# Integrative analysis of DNA methylation and gene expression in Schizophrenia.

**Estefanía Irene Eugui Anta**

Master in Bioinformatics and Biostatistics
Master Final Thesis

**Mentor: Helena Brunel Montaner**

**Coordinator: David Merino Arranz**

4 June 2019

# FICHA DEL TRABAJO FINAL

| | |
|---|---|
| **Título del trabajo:** | Integrative analysis of DNA methylation and gene expression in Schizophrenia |
| **Nombre del autor**: | Estefanía Eugui Anta |
| **Nombre del consultor/a:** | Helena Brunel Montaner |
| **Nombre del PRA:** | David Merino Arranz |
| **Fecha de entrega (mm/aaaa):** | 06/2019 |
| **Titulación:** | Máster en Bioinformática y Bioestadística |
| **Área del Trabajo Final:** | Master Final Thesis |
| **Idioma del trabajo:** | Inglés |
| **Palabras clave:** | Schizophrenia, methylation, gene expression |

## Resumen

Los estudios de asociación de genoma completo han identificado varios polimorfismos de un solo nucleótido (SNP) asociados con esquizofrenia (SZ). La evidencia sugiere la existencia de una compleja conexión entre los SNPs y la regulación epigenética de la expresión génica, que no ha sido resuelta hasta el momento. Los análisis integrativos con datos genéticos de estudios multi-ómicos podrían constituir un enfoque importante para dilucidar cómo diferentes SNPs asociados con la SZ afectan al fenotipo de la enfermedad a través de la regulación transcripcional. Para comprobar esta hipótesis, se realizó un análisis cuantitativo de asociación en una cohorte de 10 sujetos con SZ con el objetivo de identificar SNPs y sitios de metilación asociados a la gravedad de la enfermedad. En primer lugar, se identificó un SNP significativo asociado con SZ ($P$-valor ajustado $< 8{\times}10^{-8}$), ubicado en una región no codificante del cromosoma 16 próximo a un gen lncRNA, un tipo de genes conocidos por estar disregulados en la enfermedad. Un análisis integrativo de los datos genéticos y de metilación obtenidos ($P < 0.01$), permitió identificar 341 genes comunes a SNPs y sitios de metilación asociados a SZ. Finalmente, se realizó un análisis integrativo de genes diferencialmente expresados empleando datos de estudio previo de RNA-seq. De los 341 genes totales, 16 fueron identificados como diferencialmente expresados. Notablemente, 3 de ellos (*SHANK2*, *SGK1* y *TCN2*) han sido descritos previamente en la literatura como genes involucrados en SZ. La metodología presentada aquí podría constituir una herramienta novedosa y útil para avanzar en el conocimiento de la fisiopatología de la SZ.



## Abstract

Genome-wide association studies have identified a number of single nucleotide polymorphisms associated with Schizophrenia (SZ). Moreover, increasing body of evidence suggests a complex connection of SNPs and epigenetic regulation of gene expression, which, up to now, is not fully understood. Integrative analyses that use genetic data from multi-omics studies to detect DNA methylation sites associated with gene expression and SZ phenotype might constitute a major approach able to elucidate how SZ-associated SNPs affect the disease traits throughout genetic regulation of transcriptional output. To test this hypothesis we performed an exploratory integrative quantitative association analysis to obtain summary statistics data for SZ severity associated SNPs and methylation sites of 10 drug-naïve SZ cases. We firstly identified a significant SZ-associated SNP (adjusted $P < 8{\times}10^{-8}$), located on a non-coding region of chromosome 16 downstream a lncRNA gene (Long intergenic non-coding RNAs), a type of genes known for being dysregulated in SZ. At a less restrictive significant $P$ ($< 0.01$), we found 341 common genes to significant SNP and CpG SZ-associated sites. We further investigate the association of these genes with a previous SZ-RNA sequencing study containing a set of 200 up and down regulated genes. 16 genes were identified as being differentially expressed in SZ. Remarkably, 3 of them (*SHANK2*, *SGK1* and *TCN2*) had been previously described in the literature

as being involved in SZ. The methodology presented here, might be novel and useful tool to further dilucidate SZ physiopathology.

# Contents

# List of Figures



# List of Tables

# 1 Introduction

## 1.1 Context for this work

### 1.1.1 Schizophrenia: Epidemiology and symptoms

Schizophrenia (SZ) constitutes a complex, debilitating and chronic psychiatric disorder, with a lifetime risk of approximately 0.7% [1], and a worldwide prevalence of about 1% [2]. Moreover, it is among the top ten leading causes of disability by mental and neurological disorders in European countries with the subsequent socioeconomic burden not only for patients but also for the wider community [3]. SZ is characterized by the combined presence of a variety of symptoms, including positive (eg. hallucinations, delusions), negative (eg. anhedonia, blunted affect, emotional withdrawal) and cognitive (eg. deficits in executive function, working memory, poor attention) symptoms, as well as motor and mood symptoms [4]. The ratio and severity of symptoms is greatly variable depending on each individual, so that SZ is a complex and heterogenic disorder [5].

