = TRUE) # phenotype file from initial analysis

csa <- read.table("/home/hgedik/converge_csa_17Feb2017_n10502.txt",sep = "", header = TRUE) # phenotype file with CSA and PCs columns after all QC steps completed

data3 <- data2[, c("ID_1", "AAO", "AAO.raw", "AAO.noCSA")] # Subsetting first phenotype file for age at onset

data3[, 5] <- rep("NA", 11443) # Adding a column for getting as much AAO from the phenotype file filling the gaps.

data3$V5 <- as.numeric(data3$V5)

# Filling gaps in AAO column from different columns

for (i in 1:11443)
{
  if (length(as.data.frame(unique(t(data3[i, -1])))[, 1]) == 2) {
    data3[i, 5] <- as.numeric(na.omit(unique(t(data3[i, -1]))))[1]
  }
  else {
    data3[i, 5] <- as.numeric(unique(t(data3[i, -1])))[1]
  }
}

data3 <- data3[, -c(2:4)]

colnames(data3)[c(1:2)] <- c("IID", "AAO")

colnames(csa)[1] <- c("IID")

data3 <- data3[data3$IID %in% csa$IID,] # Subsetting before QC phenotype data frame into a data frame without individuals excluded after QC steps

data3 <- merge(data3, csa, by = c("IID"), all.x = T) # Merging data frames by individual ID

data3 <- data3[, c(1, 2, 4:7)] # Excluding extra columns without need

colnames(data3)[c(1:6)] <- c("IID", "AAO", "MDD", "PC1", "PC2", "CSA") # Making sure that column names are exactly the same in the following steps

datacase <- data3[data3$MDD == 1,] # Cases are subsetted in one data frame

datacase<-subset(datacase, AAO>=18)

age <- read.table("/home/hgedik/converge_n10502.mdd_plus_covars_v2.txt", sep = "", header = TRUE) # Another phenotype file with age at interview column of the control group

age <- age[, c(1, 3, 17)] # Subsetting the data frame for only the age, status and individual ID columns

colnames(age)[1:3] <- c("IID", "MDD", "AAO") # Make sure columns' names match

datacon <- data3[data3$MDD == 0,] # Controls are subsetted
```

```
agecon <- age[age$MDD == 0,] # Controls from the phenotype file with age at onset column

mergeCon <- merge(datacon, agecon, by = c("IID"), all.x = T) # Merging two phenotype data frames

mergeCon <- mergeCon[!(is.na(mergeCon$MDD.y) | mergeCon$MDD.y == ""),] # Excluding rows with unidentified case-control status

mergeCon <- mergeCon[, c(1, 3:6, 8)]

colnames(mergeCon)[1:6] <- c("IID", "MDD", "PC1", "PC2", "CSA", "AAO")

datacon <- mergeCon # Final version of the control dataframe

phenodat <- rbind(datacase, datacon) # Combining rows from control and case data frames

setwd("/home/projects/CONVERGE/Genos/postExcl/recodeA/") # Setting the working directory where genotype files are located

file_list <- list.files() # Listing genotype files for looping over to run the Cox regression analysis

file_list <- file_list[-c(1)] # Individual's genotype file (with .raw extension) name as character array

exfile_list <- grep("chr10+", file_list, perl = TRUE, value = TRUE)

file_list2 <- file_list[!file_list %in% exfile_list]

file_list <- c(exfile_list, file_list2)

file_list2<-gsub(".raw", ".bim", file_list) # Genotype file of SNPs with .bim extension as list of characters


if (iseed<=27)
{
  setwd("/home/projects/CONVERGE/Genos/postExcl/recodeA/")

  x6570 <- read.table(paste(file_list[iseed]), sep = "", header = T) # Reading the genotype file into R

  x6570 <- x6570[, -c(1, 3:6)] # Excluding FID and other columns without need for the regression analysis

  x6570wo <- x6570[x6570$IID %in% datacase$IID,]

  x6570wo <- merge(datacase, x6570, by = "IID") # Merging phenotype and genotype data

  x6570 <- x6570[x6570$IID %in% phenodat$IID,]

  x6570 <- merge(phenodat, x6570, by = "IID") # Merging phenotype and genotype data

  x6570$AAO <- as.numeric(x6570$AAO) # Make sure that AAO column is numeric
  x6570wo$AAO <- as.numeric(x6570wo$AAO)

  setwd("/home/projects/CONVERGE/Genos/postExcl/")

  dat <- read.table(paste(file_list2[iseed]), sep = "", header = F)

  dtf <- dat[rep(seq_len(nrow(dat)), each = 2), ]
  dtf <- dtf[, -c(3)]

  colnames(dtf)[c(1:5)] <- c("CHR", "SNP", "POS", "AL1", "AL2")

  # Column number which designate the SNP number of each genotype file and also the number of regression run in the following for loop.

  nrep<-length(names(x6570))
```

```
  #Two empty dataframes for storing the results of regression summary statistics of AS analysis and CC analysis
respectively

  d301c <- data.frame(coef = rep(0, (nrep - 6)),expcoef = rep(0, (nrep - 6)),secoef = rep(0, (nrep - 6)),Zscore = rep(0, (nrep -
6)), pvalue = rep(0, (nrep - 6)))
  d301d <- data.frame(coef = rep(0, (nrep - 6)),expcoef = rep(0, (nrep - 6)),secoef = rep(0, (nrep - 6)),Zscore = rep(0, (nrep -
6)), pvalue = rep(0, (nrep - 6)))

  for (i in 7:nrep) {

    G5  <- coxph(Surv(x6570$AAO, x6570$MDD) ~  x6570[, i] + PC1 + PC2 + CSA, data = x6570)
    d301c[i-6, ] <- c(summary(G5)$coefficients[1, 1],summary(G5)$coefficients[1, 2],summary(G5)$coefficients[1,
3],summary(G5)$coefficients[1, 4],summary(G5)$coefficients[1, 5])

    G5  <- coxph(Surv(x6570wo$AAO, x6570wo$MDD) ~  x6570wo[, i] + PC1 + PC2 + CSA, data = x6570wo)
    d301d[i-6, ] <- c(summary(G5)$coefficients[1, 1],summary(G5)$coefficients[1, 2],summary(G5)$coefficients[1,
3],summary(G5)$coefficients[1, 4],summary(G5)$coefficients[1, 5])
  }

  d301c <- cbind(dtf, d301c)
  d301d <- cbind(dtf, d301d)

  setwd("/home/hgedik/CSA/")
  write.table(d301c,paste("CSA", seq(1:598)[iseed], ".txt", sep = ""),row.names = F,quote = F)

  setwd("/home/hgedik/CSA/woCSA/")
  write.table(d301d,paste("CSA", seq(1:598)[iseed], ".txt", sep = ""),row.names = F,quote = F)

} else
{

  setwd("/home/projects/CONVERGE/Genos/postExcl/recodeA/")

  x6570 <- read.table(paste(file_list[iseed]), sep = "", header = T) # Reading the genotype file into R

  x6570 <- x6570[, -c(1, 3:6)] # Excluding FID and other columns without need for the regression analysis

  x6570wo <- x6570[x6570$IID %in% datacase$IID,]

  x6570wo <- merge(datacase, x6570, by = "IID") # Merging phenotype and genotype data

  x6570 <- x6570[x6570$IID %in% phenodat$IID,]

  x6570 <- merge(phenodat, x6570, by = "IID") # Merging phenotype and genotype data

  x6570$AAO <- as.numeric(x6570$AAO) # Make sure that AAO column is numeric
  x6570wo$AAO <- as.numeric(x6570wo$AAO)

  setwd("/home/projects/CONVERGE/Genos/postExcl/")

  dat <- read.table(paste(file_list2[iseed]), sep = "", header = F)

  dat <- dat[, -c(3)]
  colnames(dat)[c(1:5)] <- c("CHR", "SNP", "POS", "AL1", "AL2")

  # nrep is the column number which designate the SNP number of each genotype file and also the number of regression run
in the following for loop.

  nrep<-length(names(x6570))


  d301c <- data.frame(coef = rep(0, (nrep - 6)),expcoef = rep(0, (nrep - 6)),secoef = rep(0, (nrep - 6)),Zscore = rep(0, (nrep -
6)),pvalue = rep(0, (nrep - 6)))
  d301d <- data.frame(coef = rep(0, (nrep - 6)),expcoef = rep(0, (nrep - 6)),secoef = rep(0, (nrep - 6)),Zscore = rep(0, (nrep -
6)), pvalue = rep(0, (nrep - 6)))
```

```
  for (i in 7:nrep) {
    G5  <-coxph(Surv(x6570$AAO, x6570$MDD) ~  x6570[, i] + PC1 + PC2 + CSA, data = x6570)
    d301c[i - 6, ] <- c(summary(G5)$coefficients[1, 1],summary(G5)$coefficients[1, 2],summary(G5)$coefficients[1,
3],summary(G5)$coefficients[1, 4],summary(G5)$coefficients[1, 5])

    G5  <- coxph(Surv(x6570wo$AAO, x6570wo$MDD) ~  x6570wo[, i] + PC1 + PC2 + CSA, data = x6570wo)
    d301d[i-6, ] <- c(summary(G5)$coefficients[1, 1],summary(G5)$coefficients[1, 2],summary(G5)$coefficients[1,
3],summary(G5)$coefficients[1, 4],summary(G5)$coefficients[1, 5])
  }

  d301c <- cbind(dat, d301c)
  d301d <- cbind(dat, d301d)

  setwd("/home/hgedik/CSA/")
  write.table(d301c,paste("CSA", seq(1:598)[iseed], ".txt", sep = ""),row.names = F,quote = F)

  setwd("/home/hgedik/CSA/woCSA/")
  write.table(d301d,paste("CSA", seq(1:598)[iseed], ".txt", sep = ""),row.names = F,quote = F)