### 1.1.2 Schizophrenia: Aetiology

Two main competiting models have come forward as to explain the aetiology of SZ. The neurodevelopmental models attribute SZ to alterations in the prenatal-to-early adolescent development. This model states that genetic and environmental risk factors during prenatal, perinatal, and early adolescence periods, act as insults altering the natural developmental trajectory of the brain and leading to the onset of the disease during adolescence and young adulthood [6]. Yet, the neurodegenerative model describes SZ as a disease of progressively unfavorable neurodegenerative course [7]. This hypothesis has its origins in the the descriptions given by some psychiatrist in the early 19th century, of SZ as "*dementia praecox*", depicting thus a progressive deteriorating disease with no recover. Although both models are 2 competiting on the etiology and clinical course of this disorder, a third unifying hypothesis has been proposed conceptualizating SZ as a progressive neurodevelopmental disorder [8, 9]

### 1.1.3 Neurochemical models of Schizophrenia

Two of the most influential hypotheses concerning the neurobiology underlying this disorder involve dopamine and glutamate neurotransmiters [10]. The dopamine hypothesis of SZ initially arose from the evidence that the administration of amphetamines and similar compounds that increase extracellular concentrations of dopamine induce psychotic symptoms [11]. This hypothesis claims that hyperactivity of dopamine D2 receptor neurotransmission in subcortical and limbic brain regions contributes to positive symptoms of SZ, whereas negative and cognitive symptoms of the disorder can be attributed to hypofunctionality of dopamine D1 receptor neurotransmission in the prefrontal cortex [12].

The glutamate hypothesis was based on the observation that psychotic symptoms induced by antagonists at the NMDA glutamate receptor like ketamine, closely resemble both the positive and negative symptoms of schizophrenia [13]. Broadly, this theory points to a dysfunction of glutamatergic neurotransmission as responsible for in the etiology of the disease [14]. Although, glutamate hypothesis has become increasingly popular over the last years, there are still some inconsistencies with this model. Therefore, an unifying theory involving both neurotransmitters have been proposed [10]

### 1.1.4 Genes and Environment in Schizophrenia

Nowadays, it is widely accepted that the main risk for SZ is to share genetic variability with an affected person [15]. In fact, monozygotic twin studies have found that the heritability of SZ is around 80% and environmental influence has been estimated as 20% [16]. These findings are consistent with a view of schizophrenia as a complex trait that results from genetic and environmental etiological influences. Basing on the high heritability of SZ, there have been many efforts to discover the causative genetic factors and candidate gene studies have been a main approach. However, the identification of single candidate genes is complicated because of the existence of multiple genes interactions along with environmental influences. As a result of this, studies using a candidate gene approach have been confusing and no genes have been unequivocally associated to SZ [17].

### 1.1.5 GWAS studies and DNA methylation in Schizophrenia

Contrary to single candidate approach, GWAS studies have been able to identify many SZ susceptibility loci [18]. They have also shown that the level of DNA methylation, is partly associated with proximate SNPs [19]. DNA methylation is a major epigenetic mechanism consisting on the covalent union of a methyl group on cytosines followed by guanine residues. This type of methylation is referred to as CpG methylation, and cytosine methylated at the fifth carbon of the pyrimidine ring is called 5-methylcytosine (Figure1 A,B).

DNAm and other epigenetic phenomena are important mechanisms of transcriptional regulation. In particular, DNAm regulates gene expression by recruiting proteins involved in gene repression or by inhibiting the binding of transcription factors to DNA [20]. Epigenetic processes can be modified by environment and have been postulated as links between environmental exposures, genetic risk, and SZ [21]. Increasing body of evidence suggests a complex connection of SNPs and epigenetic regulation of gene expression, which, up to now, is not completely understood. Furthermore, most of the disease-associated variants are located in non-coding regions [22], which causes difficulties in clarifying their biological effects on the disease pathogenesis.

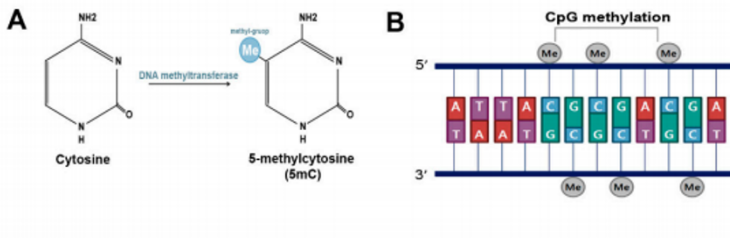


Figure 1: **DNA methylation and demethylation**. **(A)** DNA methylation occurs at the fifth carbon of cytosine and leads to the formation of 5-methylcytosine; **(B)** DNA methylation is predominantly found at CpG sites. Adapted from *Jang et al., 2017.*

### 1.2 Justification for this work

The identification of genes and regulatory elements underlying the associations between genetic susceptibility loci and psychiatric phenotype, might be essential to understand the aetilogy of complex-trait diseases such as SZ [21]. To the best of our knowledge, few studies have investigated how distinct SNPs linked with psychiatric disorders are associated with epigenetic marks with relevance for gene expression [21, 23]. In this context, and given the reduced sample size (n =10) available for this study, this work constitutes an exploratory analysis that might be used in the future with larger sample sizes to provide a novel insight on how genetic variants may contribute to the disorder through epigenetic regulation of gene expression.

## 2 Objectives

### 2.1 General objective

To perform an exploratory integrative quantitative association analysis that includes summary statistics data for SZ severity associated SNPs, methylation sites and mRNA expression levels, to identify the effect of genetic variants on SZ phenotype through epigenetic regulation of transcription.

### 2.2 Specific objectives

The specific objectives of this work are as follows:

1. To perform a SNPs association analysis to identify which genetic loci are associated with SZ severity.
2. To perform a quantitative association analyses to identify which CpG sites are associated with SZ severity.

3. To integrate genetic and methylation data in order to identify common genes in both analyses.

4. To perform an integrative analysis with differentially expressed genes previously described in SZ.

# 3 Approach and follow-up method

Integrative analyses of omics data constitutes a major challenge in Bioinformatics. Omic integration is a complex process requiring thorough study design and extensive data analysis and it is still an expanding field of research. To the best of our knowledge few studies have approach omics integration analyses in SZ. One of the few, it is the work of *Montano et al., 2016*; in which a similar analysis to the one we present here was carried out. For this reason, we considered that the most appropiate strategy to achive the aforementioned objectives should have been based on the work of these authors. Although, they were several differences regarding the data size of both studies, we were able to adapt the analysis workflow to our own dataset during the development of the analysis method (see figure 2) .

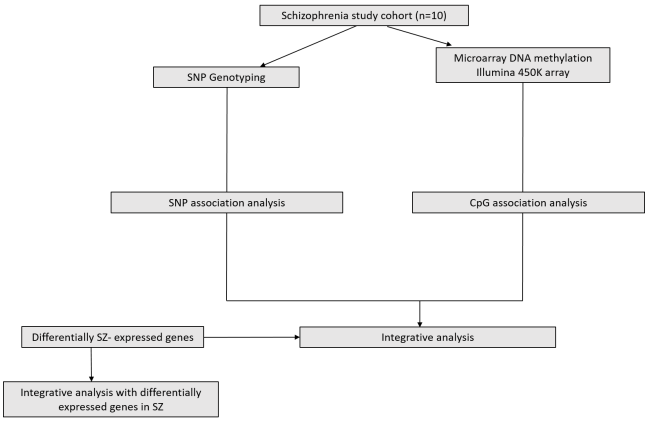


Figure 2: **Analysis workflow**. Adapted from *Montano et al., 2016*.

# 4 Work planning

To accomplish the proposed objectives according to the different deadlines of the project, we proposed a final updated task calendar shown in figure 3. The four tasks corresponded to each one of the objectives of the study. A brief description of the tasks and the duration assigned to each of them are listed below.

- **Task 1:** SNP quantitative association analysis to identify which genetic loci are associated with SZ severity. Assigned duration: 19/03/19 - 03/05/19.
- **Task 2:** Quantitative association analyses to identify which CpG sites are associated with the SZ severity. Assigned duration: 19/03/19 - 03/05/19.
- **Task 3:** Integration of genetic and methylation data in order to identify common genes accross both datasets Assigned duration: 04/05/19 - 13/05/19.
- **Task 4:** Perform an integrative analysis with differentially expressed genes found in SZ. Assigned duration: 14/05/19 - 18/05/19.

Tasks 1 and 2 were developed parallely and were finished prior to task 3. Based on the characteristics of the analyses that tasks 1 and 2 required, we estimated that the time assigned to both of them should be longer in comparison to the rest of the tasks. Both tasks 1 and 2 were finished by the delivery time of PAC2: Phase 1. Task 3 was dependant of tasks 1 and 2 and required a shorter period of time to be completed. Task 4 accounted for the last part of the project and to be completed, the accomplishment of the prior tasks was needed. Tasks 3 and 4 were completed prior to the delivery time of PAC3.

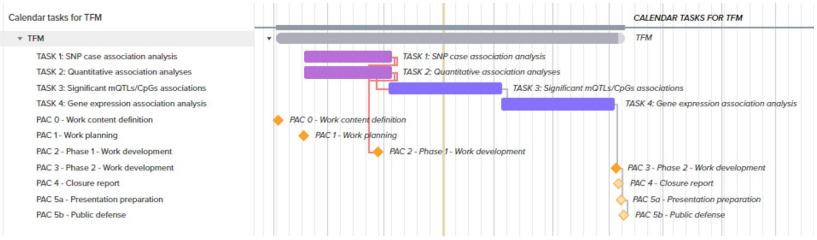


Figure 3: **Tasks calendar**. Figure shows the tasks calendar proposed in order to fulll the deadlines of the project.