}
```

II. Query gene list (alphabetical order) of FS analysis for g:GOSt gene set enrichment analysis

Table.7 Query gene list for FS analysis

AACSP1	AASDH	ABCA1	ABCB5	ABHD15-AS1	AC003084.2	AC005276.1
AC005394.1	AC005592.2	AC005616.1	AC006041.1	AC006372.4	AC006548.26	AC006548.28
AC006946.12	AC006946.16	AC006946.17	AC007131.1	AC007551.2	AC007682.1	AC009120.4
AC009501.4	AC009961.3	AC010127.3	AC010731.3	AC010731.4	AC017060.1	AC018359.1
AC018359.3	AC018731.3	AC018755.18	AC053503.2	AC064853.2	AC064875.2	AC068196.1
AC073283.4	AC087430.1	AC090505.6	AC091878.1	AC092071.1	AC092384.1	AC092684.1
AC096558.1	AC098784.1	AC104820.2	AC135999.2	ACO2	ACOT9	ACOXL
ACSS3	ADAM23	ADAMTS9-AS2	ADGRE5	ADGRV1	ADK	AF064858.6
AGBL1	AGBL1-AS1	AKAP13	AL021546.6	AL022397.1	AL163953.3	ALLC
ALPK1	AMFR	ANGPT1	ANK2	ANKRD9	ANKRD22	ANKS1A
ANP32B	AP000695.6	AP000696.2	AP001043.1	APRT	AQP6	AQP7
AQP10	ARF4P3	ARHGEF3	ARHGEF10	ARMC7	ARMCX4	ASB3
ASB5	ASCC2	ASCC3	ASH2L	ASIP	ASS1	ASTE1
ATF1P1	ATP2C1	ATP8A2	ATP8B2	ATP10B	ATXN7L1	AXDND1
BASP1	BBX	BICD1	BLOC1S1	BMP2K	BMP6	BMP8B
BMPR1B	BMS1P20	BNIP3L	BPIFB1	BRINP2	C2ORF54	C2ORF83
C5ORF66	C6	C8ORF86	C9ORF72	C9ORF129	C10orf103	C16ORF90
C19ORF44	C19ORF84	CACNB1	CACNB4	CADM1	CADPS	CADPS2
CASC4	CCDC3	CCDC146	CCL17	CCL24	CD19	CD47
CD200R1	CD200R1L	CD276	CD302	CDC42BPA	CDC42EP3	CDH13
CDHR3	CDK13	CDK14	CEBPG	CECR7	CEP83	CETN3
CFAP70	CFAP77	CHCHD3	CHM	CHPT1	CKLF-CMTM1	CLEC4C
CMIP	CNBD1	CNGB1	CNR2	CNTLN	CNTNAP2	COL4A2
COL13A1	COL23A1	COLEC12	COLQ	COQ3	COQ8A	CPA6
CPNE2	CPSF7	CPVL	CRADD	CSMD1	CSMD2	CSRNP3
CTA-747E2.10	CTAGE1	CTB-1I21.2	CTB-13L3.1	CTB-91J4.1	CTC-232P5.1	CTC-338M12.9
CTC-436P18.1	CTC-471J1.2	CTC-512J12.4	CTC-513N18.7	CTC-548K16.5	CTCFL	CTD-2009A10.1
CTD-2021H9.1	CTD-2086L14.1	CTD-2143L24.1	CTD-2357A8.2	CTD-2373J6.1	CTD-2374C24.1	CTD-2516F10.2
CTD-2528L19.3	CTD-2528L19.4	CTD-2528L19.6	CTD-2535I10.1	CTD-2554C21.1	CTD-2560C21.1	CTD-2619J13.3
CTD-3064H18.4	CTD-3222D19.2	CTNNA3	CYFIP1	CYP2AB1P	CYP2B7P	CYSTM1

CYTH1	DACH1	DACH2	DAPK2	DCAF8L2	DCC	DCLK1
DDX11-AS1	DDX53	DEFB110	DGKB	DGKG	DGKI	DIAPH3
DISC1FP1	DISP3	DKK3	DLGAP1	DNAJC5	DNAJC12	DNAJC19P1
DNM3	DNMT1	DOCK3	DOK6	DOPEY2	DPH6-AS1	DSCAML1
DUTP3	DYM	DYTN	EBF1	EDARADD	EEF1A1P22	EEF1B2P3
EFCAB6	EFHD2	EGFEM1P	EHBP1	EHMT1	EIF4EBP2P3	ELMO1
EMCN	ENPP6	ENSG00000143248	ENSG00000182957	ENSG00000206885	ENSG00000239265	ENSG00000271798
EPCAM	EPM2A	EPSTI1	ERC2	ERO1A	ERVMER61-1	ESR1
ESR2	ETS2	ETV5	ETV5-AS1	ETV6	EXOC2	F2RL2
F13A1	FAAP24	FAM43B	FAM66C	FAM69C	FAM78B	FAM90A1
FAM135A	FAM160A2	FAM160B2	FAM174B	FAM181A-AS1	FAM189A1	FAM214A
FAR2	FBLN2	FBXL17	FBXW7	FGF19	FGFR1OP2P1	FHIT
FHL5	FMN2	FMO7P	FOXG1-AS1	FOXP2	FRMD1	FRMD4A
FRMD5	FSTL4	FYN	GABRB1	GABRB3	GALNS	GALNT13
GAPDHP22	GAPDHP55	GAS2	GFRA2	GGA3	GLB1L2	GLB1L3
GLI2	GNAO1	GNPTAB	GNRHR	GOLPH3L	GPC5	GPC5-AS2
GPHN	GPR18	GPR183	GRAMD1B	GRID2	GRIK2	GRIK4
GRM7	GSTM1	GSTM2	GSTM5	GTF2F2P2	GTPBP2	GTPBP3
H2AFY	H2AFZP3	HCG2040054	HCN1	HCRTR2	HDAC4	HDAC9
HERC4	HERPUD2	HHAT	HMGB1P1	HMGB3P4	HMGN2P22	HMGN2P35
HNRNPA1P54	HORMAD1	HS3ST4	HSPB8	HSPE1P19	IGFBP7	IGSF9B
IGSF11	IL1RL1	IL10RA	IL17RA	IL18R1	INPP1	IPCEF1
IQCA1L	IQGAP2	ITGA7	ITM2A	KAZN	KCNA5	KCND2
KCNH1	KCNJ6	KCNK12	KCNMB2	KCNMB2-AS1	KCTD8	KIAA0040
KIAA0232	KIAA0930	KIAA1211	KIAA1462	KIFAP3	KIRREL	KLHL23
KREMEN1	KRTAP21-4P	L3MBTL4	LA16C-306E5.3	LANCL1-AS1	LCOR	LHPP
LIFR	LILRP1	LIM2	LINC00238	LINC00276	LINC00378	LINC00434
LINC00504	LINC00511	LINC00520	LINC00536	LINC00882	LINC00923	LINC01102
LINC01122	LINC01170	LINC01310	LINC01314	LINC01377	LINC01435	LINC01482
LINGO2	LL22NC03-80A10.6	LLGL2	LMCD1	LMCD1-AS1	LMNTD1	LPAL2
LPGAT1	LPIN1	LPIN2	LPP	LRP1	LRP1-AS	LRP1B
LRRC1	LRRIQ4	LRRTM4	LRSAM1	LUZP2	LY75-CD302	LYAR
MAD2L1BP	MAGEC3	MALRD1	MAPK6	MAPKAP1	MARCH7	MBP
MCF2	MCHR2	MECOM	METAZOA_SRP	MGAT5B	MIR585	MIR605
MIR607	MIR623	MIR3681HG	MIR4448	MIR4453	MIR4697HG	MIR4754
MIR5689HG	MIRLET7DHG	MLEC	MOB3B	MPP4	MPPE1	MRC2
MROH2B	MRPL22P1	MRPL48P1	MRPL49P1	MRPS21P8	MSH2	MSRA
MTCYBP45	MTMR3	MYO18B	MYRIP	NA	NALCN	NALCN-AS1

NANS	NAV2	NEK11	NELL1	NFATC2IP	NFIA	NFIA-AS1
NIPSNAP1	NLGN1	NOSTRIN	NOVA1-AS1	NPAS3	NRG2	NRG3
NTM	NTRK3	NUDT18	NUF2	NUP85	NXPH1	OCA2
OLFML2A	OPRM1	OR2AH1P	OR2D2	OR5AR1	OR10Y1P	OR51R1P
OSBP2	OXCT2	PAH	PALM2	PARK2	PARL	PARVA
PAX5	PAXIP1	PAXIP1-AS2	PBX3	PCAT29	PCDH15	PCLO
PCNP	PCSK5	PCYT1B	PCYT1B-AS1	PDE4B	PDE4D	PDE10A
PDIA5	PDZD7	PEPD	PEX14	PFDN1	PGM1	PHBP4
PHOSPHO2	PIK3C2G	PIK3C3	PITPNC1	PLA2G4E	PLA2G4E-AS1	PLA2G4F
PLA2R1	PLEKHA7	PLGRKT	PNP	POLA1	POLR2A	POTEI
PPARGC1A	PPFIA2	PPFIBP1	PRAMEF25	PRB4	PRDM16	PRDX4
PRELID2	PRICKLE2	PRKACB	PRKG1	PRLR	PROS2P	PRPH
PRTFDC1	PSD3	PSMA2P1	PSME4	PTCHD1-AS	PTPRC	PTPRD
PTPRN2	PVT1	QKI	RABEP2	RACGAP1	RAD1P1	RAD51B
RALYL	RANP2	RAP1GAP2	RAPGEF5	RBFOX1	RERGL	RGS6
RGS7	RHPN2	RIT2	RN7SL66P	RN7SL301P	RN7SL344P	RN7SL394P
RN7SL561P	RN7SL756P	RNF152	RNF220	RNF225	RNU4-71P	RNU6-230P
RNU6-235P	RNU6-310P	RNU6-331P	RNU6-481P	RNU6-523P	RNU6-638P	RNU6-651P