# 5 Brief summary of the products obtained

At the end of the present project, we have obtained the following items:

- Phenotype assesment and categorization scale for the study cohort.
- Significant SNP sites associated with SZ severity at different significance values ($P< 0.001$; $P< 0.01$; adjusted-$P= 1\times10^{-7}$).
- Significant DNA methylation sites associated with SZ severity ($P< 0.01$)
- Output for the results of the genetic and methylation data integration analysis at two different significance values ($P< 0.001$; $P< 0.01$)
- Output for the results of the differentially expressed genes integration analysis at two different significance values (textitP$< 0.001$; $P< 0.01$)

# 6 Brief description of the other chapters in the memory

Other chapters in the memory have the following contents:

- **Chapter 7:** *Material and Methods*
  This section contains two main subsections. The first one (*Material*) details data characteristics regarding the sample cohort such as demographic and phenotypic values are detailed. Also, a description of the genotyping, methylation and RNA sequencing data employed for the analysis is presented. The second section (*Methods*) accounts for the statistical analyses performed in order to fullfil the objectives of this study.
- **Chapter 8:** *Results*
  This section contains the main outcomes of the statistical analyses performed in the previous section. More specifically, two main outcomes are listed and described: the results of the integrative analysis of genetic and methylation data and also the output for the integrative analysis with differentially expressed genes.
- **Chapter 9:** *Discussion* This section includes a discussion focused on the 2 main findings of this work. More specifically, the results of the unique associated SZ SNP found to be significant after multiple correction is addressed. Althoug, the main discussion has been based on the dataset of 3 differentially expressed genes found in our analysis.

# 7 Material and Methods

Demographic, clinical and genetic data were kindly provided by Professor Benedicto Crespo-Facorro of University Hospital Marques de Valdecilla (Cantabria, Spain).

## 7.1 Subjects

The cohort analyzed in the present study included drug-naive schizophrenia male patients (n=10) aged between 20 and 43 (mean=30,4; SD= 9,8) , obtained from an ongoing epidemiological and three-year longitudinal intervention program of first-episode psychosis (PAFIP: *Programa Atención Fases Iniciales de Psicosis*) at the outpatient clinic and the inpatient unit at the University *Hospital Marques de Valdecilla* (Cantabria, Spain). The study procedures were approved by the medical faculty ethical committee, and written informed consent was obtained from all study participants.

## 7.2 Phenotype assesment and categorization

Subjects were evaluated at baseline by clinical examiners to confirm the diagnosis of schizophrenia, according to the DSM-IV and the ICD-10 criteria (no

data available). They were also evaluated with the BPRS. This scale [24], first published in 1962 is based on a 18-items-questionnaire that measures psychotic symptoms on SZ patients rated on a seven-point scale (1, not present; 2, very mild; 3, mild; 4, moderate; 5, moderately severe; 6, severe; 7, extremely severe). Accordingly, possible BPRS total scores range from 18 to 126. Although, psychometric properties of BPRS scale in terms of reliability, validity and sensitivity have been extensively examined [25], the clinical meaning of its total score and cut-off values used to define the severity of the disease remains unclear [26]. Consequently, for the purpouse of this work we assessed SZ severity cohort by rating BPRS scores as reported in the work of *Leucht et al.* [26]. The authors categorized BPRS scores according to CGI scale [27]. This scale is to some extent more informative than BPRS since it describes a patient's overall clinical state as a global impression by the rater. Overall,the cohort SZ severity was assessed based on the following linking of CGI score and BPRS total score at baseline (see Tables 1 and 7.2 ):

- **Mildly ill** on the CGI (CGI score 3) approximately corresponds to a BPRS total score of 32 at baseline.
- **Moderately ill** on the CGI (CGI score 4) corresponded to a BPRS total score of 44 at baseline.
- **Markedly ill** (CGI score 5) corresponded to a BPRS total score of 55 at baseline.
- **Severely ill** (CGI score 6) corresponded to a BPRS total score of 70 at baseline.
- **Extremely ill** (CGI score 7) corresponded to a BPRS total score of 85 at baseline.

| **Study cohort (n = 10)** | | | |
|---|---|---|---|
| | Age | BPRS score | CGI score |
| Mean | 30, 4 ± 9.8 | 62,1 ± 19,5 | 4,8 ± 1,3 |
| Maximun value | 43,6 | 100 | 7 |
| Minimum value | 20,3 | 40 | 3 |

Table 1: **Demographic and clinical features of study cohort.** Mean ± standard deviation and maximum and minimum values are shown for each variable.

| Subject | BPRS score | CGI score | Severity |
|---|---|---|---|
| 1 | 46 | 4 | Moderate ill |
| 2 | 40 | 3 | Mildly ill |
| 3 | 58 | 5 | Markedly ill |
| 4 | 77 | 6 | Severely ill |
| 5 | 66 | 5 | Markedly ill |
| 6 | 82 | 6 | Severely ill |
| 7 | 62 | 5 | Markedly ill |
| 8 | 42 | 3 | Mildly ill |
| 9 | 48 | 4 | Moderate ill |
| 10 | 100 | 7 | Extremely ill |

Table 2: **Categorization of SZ phenotype.** Table shows baseline BPRS scores for each cohort individial. CGI corresponding scores along with SZ severity phenotype are also shown.