RNU6-666P	RNU6-712P	RNU6-743P	RNU6-815P	RNU6-918P	RNU6-988P	RNU6-1104P
RNU6-1250P	RNU7-48P	ROBO2	ROR1	ROR2	RORA	RP1-41C23.1
RP1-97J1.2	RP1-118J21.24	RP1-118J21.25	RP1-164F3.8	RP3-434P1.6	RP3-468K18.7	RP4-536B24.3
RP4-536B24.4	RP4-597J3.1	RP4-725G10.3	RP4-736H5.3	RP4-785G19.5	RP5-837O21.2	RP5-887A10.1
RP5-933B4.1	RP5-937E21.1	RP5-972B16.2	RP5-1069C8.2	RP5-1101C3.1	RP11-1L9.1	RP11-6N13.1
RP11-9L18.2	RP11-9L18.3	RP11-14I17.3	RP11-17P16.1	RP11-17P16.2	RP11-22B23.1	RP11-22B23.2
RP11-22H5.2	RP11-24J19.1	RP11-24P4.1	RP11-30G8.2	RP11-30J20.1	RP11-32D16.1	RP11-42H13.1
RP11-42O15.2	RP11-50B3.2	RP11-57C13.3	RP11-57C13.6	RP11-57G10.8	RP11-61O1.1	RP11-61O1.2
RP11-64B16.5	RP11-72M17.1	RP11-75I2.3	RP11-80H8.4	RP11-80I3.1	RP11-93G5.1	RP11-111D3.2
RP11-114H23.1	RP11-115A15.4	RP11-120M18.2	RP11-121G22.3	RP11-122G18.7	RP11-131H24.5	RP11-132A1.6
RP11-137P24.5	RP11-141M1.3	RP11-152L20.3	RP11-161H23.9	RP11-165E7.1	RP11-166N6.2	RP11-168O22.1
RP11-174O3.3	RP11-180K7.1	RP11-217B7.2	RP11-217L21.1	RP11-234A1.1	RP11-237D3.1	RP11-239C9.1
RP11-243A14.1	RP11-244F12.1	RP11-264B17.2	RP11-264B17.3	RP11-264B17.4	RP11-264B17.5	RP11-265N7.2
RP11-266K4.1	RP11-276E17.2	RP11-278H7.1	RP11-290L1.7	RP11-306G20.1	RP11-307N16.6	RP11-314P15.2
RP11-342M3.5	RP11-347H15.1	RP11-354I13.2	RP11-359E10.1	RP11-364B6.2	RP11-369E15.2	RP11-369E15.4

RP11-385J1.3	RP11-386M24.9	RP11-396O20.2	RP11-397C12.1	RP11-403P17.2	RP11-406O16.1	RP11-410D17.2
RP11-413H22.2	RP11-420K14.8	RP11-420N3.2	RP11-420N3.3	RP11-421P23.1	RP11-429A20.2	RP11-429A20.3
RP11-429A20.4	RP11-436D23.1	RP11-436F21.1	RP11-443C10.1	RP11-443C10.2	RP11-443C10.3	RP11-444A22.1
RP11-446J8.1	RP11-451O13.1	RP11-453E17.3	RP11-453E17.4	RP11-462G2.2	RP11-466A19.6	RP11-466A19.8
RP11-467H10.2	RP11-474D1.3	RP11-474D1.4	RP11-479J7.1	RP11-486E2.1	RP11-486G15.1	RP11-486G15.2
RP11-503E24.2	RP11-513G11.3	RP11-525K10.2	RP11-525K10.3	RP11-535A5.1	RP11-535C21.3	RP11-539E19.2
RP11-543A18.1	RP11-543H23.2	RP11-550A9.1	RP11-552D8.1	RP11-565A3.2	RP11-570L15.2	RP11-589M4.2
RP11-589M4.3	RP11-613M10.9	RP11-622A1.1	RP11-624L4.1	RP11-646I6.5	RP11-650J17.1	RP11-665G4.1
RP11-671P2.1	RP11-696F12.1	RP11-713P17.3	RP11-713P17.5	RP11-728G15.1	RP11-737O24.3	RP11-748L13.7
RP11-753E22.2	RP11-753E22.3	RP11-757O6.4	RP11-777F6.3	RP11-793A3.2	RP11-797H7.6	RP11-813I20.2
RP11-815J21.1	RP11-817I4.1	RP11-817I4.2	RP11-849N15.4	RP11-849N15.5	RP11-855A2.1	RP11-855A2.3
RP11-862P13.1	RP11-863N1.4	RP11-871F6.3	RP11-936I5.1	RP11-1398P2.1	RP13-497K6.1	RPH3A
RPL7AP31	RPL7L1P10	RPL7P13	RPL12P8	RPL17P35	RPL23AP29	RPL30P13
RPL31P52	RPS3AP38	RPS4XP20	RPS4XP23	RPS5	RPS6KA2	RPS6P23
RPS8P4	RPSAP31	RPTOR	RREB1	RSU1	RUNX1	RYR2
RYR3	S1PR2	SATB1-AS1	SCARB1	SCARNA23	SCFD2	SCHIP1
SCN3B	SCN7A	SCN9A	SDHAF2	SDHC	SDK2	SEC14L5
SEC22A	SEL1L	SEMA5A	SEPT7	SEPT7-AS1	SEPT9	SERTAD4
SFXN3	SGCD	SGCG	SGCZ	SGK223	SGO1-AS1	SHANK2
SHB	SIGLEC5	SIPA1L2	SIPA1L3	SIRPB3P	SIRT1	SIX4
SLC2A13	SLC8A1-AS1	SLC10A2	SLC12A8	SLC13A3	SLC14A2	SLC14A2-AS1
SLC15A1	SLC20A2	SLC22A2	SLC22A3	SLC22A10	SLC22A14	SLC22A23
SLC24A2	SLC25A18	SLC25A37	SLC27A6	SLC29A4	SLC35D2	SLC35F1
SLC44A1	SLC52A3	SLIT1	SLIT3	SMAD4	SMAP1	SMCO4
SMIM19	SMKR1	SNHG18	SNRPD3	SNTG1	SOCS4	SORCS1
SORCS2	SOX9-AS1	SPC25	SPNS1	SPPL3	SPTBN1	SRGAP1
ST3GAL1P1	ST6GALNAC3	ST13P19	STAC	STAM2	STAP1	STARD13
STAT6	STEAP1B	STOML3	STUM	STYX	SVEP1	SWAP70
SYNE2	SYNJ2	SYT9	SYT14	TAS2R1	TBC1D5	TBXAS1
TDRD9	TEAD1	TECPR2	TEKT3	TENM2	TENM4	TIAL1
TIGAR	TM4SF19-AS1	TM4SF19-TCTEX1D2	TMCO4	TMEM55B	TMEM94	TMEM132B
TMEM132D	TMEM135	TMEM185A	TMEM261	TNR	TP53I13	TPM3P3
TRIB3	TRIM14	TRIM21	TRMT10C	TRPC5	TRPS1	TSPAN7
TSPAN8	TTC7B	TTC21B	TTC39B	TWSG1	TXNL1	UBA6
UBA6-AS1	UBAC2	UBAC2-AS1	UBE2CP3	UBXN7	UBXN7-AS1	UCKL1

UHRF1BP1	ULK4	UNC5C	VCAN	VN1R9P	VN1R31P	VN1R32P
VPS54	VWDE	WDHD1	WDR7	WDR25	WDR70	WIPI1
WNK2	Y_RNA	YES1	YRDCP3	ZBED3-AS1	ZC2HC1B	ZCCHC24
ZFP30	ZMAT4	ZMYND12	ZNF100	ZNF124	ZNF229	ZNF423
ZNF500	ZNF525	ZNF540	ZNF571	ZNF571-AS1	ZNF573	ZNF584
ZNF607	ZNF626	ZNF670	ZNF670-ZNF695	ZNF733P	ZNF781	ZNF793
ZNF804A	ZNF837	ZNF841	ZSCAN5A			

III. Query gene list (alphabetical order) of CC analysis for g:GOSt gene set enrichment analysis

Table.8 Query gene list for CC analysis

ABCC3	ABCD2	ABCG4	ABHD2	AC003092.1	AC004901.1	AC004941.3
AC005034.2	AC005062.2	AC005387.2	AC005775.2	AC006042.7	AC006372.5	AC006466.5
AC006548.28	AC007128.1	AC007349.7	AC007682.1	AC009403.2	AC010136.2	AC011288.2
AC012074.2	AC012354.6	AC012368.1	AC012501.2	AC013463.2	AC018816.3	AC018890.6
AC020743.2	AC024560.3	AC037459.4	AC069277.2	AC079610.2	AC092684.1	AC092687.3
AC092798.2	AC096579.15	AC096649.3	AC100802.3	AC104306.4	AC104667.3	AC118754.4
AC132008.1	ACA64	ACEA_U3	ACSL3	ACTBP2	ADAMTS16	ADAMTSL1
ADCY9	ADGB	ADGRD1	ADGRE5	ADGRL2	AEBP2	AGAP3
AGBL1	AGFG1	AHCYL2	ALDH1A2	ALK	ALPK2	AMPH
ANGPT2	ANKRD7	ANKS1A	ANXA2P3	AP000282.2	AP000462.3	AP000619.6
AP001347.6	AP002856.4	AP1G1	AQP4-AS1	ARF4P3	ARFIP2	ARL8A
ARL15	ARSB	ASB11	ASPN	ASTN2	ATG13	ATP2B2
ATP5B	ATP6V0A4	ATP6V1C2	ATP8A2P3	ATP8B4	ATP10B	B3GLCT
B3GNTL1	BACE1	BACE1-AS	BAZ2A	BBC3	BCL11B	BEND3P2
BIN3	BMP6	BMX	BNIP3P13	BPI	BRCC3	BTBD11
BTC	C1ORF50	C2CD2L	C2ORF88	C3	C4ORF22	C6ORF229
C7ORF72	C8ORF37-AS1	C8ORF58	C9ORF57	C10ORF90	C10ORF113	C11ORF40
C12ORF40	C16ORF90	C19ORF60	CACNA1C	CACNA1C-IT2	CACNA1G	CACNA1H
CADM1	CARD14	CARMIL1	CBL	CCAR2	CCBE1	CCDC7
CCDC18	CCDC18-AS1	CCDC30	CCDC40	CCDC61	CCDC88C	CCDC105
CCDC146	CCDC153	CCDC171	CCDC190	CCNL1	CCNYL1	CCSER1
CD1D	CD44	CD300LF	CDC5L	CDC20	CDC42SE2	CDCA3
CDH13	CDH18	CDH20	CDHR3	CDK5	CDYL2	CELF2
CENPP	CEP78	CEP128	CEP164P1	CEP250	CERK	CERS4
CETP	CFAP47	CH17-262A2.1	CHCHD3	CHD6	CHD7	CHN2
CLEC4C	CLNK	CLRN3	CMB9-94B1.2	CMC4	CMIP	CNTLN
CNTN4	CNTN5	COBL	COL4A4	CORO2B	CPA6	CPLX2
CPXM2	CRLF1	CSGALNACT1	CSMD1	CSMD3	CSRNP3	CTA-221G9.12
CTA-392E5.1	CTB-22K21.2	CTB-52I2.4	CTB-118P15.2	CTB-158E9.1	CTBP2P8	CTC-360J11.5
CTC-360J11.6	CTC-394G3.2	CTC-432M15.3	CTC-459M5.2	CTC-471J1.8	CTC-512J12.4	CTC-512J12.6
CTC-548K16.5	CTD-2007H18.1	CTD-2010I16.1	CTD-2043I16.1	CTD-2066L21.3	CTD-2147F2.1	CTD-2297D10.1

CTD-2308N23.2	CTD-2354A18.1	CTD-2525I3.2	CTD-2534I21.8	CTD-2619J13.3	CTD-2619J13.13	CTD-3006G17.2
CTD-3020H12.3	CTD-3020H12.4	CTD-3128G10.7	CTNND2	CTSB	CXCL13	CYB5AP4
CYP7B1	DAB1	DACH1	DCC	DCUN1D5	DDX11	DDX19A
DDX39A	DENND1A	DESI1	DIO2-AS1	DLEU1	DLGAP1	DMD
DMXL2	DNAH5	DNAJC3	DOCK1	DOCK8	DOK7	DPH6
DPH6-AS1	DPP6	DPYD	DR1	DRC1	DSCAM	DTNB
DUXA	DYNC2H1	ECM2	EEF2KMT	EEFSEC	EFR3A	EIF2D
EIF4A3	ELF2	ELL2P1	ELMOD1	ELOVL1	EMID1	EMP1
ENOX1	ENPEP	ENPP7P14	EPB41L4A	ERC1	ERC2	ERG
ERMAP	ESR1	ESRRG	EXT1	EZH2	F8	F11-AS1
FA2H	FADS2	FAM19A5	FAM81B	FAM129A	FAM169B	FAM189A1
FANCC	FANCD2P2	FAP	FARP1	FARS2	FASTK	FAT1
FBN1	FBXO21	FCRL2	FDPSP3	FGD2	FGD6	FGFBP1
FGFR1	FGGY	FHIT	FILIP1	FKBP8	FNIP1	FPR1
FREM3	FRMD1	FRMPD4	FSTL4	FSTL5	FTH1P24	FTLP12
FUNDC2	FXYD6	FXYD6-FXYD2	GAA	GAB4	GALNT9	GAS7
GBP1P1	GCH1	GFRA1	GJA8	GLDC	GLDN	GLYR1
GMNN	GNB3	GNG12-AS1	GNGT1	GNL2P1	GNL3L	GPATCH8
GPCPD1	GRAMD1C	GRHL2	GRIK2	GRIP1	GRM7	GRM7-AS3
GSN	GUSBP5	HACD2	HDAC4	HDAC9	HECA	HHIPL1
HINFP	HIST1H4E	HKDC1	HMCN2	HMGA2	HMGB1	HMGN1P11
HNRNPA1P60	HRCT1	HSBP1L1	HTR1DP1	HTRA1	HYI	HYI-AS1
IARS	IFT57	IGFBP7-AS1	IGKC	IGSF9B	IL1RAPL2	IMMP2L
INSR	INTS10	IPPK	IQCJ	IQCJ-SCHIP1	ITGA8	ITPR1
ITPR1-AS1	ITPR3	ITSN2	JARID2	KALRN	KANK1	KATNBL1P6
KB-1562D12.1	KCNA1	KCND3	KCNH5	KCNIP4	KCNJ3	KCNK6
KCNMA1	KCNMA1-AS1	KCNMB2	KCNMB2-AS1	KCNQ5	KIAA0232	KIAA1143
KIAA1211L	KIAA1217	KIAA1328	KIAA1549	KIAA1671	KIAA1755	KIF18B
KIF21A	KLHL1	KLRC2	KLRC3	KRT8P45	KRT18P34	KSR2
LA16C-306E5.1	LA16C-306E5.3	LAMA5	LAMC3	LDB2	LDLRAD3	LDLRAD4
LGALS9DP	LHFPL3	LINC00113	LINC00374	LINC00442	LINC00504	LINC00547
LINC00578	LINC00582	LINC00643	LINC00644	LINC00824	LINC00870	LINC00877
LINC01016	LINC01035	LINC01036	LINC01056	LINC01073	LINC01250	LINC01266
LINC01299	LINC01317	LINC01322	LINC01331	LINC01435	LINC01501	LINC01592
LINC01627	LINC01629	LINC01706	LINC01722	LINC01749	LINC01807	LINC01828
LINC01850	LINC01926	LINC01929	LINC01933	LINC01992	LINC02008	LINC02017
LINC02055	LINC02071	LINC02174	LINC02196	LMCD1-AS1	LMNTD1	LNP1
LPP	LRP1B	LRRC8B	LRRC31	LRRIQ4	LUZP2	LYSMD4
LYST	MAD1L1	MAGI2	MAP1B	MAP3K4	MAPKAP1	MARC2
MARCH1	MBOAT1	MBP	MBTPS1	MCAM	MCC	MCF2L
MCPH1	MED8	METAZOA_SRP	METTL15	MFSD9	MGAT4C	MGC16275
MICU3	MIGA1	MIR181A1HG	MIR769	MIR3651	MIR3941	MIR4308
MIR4475	MIR4479	MIR4525	MIR4662B	MIR4670	MIR4697HG	MIR4754

MIR5190	MIR6734	MIR6756	MIR7114	MIR8062	MITF	MKRN7P
MLC1	MMP1	MMP20	MOB3B	MORF4L1	MPL	MRPL45
MRPS22	MS4A4A	MSI2	MSLN	MTAP	MTCP1	MTCYBP28
MTF2	MTG2	MTHFD2P7	MUC4	MVB12B	MYH14	MYLK
MYLKP1	MYOM2	MYRIP	NALCN-AS1	NDUFA3P2	NDUFA5	NDUFS4
NEBL	NEDD4L	NGLY1	NINJ2	NIPAL1	NIPAL3	NIPBL
NIT2	NKAIN2	NLGN4X	NLRX1	NME1	NME1-NME2	NMNAT1
NNT	NOL8	NOS1	NOS2	NOX4	NPM1P30	NR2C1
NREP	NRG3	NRXN3	NSMF	NSRP1P1	NXPH1	OGN
OMD	OPHN1	OR7A3P	OR7E66P	OR7E129P	OR11K1P	OSGEP
OTOF	P3H1	P3H3	PABPC1P10	PALM2	PAPPA2	PCDH9
PCDH11X	PCDHA1	PCDHA4	PCDHA8	PCED1B	PCED1B-AS1	PCLO
PDE1C	PDE7B	PDE8B	PDLIM2	PDZD3	PELI1	PEMT
PGLYRP1	PHACTR3	PHF2P2	PIF1	PIGA	PIP5K1B	PIR
PKD1L2	PKHD1	PKIB	PKN2	PKN2-AS1	PLCB1	PLCB1-IT1
PLCH1	PLEKHA1	PLEKHA8	PLEKHG6	PLS3-AS1	PLXNC1	PM20D1
PNLIPP1	PNMAL1	POC1B	POLA1	POLR2H	POLR2M	POU5F1P6
PPP6R3	PQLC1	PQLC2L	PRDM11	PRDM16	PREX1	PRKCA
PRKG1	PRKG1-AS1	PROX1	PRPF39	PRR20A	PRR20C	PRR27
PRSS46	PRSS50	PRUNE2	PSAT1	PSD3	PSMB6	PTAR1
PTCHD1-AS	PTGES3P5	PTPRB	PTPRD	PTPRE	PTPRN2	PTPRT
PTTG1IP	PVT1	RAB9A	RAB14	RAB17	RAB37	RAB40B
RAD51B	RAD52	RAPGEF6	RASGEF1B	RASGEF1C	RASGRF2	RASSF5
RBFOX1	RBFOX3	RBM33	RBM41	RBM47	RBMS3	RELN
RERE	RFPL4AP6	RFWD2	RGS7	RHBDL2	RIMS3	RN7SKP165
RN7SKP228	RN7SKP269	RN7SL391P	RN7SL714P	RNA5SP146	RNA5SP309	RNF6
RNF38	RNF138P2	RNF165	RNF214	RNF217	RNF225	RNLS
RNU4-56P	RNU4-86P	RNU5A-2P	RNU6-7	RNU6-8	RNU6-75P	RNU6-125P
RNU6-192P	RNU6-262P	RNU6-300P	RNU6-347P	RNU6-1090P	RNU7-6P	ROBO1
ROBO2	ROGDI	RP1-34B20.4	RP1-35C21.2	RP1-68D18.4	RP1-90G24.10	RP1-91J24.3
RP1-92O14.3	RP1-92O14.6	RP1-228P16.3	RP1-228P16.4	RP1-251M9.3	RP1-292B18.4	RP1-297M16.2