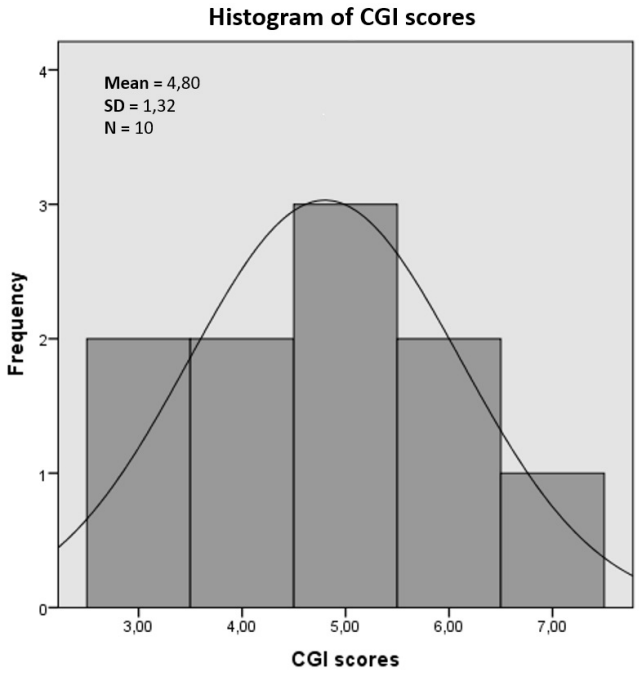


Figure 4: **Histogram of CGI scores and overlaid normal distribution curve.** As observed, phenotype status assessed by CGI scores deviates from normality.

## 7.3 Genotyping data

All study participants (n=10) were genotyped using a customised version of the Illumina Infinium® PsychArray-24 v1.2 BeadChip arrray developed by Illumina in collaboration with the Psychiatric Genomics Consortium [28]. Content for array included a extended set of proven tag single nucleotide polymorphisms (SNPs) associated with SZ disorder and other common psychiatric conditions. Genotyping information data were accessed in PLINK File-Format. PLINK is an open-source, free toolset, widely used for genome association analysis designed by Shaun Purcell [29]. It is considered as standard input format for genotyping array data since it allows to perform a wide range of basic, large-scale genetic analyses [30].

## 7.4 Methylation data

DNA methylation levels in peripheral blood were measured in all individuals (n=10) using the Infinium® Human Methylation 450K BeadChip Assay [31]. This technology quantifies methylation levels at specific loci covering over 480,000 CpG sites and targeting 96% of CpG islands in human genome [32]. The Illumina 450K BeadChip includes two distinct probe types, one for detecting 'methylated' (M) intensity and another one for detecting 'unmethylated'(U) intensity at the interrogated CpG site.To date, two methods have been proposed to measure the methylation level. The first one is called $\beta$ value, ranging from 0 to 1, where $\beta$=M/(M+U+ 100)[33]. The second method, referred by some authors as M-value[34], is the log2 ratio of the intensities of methylated probe versus unmethylated probe [35]. However, $\beta$ value method has been widely used to measure the percentage of methylation [34] and this is the method currently recommended by Illumina. [36, 37].

## 7.5 RNA sequencing data

To obtain information about differentially expressed genes in SZ, a comprehensive bibliographic search on Pubmed repository was conducted. The main objective for this search was to obtain data about transcriptome studies on blood tissue samples of SZ patients that could be employed in our analysis. A total of four RNA-seq studies were found (Table 3).

| Sample Size | SZ subjects | Platform | References |
|---|---|---|---|
| 6 (3S, 3C) | Drug naive | Illumina GA | *Xu et al, 2012* [38] |
| 76 (36S, 40C) | Drug naive | Illumina GA | *Sainz et al, 2013* [39] |
| 22 (22S) | Non drug naive | Illumina GA | *Sainz et al,2015*[40] |
| 1189 (529S, 660C) | No information | Illumina TruSeq | *Sanders et al, 2017* [41] |

Table 3: **Summary of RNA-Seq studies in SZ cohorts.** *Schizophrenia (S); healthy control (C). †Illumina Genome Analyzer (GA). Adapted from *Li et al; 2017[42]*.