RP3-331H24.5	RP3-369A17.6	RP3-399J4.2	RP3-422G23.3	RP3-428L16.2	RP3-468B3.2	RP4-569D19.5
RP4-612C19.1	RP4-612C19.2	RP4-663N10.2	RP4-669H2.1	RP4-678D15.1	RP4-713B5.2	RP4-717I23.2
RP5-864K19.4	RP5-994D16.9	RP5-994D16.12	RP5-1054A22.4	RP5-1106E3.1	RP11-6O2.2	RP11-8L2.1
RP11-8L8.2	RP11-10K17.6	RP11-19J3.5	RP11-21L1.1	RP11-25L3.1	RP11-25L3.3	RP11-28A22.2
RP11-32K4.1	RP11-33I11.2	RP11-53B2.1	RP11-61G19.1	RP11-62C3.6	RP11-62C3.8	RP11-62C3.10
RP11-63E5.6	RP11-64I5.1	RP11-70F11.8	RP11-74K11.2	RP11-75C9.2	RP11-76E12.1	RP11-78A18.2
RP11-81A22.5	RP11-83M16.6	RP11-84A14.4	RP11-87F15.2	RP11-93K22.6	RP11-97E7.1	RP11-98E6.1
RP11-105N14.2	RP11-110J1.2	RP11-115J16.2	RP11-118H4.1	RP11-135F9.3	RP11-143K11.5	RP11-145E5.5

RP11-150C16.1	RP11-154D6.1	RP11-154D17.1	RP11-158D2.2	RP11-161D15.1	RP11-161D15.2	RP11-171I2.1
RP11-172F10.1	RP11-173A6.3	RP11-175E9.1	RP11-175E9.2	RP11-192P3.4	RP11-196H14.3	RP11-196H14.4
RP11-202H2.1	RP11-203M5.6	RP11-203M5.7	RP11-204C23.1	RP11-208K4.1	RP11-212E8.1	RP11-213G6.2
RP11-215A19.2	RP11-215D10.1	RP11-231C18.3	RP11-235P11.1	RP11-239H6.2	RP11-242C24.2	RP11-242C24.3
RP11-242P2.1	RP11-243M5.5	RP11-252C24.3	RP11-255E6.6	RP11-259A24.1	RP11-263G22.1	RP11-264M12.4
RP11-266O8.1	RP11-277P12.6	RP11-280O1.2	RP11-281H11.1	RP11-283G6.5	RP11-283G6.6	RP11-285G1.15
RP11-305B6.1	RP11-305B6.3	RP11-305P14.1	RP11-315L6.1	RP11-317M11.1	RP11-326A19.3	RP11-331H2.4
RP11-334C17.3	RP11-334C17.5	RP11-334C17.6	RP11-342M1.3	RP11-342M21.2	RP11-355I22.2	RP11-359B12.2
RP11-363G15.2	RP11-366L20.3	RP11-375N15.1	RP11-382E9.1	RP11-384F7.1	RP11-385J1.3	RP11-391L3.4
RP11-391L3.5	RP11-391M7.3	RP11-404O13.4	RP11-405M12.2	RP11-410D17.2	RP11-429A20.3	RP11-430H10.2
RP11-430H10.4	RP11-431D12.1	RP11-433A10.3	RP11-434D9.2	RP11-435F13.2	RP11-436D23.1	RP11-436F21.1
RP11-442J17.3	RP11-445O3.2	RP11-485F13.1	RP11-492I21.1	RP11-497K15.1	RP11-505D17.1	RP11-507B12.2
RP11-510J16.5	RP11-513H8.1	RP11-525K10.3	RP11-526F3.1	RP11-529K1.3	RP11-531H8.1	RP11-531H8.2
RP11-538C21.2	RP11-548M13.1	RP11-550I24.2	RP11-550I24.3	RP11-555M1.3	RP11-558A11.2	RP11-563P16.1
RP11-570K4.1	RP11-586K2.1	RP11-593P24.4	RP11-597G23.1	RP11-638L3.1	RP11-654A16.3	RP11-655H13.2
RP11-657O9.1	RP11-669I1.1	RP11-669N7.2	RP11-672L10.5	RP11-673E1.4	RP11-677O4.5	RP11-689P11.2
RP11-689P11.3	RP11-690J15.1	RP11-691H4.4	RP11-707P17.1	RP11-736K12.1	RP11-764D10.2	RP11-767N15.1
RP11-779P15.2	RP11-817J15.2	RP11-817J15.3	RP11-849N15.4	RP11-901H12.1	RP11-1016B18.1	RP11-1082L8.3
RP11-1094M14.5	RP11-1094M14.8	RP11-1105O14.1	RP11-1365D11.1	RP13-143G15.4	RP13-580F15.2	RPIAP1
RPL7AP8	RPL21P39	RPL26P9	RPL31P13	RPL36AP10	RPL38	RPN1
RPRM	RPS5	RPS26P54	RPUSD4	RSU1	RTN4	RTN4R
RUNX1	RXFP2	SAE1	SAGE4P	SATB1-AS1	SCARNA11	SCFD2
SCG5	SCHLAP1	SCML2	SCNN1A	SDC3	SDK1	SEC14L5
SEH1L	SEMA4D	SERINC2	SETD7	SETDB1	SGCD	SGCZ
SGO1-AS1	SHANK2	SHISA9	SIX4	SLAIN2	SLC1A2	SLC1A6
SLC2A14	SLC4A2	SLC5A4	SLC6A5	SLC6A15	SLC22A2	SLC22A10
SLC24A2	SLC24A3	SLC27A6	SLC28A3	SLC35G4	SLC38A11	SLC44A3
SLC44A3-AS1	SLFN12L	SLIT3	SMAD3	SNORA31	SNORA64	SNORA67
SNORA84	SNX2P2	SNX29	SORBS2	SORCS1	SORCS2	SPAAR
SPAG16	SPATA4	SPATC1L	SPCS3	SPHKAP	SPNS3	SPOCK1
SPOCK3	SRGAP1	SSPN	SSTR1	SSU72P8	ST6GALNAC5	STAC
STARD4-AS1	STAT4	STK35	STK39	STON2	STXBP4	SULT1C2

SUMF1	SUPT6H	SVBP	SVILP1	SYN3	SYNPO2	SYT6
SYT9	SZT2	TAF9BP1	TBC1D19	TBC1D24	TBCA	TCEANC
TCP11	TEC	TEKT5	TENM2	TENM3	TESC-AS1	TESPA1
TEX13A	TFEC	THPO	THSD7A	THSD7B	TIAM1	TIE1
TIMM10B	TM4SF5	TMED5	TMEM59L	TMEM132B	TMEM132D	TMEM132E
TMEM150C	TMEM229B	TMEM268	TMEM269	TMTC2	TNFRSF1A	TNFRSF8
TNFRSF10D	TNFSF9	TNKS	TNPO3	TNS1	TOB2	TOM1L1
TOMM70	TOPORSLP1	TPGS1	TPGS2	TPP2	TPTE2	TRAF2
TRAPPC9	TRHDE	TRIB3	TRIM9	TRIM66	TRIP4	TRMT44
TRPC4	TRPC6	TRPM2	TRPV4	TSHZ2	TSHZ3	TSNAX-DISC1
TSPAN11	TSPAN13	TTN	TTN-AS1	TTYH2	TUSC3	TXNL4A
U3	U8	UBE2E2	UBR1	UHRF1BP1L	UMAD1	UPP1
UQCC1	USH1C	USP5	USP43	UTP18	UTRN	VAT1L
VCAM1	VEGFD	VEPH1	VRK2	VWA8	VWC2	WASF1P1
WBSCR17	WDFY4	WDR7	WDR33	WFDC1	WLS	WTAPP1
WWOX	XRCC6	Y_RNA	YAP1	YBX1P10	ZBTB7C	ZBTB20
ZBTB20-AS1	ZC3H4	ZCWPW2	ZDHHC23	ZEB1-AS1	ZFPM2	ZMIZ1-AS1
ZNF3	ZNF285	ZNF365	ZNF385D	ZNF501	ZNF534	ZNF595
ZNF609	ZNF718	ZNF804A	ZNF836	ZNF837	ZPBP	

IV. Permutation results

Table.9 Permutation results of 4 most significant SNPs with very low MAF (<0.01)

SNP	CHR	p-value (Cox)	Minor allele frequency (MAF) / count	Permutated p -value
rs192512830	X	7.89×10^{-10}	10^{-3} / 11	**2.38×10^{-5}**
rs117319230	4	1.34×10^{-9}	5.71×10^{-4} / 6	**1.56×10^{-5}**
rs148193623	2	4.48×10^{-8}	10^{-3} / 11	**1.34×10^{-5}**
rs6915535	6	2.77×10^{-8}	7.52×10^{-4} / 8	**4.69×10^{-6}**

SNP: Single Nucleotide Polymorphism **MAF**: Minor Allele Frequency, **CHR**: Chromosome

V. Power analysis of GWSA

The assumptions for the power analysis were the additive genetic model, proportional hazards ratio, biallelic variant in Hardy-Weinberg equilibrium. Power was estimated **(Fig. 25)** using survSNP R library: Power Calculations for SNP Studies with Censored Outcomes (Owzar, Li, Cox, & Jung, 2012). The FS analysis had around 10,000 sample with 5,000 cases (0.5, the rate of event observation to all observations) and CC had approximately 5,000 all cases (event rate is ~0.99).

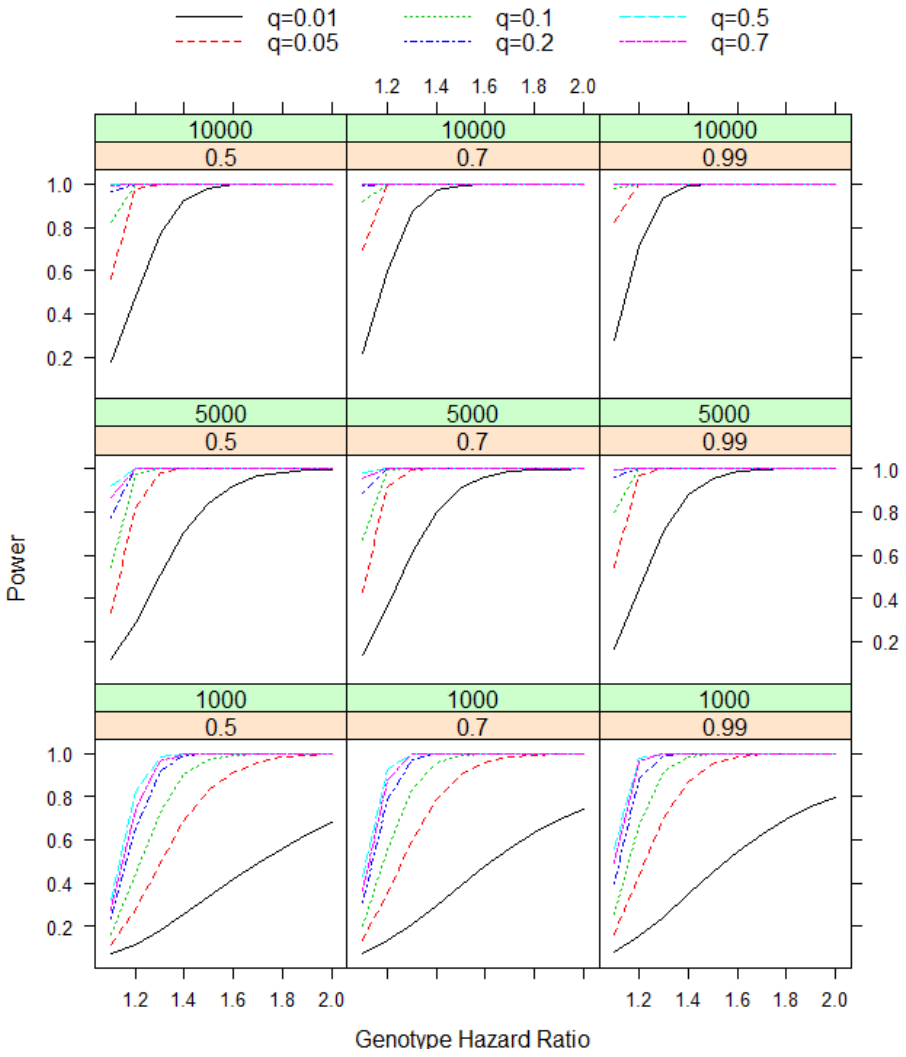
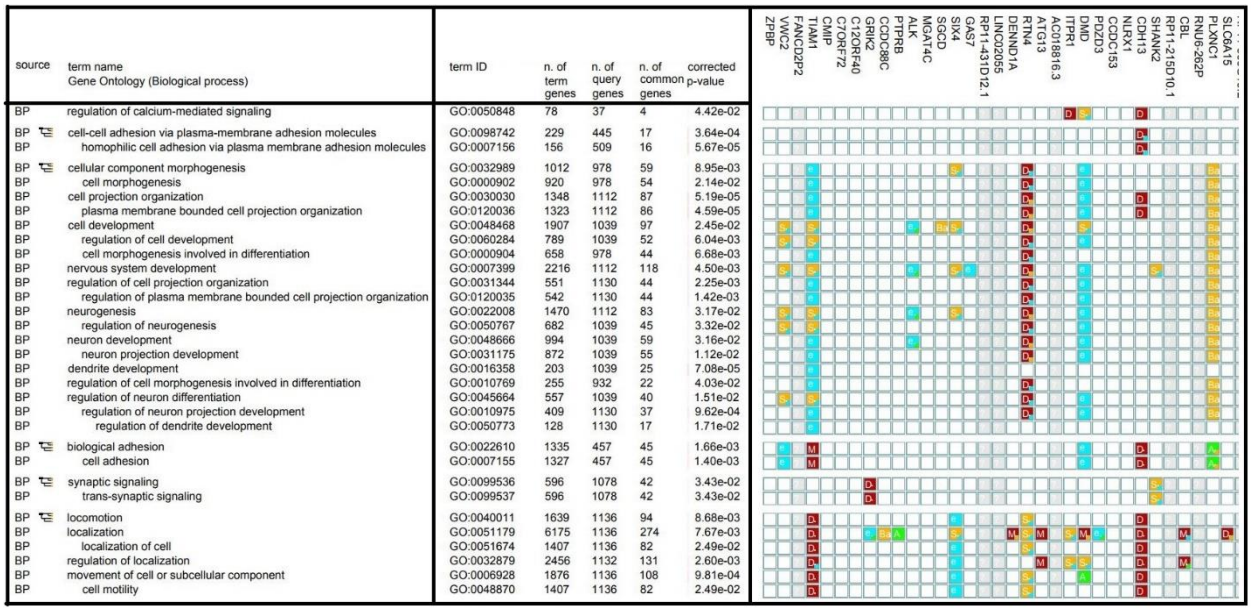


Figure.24 Power analysis. The green strip denotes sample size (1,000, 5,000 and 10,000) and the brown one event rate (0.5, 0.7 and 0.99), q denotes the allele frequency

So, the top left corner and the middle row of third column of the **Fig.25** represent the FS and CC analysis respectively. The power analysis suggests that these two analyses are well-powered (> 0.8) in variants with Genotype hazard ratio > 1.4 and MAF > 0.01. For variants with small hazard ratio, the power of two analyses is steeply reduced below 80% for rarer variants.

VI. Remaining Gene Enrichment (g:GOSt) Results

The first group of figures are the graphical outputs of gene list queries of the FS analysis. All the gene list queries per term are the same within each two Cox PH Analyses. This means each analysis has its own gene list query. These two gene lists can be found in Appendices section at the end of the thesis.



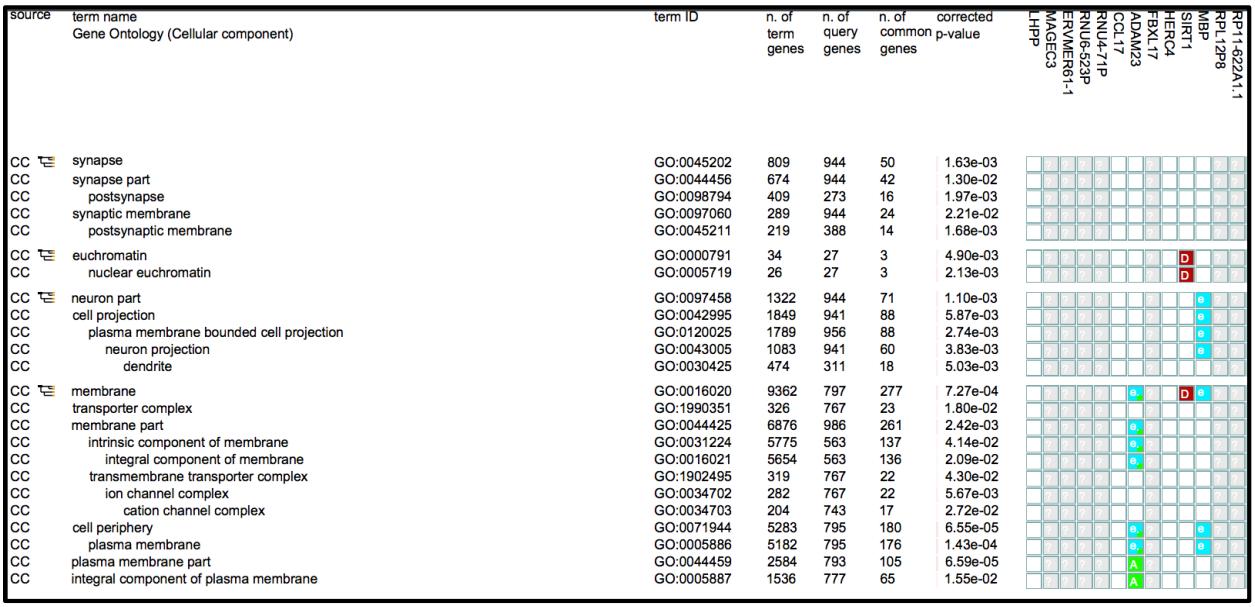
source	term name Gene Ontology (Biological process)	term ID	n. of term genes	n. of query genes	n. of common genes	corrected p-value
BP	regulation of calcium-mediated signaling	GO:0050848	78	37	4	4.42e-02
BP	cell-cell adhesion via plasma-membrane adhesion molecules	GO:0098742	229	445	17	3.64e-04
BP	homophilic cell adhesion via plasma membrane adhesion molecules	GO:0007156	156	509	16	5.67e-05
BP	cellular component morphogenesis	GO:0032989	1012	978	59	8.95e-03
BP	cell morphogenesis	GO:0000902	920	978	54	2.14e-02
BP	cell projection organization	GO:0030030	1348	1112	87	5.19e-05
BP	plasma membrane bounded cell projection organization	GO:0120036	1323	1112	86	4.59e-05
BP	cell development	GO:0048468	1907	1039	97	2.45e-02
BP	regulation of cell development	GO:0060284	789	1039	52	6.04e-03
BP	cell morphogenesis involved in differentiation	GO:0000904	658	978	44	6.68e-03
BP	nervous system development	GO:0007399	2216	1112	118	4.50e-03
BP	regulation of cell projection organization	GO:0031344	551	1130	44	2.25e-03
BP	regulation of plasma membrane bounded cell projection organization	GO:0120035	542	1130	44	1.42e-03
BP	neurogenesis	GO:0022008	1470	1112	83	3.17e-02
BP	regulation of neurogenesis	GO:0050767	682	1039	45	3.32e-02
BP	neuron development	GO:0048666	994	1039	59	3.16e-02
BP	neuron projection development	GO:0031175	872	1039	55	1.12e-02
BP	dendrite development	GO:0016358	203	1039	25	7.08e-05
BP	regulation of cell morphogenesis involved in differentiation	GO:0010769	255	932	22	4.03e-02
BP	regulation of neuron differentiation	GO:0045664	557	1039	40	1.51e-02
BP	regulation of neuron projection development	GO:0010975	409	1130	37	9.62e-04
BP	regulation of dendrite development	GO:0050773	128	1130	17	1.71e-02
BP	biological adhesion	GO:0022610	1335	457	45	1.66e-03
BP	cell adhesion	GO:0007155	1327	457	45	1.40e-03
BP	synaptic signaling	GO:0099536	596	1078	42	3.43e-02
BP	trans-synaptic signaling	GO:0099537	596	1078	42	3.43e-02
BP	locomotion	GO:0040011	1639	1136	94	8.68e-03
BP	localization	GO:0051179	6175	1136	274	7.67e-03
BP	localization of cell	GO:0051674	1407	1136	82	2.49e-02
BP	regulation of localization	GO:0032879	2456	1132	131	2.60e-03
BP	movement of cell or subcellular component	GO:0006928	1876	1136	108	9.81e-04
BP	cell motility	GO:0048870	1407	1136	82	2.49e-02

Figure.25 g:GOSt results for *Biological Process* GO terms in CC analysis.

The *Biological Process* GO term was the only option selected as the gene set enrichment term on the options panel. The far-left of the same figure has the first column as the GO category. The second column designates the term names, which are in hierarchical order. The far right-hand side is the color-coded matrix of gene annotations with column heads as gene names. (**Fig.22** explains each annotation).

The *Biological Process* GO category had enriched GO terms in the query gene list of CC analysis (**Fig.24**). Plasma membrane bounded cell projection organization (GO:0120036, q = 4.59×10^{-5}), dendrite development (GO:0016358, q = 7.08×10^{-5}), and homophilic cell adhesion via

plasma membrane adhesion molecules (GO:0007156, $q = 5.67x10^{-5}$) were the top three significant GO terms found as enriched in the CC query gene list.



source	term name Gene Ontology (Cellular component)	term ID	n. of term genes	n. of query genes	n. of common genes	corrected p-value
CC	synapse	GO:0045202	809	944	50	1.63e-03
CC	synapse part	GO:0044456	674	944	42	1.30e-02
CC	postsynapse	GO:0098794	409	273	16	1.97e-03
CC	synaptic membrane	GO:0097060	289	944	24	2.21e-02