The study of *Sainz et al, 2015* analysed differentially expressed genes in SZ after treatment with antypschicotics, while the other two (*Xu et al, 2012; Sainz et al, 2015*) were conducted on drug naive SZ subjects. There was no information regarding any antypsychotic prescription for the study of *Sanders et al, 2017*. Given that our study cohort was composed of 10 SZ subjects that had not taken previous antypschotic medication, the first study was discarded. Out of the other three, the ones with larger sample sizes (*Sainz et al, 2013; Sanders et al, 2017*) were selected. The first one analyzed the blood transcriptome of 36 drug naive schizophrenia patients and 40 healthy matched controls by next-generation sequencing. Among the 22.278 genes analyzed, the authors found significant differential expression (adjusted $P < 0.05$) in 200 genes. The second one undertook an RNA seq-based transcriptomic proling study on a sample of 529 schizophrenia cases and 660 controls. A total of 1058 genes were differentially expressed by affection status after Bonferroni adjustment ($P< 2.36\times10^{-6}$). Among these genes, 361 were downregulated and 697 were upregulated in cases compared to controls.

## 7.6 Statistical Analysis

### 7.6.1 SNPs association analysis to identify genetic loci associated with SZ severity.

- **Association Analysis**

For the association analysis of SNPs and SZ phenotype assesed by CGI score, PLINK software (v1. 90b68) implemented in Linux (Ubuntu distribution 16.04.5), was employed. A total of 624,694 genotyped SNPs for each individual were analysed using a quantitative trait association analysis.

- **Annotation of genotyped SNPs**

To identify which genes were associated with genotyped SNPs, different gene annotation strategies were approached. These strategies are briefly described hereunder.

1. Annotation Strategy based on **Biomart** package.

   Biomart package [43] from Bioconductor software suite (release version 3.9) was firstly employed to access Ensembl database in order to obtain gene annotation for the SNP output based on human genome assembly GRCh37 (hg19). However, method did not provide accurate UCSC gene ID.

2. Annotation Strategy based on **Illumina Psycharray-24 kit**.

   To the best of our knowledge, all study participants (n=10) were genotyped using a customised version of Illumina Infinium® PsychArray-24 BeadChip arrray developed by Illumina in collaboration with the Psychiatric Genomics Consortium [28]. Given that no specific information regarding the array version and customization was provided, the next annotation strategy was based on the gene annotation files provided by Illumina Webpage [44] for the following PsychArrays-24 kit versions:

   - Infinium PsychArray A
   - Infinium PsychArray v1.2
   - Infinium PsychArray v1.3

   Althoug this method did provide accurate UCSC gene ID description for most of the genotyped SNPs, some of them were not present in the gene annotation files for any of the PsychArray versions.

3. Annotation Strategy based on **HumanOmniExpress-24 v1.1 BeadChip.**.

   Next annotation strategy was developed using gene annotation files for HumanOmniExpress-24 v1.1 BeadChip Array provided by Illumina [45]. This latter array was selected since it contains a wider number of SNPs than Infinium PsychArray versions. Again, some of the genotyped SNPs were not present in the gene annotation files for this array.

4. Annotation Strategy based on **genomic coordinates provided by PLINK files**.

   Last annotation strategy was based on finding genomic position coordinates for genotyped SNPs. These coordinates were accessed through the information provided in the initial PLINK files, specifically in the .bim extension file. Once the genomic position were annotated, UCSC gene ID were obtained using SNP Nexus online annotation tool [46].

#### 7.6.2 Quantitative association analyses to identify CpG sites associated with the SZ severity.

• **Association Analysis**

Statistical analyses for testing the quantitative association between CpG sites and SZ phenotype were performed using R, Statistical Software (version 1.1.463) [47] and SPSS software (version 24; SPSS, Armonk, NY, USA). For the purpose of this work DNA methylation levels of 485, 512 CpG sites were computed in terms of $\beta$ values. To identify methylated positions associated with SZ severity, a primary regression analysis was performed. Given the assumption of non normality for the phenotype as variable response (see Figure 4), a two-predictor logistic regression model was fitted instead a linear model as initially planned. The association between proportion methylation values (Illumina "Beta" scale) and CGI score at each CpG site was tested adjusting for age as a covariate (see formula 1). Logistic model was run in R, $\beta$ and $P$-values for the model were obtained.

*Logistic regression modeling formula for testing the quantitative association between CpG sites and SZ phenotype:*

$$\log(\text{phenotype}) \sim \beta_0+\beta_1(DNAm)+\beta_2(age) \quad (1)$$

• **Annotation of CpGs**

Gene annotation for CpG sites was performed with Bioconductor package (release version 3.9). UCSC genes IDs and coordinates for the CpG sites were accessed through R library for Illummina Human Methylation 450k Array for genome assembly hg19.

# 8 Results

• **Identification of significant SNPs**

The initial evaluation of statistically significant SNPs was perfomed after adjustment for multiple testing based on Bonferroni correction ($P < 8\times10^{-8}$). Only one SNP, rs9936526, located in a non coding region of chromosome 16, yielded a significant $P(\ 3 \times 10^{-8})$. The statistical significance was later set to less restrictive $P$-values. For that, a non corrected $P$ of 0.001 was applied. This resulted in 258 significant SNPs located on 104 genes. A further significance analysis with a non corrected $P$ of 0.01 yielded 6437 statistically significant SNPs located on 1586 genes.