CC	postsynaptic membrane	GO:0045211	219	388	14	1.68e-03
CC	euchromatin	GO:0000791	34	27	3	4.90e-03
CC	nuclear euchromatin	GO:0005719	26	27	3	2.13e-03
CC	neuron part	GO:0097458	1322	944	71	1.10e-03
CC	cell projection	GO:0042995	1849	941	88	5.87e-03
CC	plasma membrane bounded cell projection	GO:0120025	1789	956	88	2.74e-03
CC	neuron projection	GO:0043005	1083	941	60	3.83e-03
CC	dendrite	GO:0030425	474	311	18	5.03e-03
CC	membrane	GO:0016020	9362	797	277	7.27e-04
CC	transporter complex	GO:1990351	326	767	23	1.80e-02
CC	membrane part	GO:0044425	6876	986	261	2.42e-03
CC	intrinsic component of membrane	GO:0031224	5775	563	137	4.14e-02
CC	integral component of membrane	GO:0016021	5654	563	136	2.09e-02
CC	transmembrane transporter complex	GO:1902495	319	767	22	4.30e-02
CC	ion channel complex	GO:0034702	282	767	22	5.67e-03
CC	cation channel complex	GO:0034703	204	743	17	2.72e-02
CC	cell periphery	GO:0071944	5283	795	180	6.55e-05
CC	plasma membrane	GO:0005886	5182	795	176	1.43e-04
CC	plasma membrane part	GO:0044459	2584	793	105	6.59e-05
CC	integral component of plasma membrane	GO:0005887	1536	777	65	1.55e-02

Figure 26 g:GOSt results for *Cellular component* in FS analysis.

In **Fig.26**, cell periphery GO term (GO:71944) is indicated as the most significant ($q = 6.55x10^{-5}$) term among *Cellular component* GO category. The FS analysis GO term query has top significant loci mostly computationally (*in silico*) annotated genes. It also shows that there are significant gene enrichments in synapse (GO:0045202, $q = 1.63x10^{-3}$) and neuron part (GO:0097458, $q = 1.13x10^{-4}$)

The most significant ($q = 5.01x10^{-4}$) GO term is substrate specific transporter activity (GO:00022892) among other terms in the *Molecular Function* GO category as shown in **Fig.27**. As a subgroup of this category, ion transmembrane activity (GO:0015075) has a moderate gene set enrichment ($q = 2.54x10^{-2}$).

source	term name Gene Ontology (Molecular function)	term ID	n. of term genes	n. of query genes	n. of common genes	corrected p-value	LHPP	MAGEC3	ERVMER61-1	RNU6-523P	RNU4-71P	CCL17
MF	metal ion transmembrane transporter activity	GO:0046873	440	508	22	1.16e-02			?			
MF	transporter activity	GO:0005215	1333	526	47	2.52e-03						
MF	substrate-specific transporter activity	GO:0022892	1139	526	44	5.01e-04						
MF	transmembrane transporter activity	GO:0022857	1037	526	40	2.22e-03						
MF	passive transmembrane transporter activity	GO:0022803	469	777	28	4.76e-02						
MF	channel activity	GO:0015267	468	777	28	4.57e-02						
MF	gated channel activity	GO:0022836	331	371	15	2.78e-02						
MF	substrate-specific transmembrane transporter activity	GO:0022891	948	526	39	6.07e-04						
MF	substrate-specific channel activity	GO:0022838	439	777	28	1.35e-02						
MF	ion transmembrane transporter activity	GO:0015075	878	520	35	4.56e-03						
MF	cation transmembrane transporter activity	GO:0008324	667	777	36	2.31e-02						
MF	ion channel activity	GO:0005216	428	777	27	2.54e-02						
MF	cation channel activity	GO:0005261	316	777	22	4.43e-02						

Figure 27 g:GOSt results for *Molecular Function* GO term in FS analysis.

source	term name Biological pathways (KEGG)	term ID	n. of term genes	n. of query genes	n. of common genes	corrected p-value	LHPP	MAGEC3	ERVMER61-1	RNU6-523P	RNU4-71P	CCL17
keg	Glycerophospholipid metabolism	KEGG:00564	95	101	4	4.83e-02		?			?	
keg	Arrhythmogenic right ventricular cardiomyopathy (ARVC)	KEGG:05412	72	457	7	3.29e-02		?			?	
keg	Dilated cardiomyopathy	KEGG:05414	89	251	6	5.00e-02		?			?	
keg	Choline metabolism in cancer	KEGG:05231	101	664	10	2.36e-02		?			?	

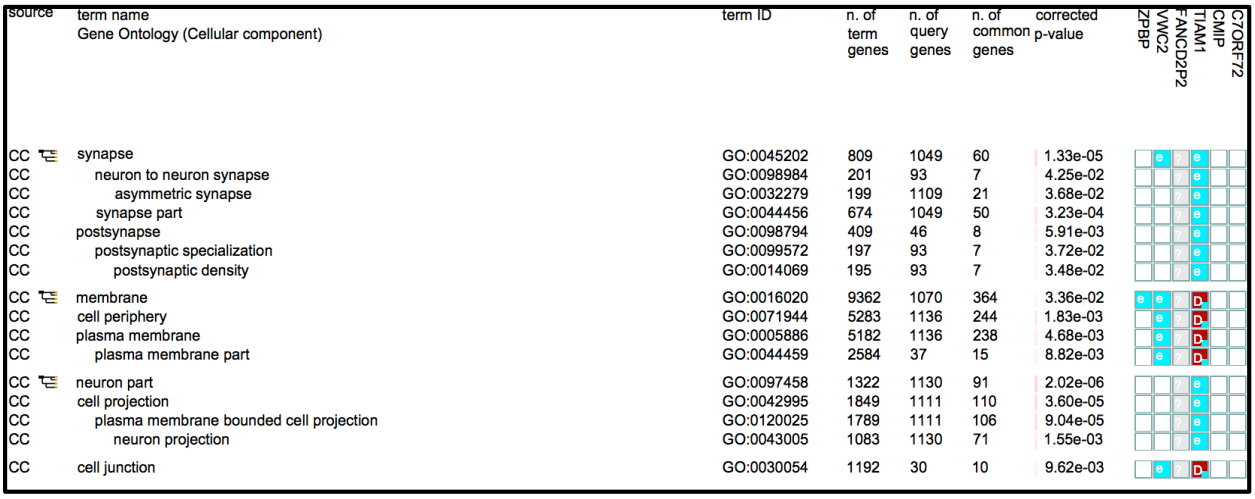
Figure 28 g:GOSt results for KEGG pathways in FS analysis.

The most significant pathway for FS analysis in KEGG pathway database is Choline metabolism in cancer (KEGG:05231, q = 2.36x10^{-2}) as shown in **Fig.28**.

source	term name Biological pathways (Reactome)	term ID	n. of term genes	n. of query genes	n. of common genes	corrected p-value	LHPP	MAGEC3	ERVMER61-1	RNU6-523P	RNU4-71P	CCL17
rea	Netrin mediated repulsion signals	REAC:418886	10	250	3	5.00e-02		?			?	
rea	Axon guidance	REAC:422475	556	513	23	3.55e-02		?			?	
rea	Neuronal System	REAC:112316	359	886	26	4.76e-03		?			?	

Figure.29 g:GOSt results for Reactome pathway in FS analysis.

Neuronal system (REAC:112316, q = 4.76x10^{-3}) was the most significantly enriched Reactome gene set in the query gene list of the FS analysis as shown in **Fig.29**.

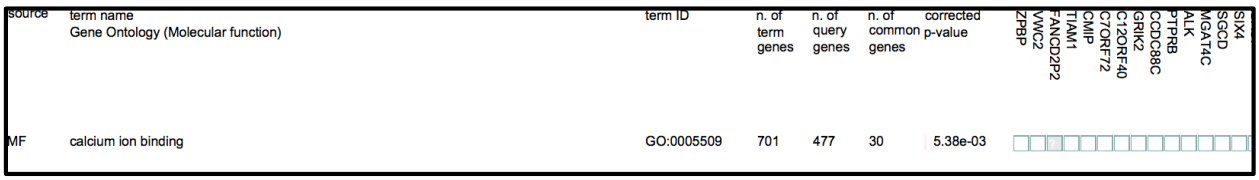


source	term name Gene Ontology (Cellular component)	term ID	n. of term genes	n. of query genes	n. of common genes	corrected p-value
CC	synapse	GO:0045202	809	1049	60	1.33e-05
CC	neuron to neuron synapse	GO:0098984	201	93	7	4.25e-02
CC	asymmetric synapse	GO:0032279	199	1109	21	3.68e-02
CC	synapse part	GO:0044456	674	1049	50	3.23e-04