- **Manhattan plot**

The results from the SNP association analysis were depicted using a Manhattan plot representation (figure 5). As it can be observed, each point represents a genetic variant. The chromosome position for each variant is showed along the X axis. The Y-axis shows the negative log-base-10 of the $P$ for each indivudal SNP measuring the strength of the association between SZ phenotype and each particular SNP. In brief, the Y axis tells how much the SZ phenotype it is associated with a particular variant. The red line shows the threshold for genome-wide significance after adjustment for multiple testing based on Bonferroni correction ($P$-value $8 \times 10^{-8}$), while the blue line corresponds to the suggestive threshold of $P$-value $1 \times 10^{-5}$).

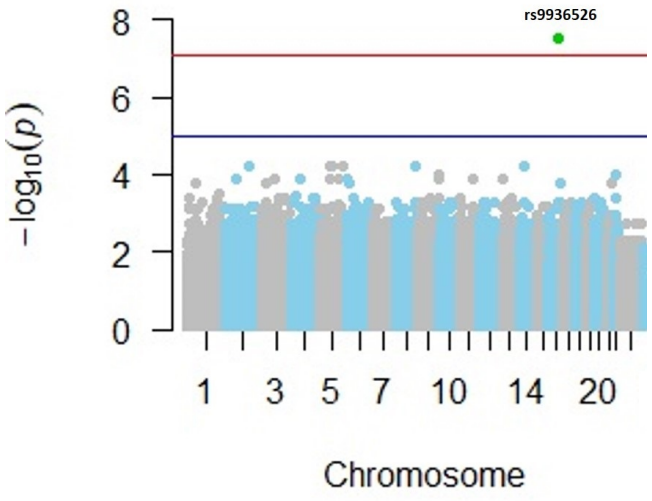


Figure 5: **Manhattan plot for SNPs association analysis.** Figure shows Manhattan plot for the results of the SNP association analysis. Significant SNP: rs9936526 (*adjusted* $P < 8{\times}10^{-8}$) located on chr 16, is highlighted in green.

- **Identification of significant CpG sites**

No stastitically significant CpG sites were found after adjustment for multiple testing based on Bonferroni correction ($P = 1{\times}10^{-7}$). Similarly to the statistical procedure conducted in the SNP statistical analysis, a threshold of 0.001 was then applied. This yielded a total of 499 suggestive CpG sites located on

408 genes. Again a further significance analysis with a non corrected $P$-value of 0.01 yielded a total of 5720 suggestive significant CpG sites on 3791 genes.

- **Manhattan plot**

A Manhattan plot representation depicting the results from the CpGs association analysis can be observed in figure 6. The Y-axis shows the negative log-base-10 of the $P$ for each indivudal CpG measuring the strength of the association between SZ phenotype and each particular methylation site. The red line shows the threshold for genome-wide significance after adjustment for multiple testing based on Bonferroni correction ($P$ $1 \times 10^{-7}$). The blue line corresponding to the suggestive threshold was set on a $P$ of $1 \times 10^{-7}$).

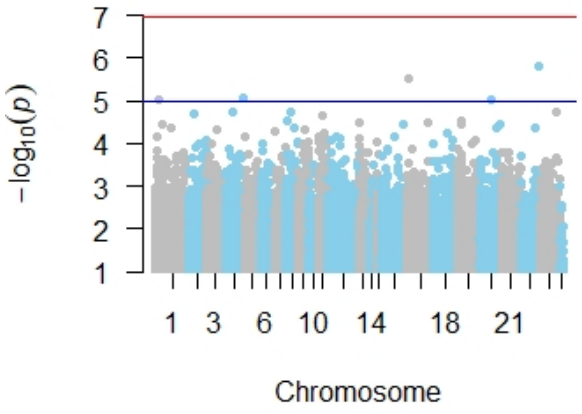


Figure 6: **Manhattan plot for CpGs association analysis.** Figure shows Manhattan plot for the results of the Cpg association analysis. No significant CpG sites were found after Bonferroni correction ($P$-value $1 \times 10^{-7}$).

## 8.1 Integration of genetic and methylation data.

To integrate data from the SNP and CpGs association analyses, statistically significant genes identified on each case were annotated for the aforementioned less restrictive $P$-value ($< 0.001$ and 0.01). Then, a search for commom genes across both gene annotations data sets was performed. Venn diagrams were plotted on each case to identify the number of overlapped genes.

- **Overlapped genes ($P$-value $< 0.001$)**

At a significance level of $P < 0.001$, 4 genes corresponding to RNFT1, CDKL1, NOTCH1 and SPATA13 overlapped for significant CpG and SNP sites, as observed in figure 7.