CC	postsynapse	GO:0098794	409	46	8	5.91e-03
CC	postsynaptic specialization	GO:0099572	197	93	7	3.72e-02
CC	postsynaptic density	GO:0014069	195	93	7	3.48e-02
CC	membrane	GO:0016020	9362	1070	364	3.36e-02
CC	cell periphery	GO:0071944	5283	1136	244	1.83e-03
CC	plasma membrane	GO:0005886	5182	1136	238	4.68e-03
CC	plasma membrane part	GO:0044459	2584	37	15	8.82e-03
CC	neuron part	GO:0097458	1322	1130	91	2.02e-06
CC	cell projection	GO:0042995	1849	1111	110	3.60e-05
CC	plasma membrane bounded cell projection	GO:0120025	1789	1111	106	9.04e-05
CC	neuron projection	GO:0043005	1083	1130	71	1.55e-03
CC	cell junction	GO:0030054	1192	30	10	9.62e-03

ZPBP, VWC2, FANCD2P2, TIAM1, CMIP, C7ORF72

Figure.30 g:GOSt results for *Cellular Component* GO terms in CC analysis.

Another broad category of Gene Ontology is *Cellular Component* **(Fig.30)**. In this GO category, we found sixteen GO terms are significantly enriched in our query gene list based on results in CC analysis. Top two significant GO terms are Neuron part (GO:0097458, $q = 2.02 \times 10^{-5}$) and Synapse (GO:0045202, $q = 1.33 \times 10^{-5}$).

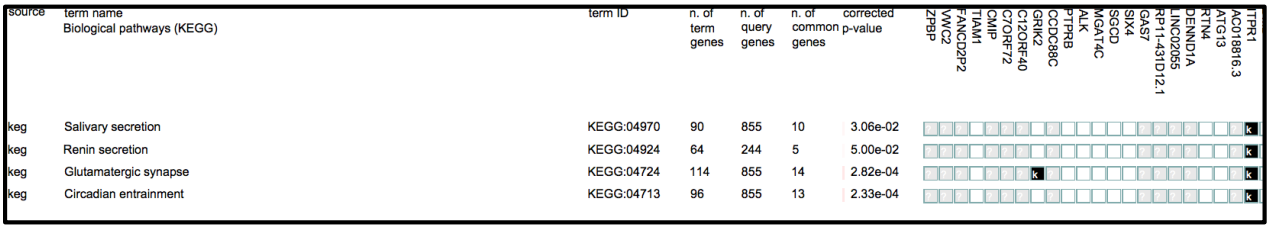


source	term name Gene Ontology (Molecular function)	term ID	n. of term genes	n. of query genes	n. of common genes	corrected p-value
MF	calcium ion binding	GO:0005509	701	477	30	5.38e-03

ZPBP, VWC2, FANCD2P2, TIAM1, CMIP, C7ORF72, C12ORF40, GRIK2, CCDC88C, PTPRB, ALK, MGAT4C, SGCD, SIX4

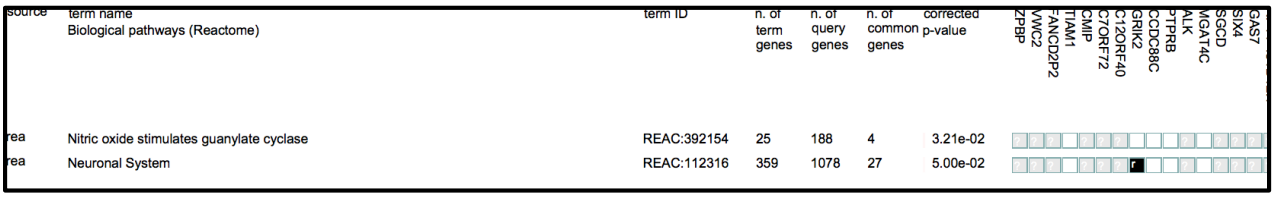
Figure.31 g:GOSt results for *Molecular Function* GO term in CC analysis.

Molecular Function term called calcium ion binding (GO:0005509) shows a significant enrichment ($q = 5.38 \times 10^{-3}$) in the query gene list of CC analysis (**Fig.31**).



source	term name Biological pathways (KEGG)	term ID	n. of term genes	n. of query genes	n. of common genes	corrected p-value
keg	Salivary secretion	KEGG:04970	90	855	10	3.06e-02
keg	Renin secretion	KEGG:04924	64	244	5	5.00e-02
keg	Glutamatergic synapse	KEGG:04724	114	855	14	2.82e-04
keg	Circadian entrainment	KEGG:04713	96	855	13	2.33e-04

Figure.32 g:GOSt results for KEGG pathway in CC analysis.



source	term name Biological pathways (Reactome)	term ID	n. of term genes	n. of query genes	n. of common genes	corrected p-value
rea	Nitric oxide stimulates guanylate cyclase	REAC:392154	25	188	4	3.21e-02
rea	Neuronal System	REAC:112316	359	1078	27	5.00e-02

Figure.33 g:GOSt results for Reactome pathway in CC analysis.

The gene enrichment results for KEGG pathways yield Glutamergic synapse (KEGG:04724, q = 2.82x10^{-4}) and Circadian entrainment (KEGG:04713, q = 2.33x10^{-4}) as statistically significant signals in CC analysis (**Fig.32**).

Reactome pathway terms Nitric oxide stimulates guanylate cyclase (REAC: 392154, q = 5x10^{-2}) and Neuronal system (REAC:112316, q = 3.21x10^{-2}) are the only two nominally significant findings from g:GOSt query of gene list of CC analysis (**Fig.33**).

In all gene set enrichment analysis from the main text, the only reported results were with $q \leq$ 0.05 after Bonferroni correction. Effective domain size for GO is 18,971. Based on this number, g:GOSt accepts P < 6.58x10^{-6} as signal of significant enrichment after Bonferroni's correction. On the other hand, the significance threshold for Benjamini-Hochberg FDR is P = 6.28x10^{-3} for GO gene set enrichment. **Table.10** is the counterpart of summary results table (**Table.5**) which adds FDR q-values for each GO term. FDR adjustments enhances the significance of the signals.

VII. Sensitivity analysis of gene set enrichment results

When interpreting the GO term gene set enrichment results, it should be noted that the queries used the electronic annotations. This means that the results were based on *in silico* gene annotations. If you avoid using electronically annotated genes in gene set enrichment of g:GOSt, this would result in non-significant enrichments in some of the terms. The repeated g:GOSt queries with adding option "No electronic GO annotations" had the following results.

For FS analysis, the only GO terms remain significant are cell periphery (*Cellular Component*, GO:0071944, q =3.84.91x10^{-4}) and substrate specific transporter activity (*Molecular Function*, GO:0022892, q = 1.91x10^{-4}). Also, axon guidance *(Biological Process*, GO:0007411, q = 1.91x10^{-2}) came out as significantly enriched in FS analysis. This is somewhat different than the primary results for FS analysis on gene list query for GO term category Biological Process.

For CC analysis, GO terms remain significant are neuron part (*Cellular Component*, GO:0097456, q = 3.89x10^{-4}), lipid transporter activity (*Molecular Function*, GO:0005319, q = 6.27x10^{-3}) and regulation of neuron differentiation (*Biological Process*, GO:0045664, q = 5.15x10^{-4}).