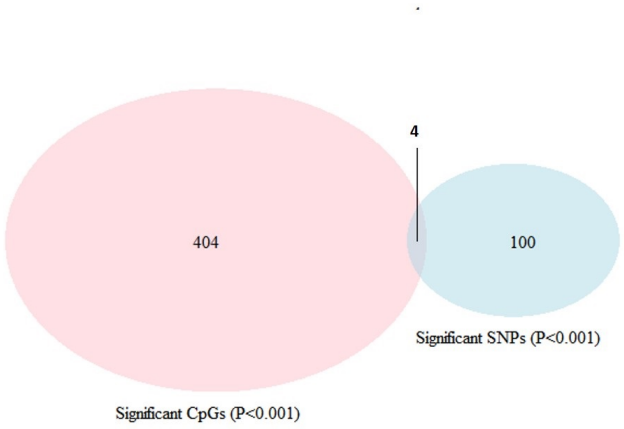


Figure 7: **Venn diagram for significant SNP and CpG sites ($P < 0.001$).** As observed, 4 genes overlap overlapped for significant CpG and SNP sites.

- **Overlapped genes ($P$-value $< 0.01$)**

At a less restrictive significance level of $P < 0.01$, 341 genes overlapped for significant CpG and SNP sites, as observed in figure 8. These common genes included the previous 4 genes found at a significance level of $P$-value $< 0.001$.

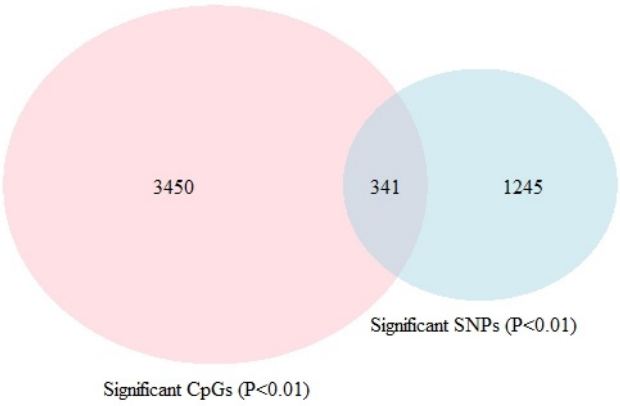


Figure 8: **Venn diagram for significant SNP and CpG sites ($P < 0.01$).** As observed, 341 genes overlap overlapped for significant CpG and SNP sites.

## 8.2 Integrative analysis with differentially expressed genes.

The 341 overlapped genes ($P < 0.01$) found in section 8.1 were used to perform an integrative analysis with differentially expressed genes in SZ from previous studies. This analysis was performed by duplicate, first with the 200 differentially expressed genes ($P$ adjusted $< 0.05$) found in the study of *Sainz et al, 2013* and later with the 1058 differentially expressed genes from the study of *Sanders et al, 2017*.

**• Integrative analysis with differentially expressed genes from *Sainz et al, 2013*.**

The 200 differentally expressed genes found in the study of *Sainz et al, 2013*, were integrate with the 341 genes ($P < 0.01$) found in our study. As it can be observed in figure 9, 5 differentially expressed genes overlapped accross both data sets.

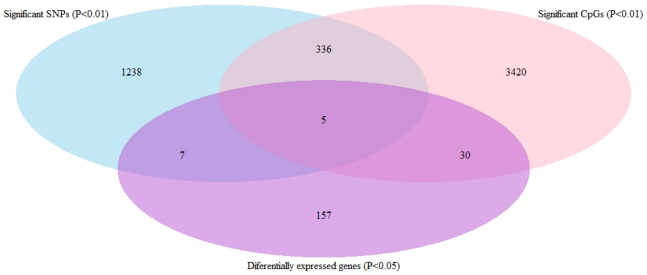


Figure 9: **Venn diagram for significant SNP, CpG sites and differentially expressed genes ($P < 0.01$).** As observed, there is an overlap of 5 differentially expressed genes with the 341 genes associated to SNPs and CpGs found in our study.

For each of the 5 genes found, table 4 shows global $P$ and $\beta$ values computed as the average values for individual SNPs and CpGs associated to each gene. $P$ and fold change values for gene expression extracted directly from *Sainz et al, 2013* are also shown.

| **Gene** | **SNP asssociation** | | **CpG association** | | **Gene expression** | |
|---|---|---|---|---|---|---|
| | $P$ global | $\beta$ global | $P$ global | $\beta$ global | $P$ global | FC |
| *ABCC13* | 0.008 | -1.793 | 0.007 | 2.912 | 0.0002 | 1.389 |
| *CSMD1* | 0.005 | 1.419 | 0.005 | -0.734 | 0.0004 | 1.572 |
| *RIMBP2* | 0.006 | -1.065 | 0.005 | 7.12 | 0.0004 | 2.205 |
| *SGIP1* | 0.005 | 1.642 | 0.002 | 21.099 | 0.0005 | 1.506 |
| *TNS1* | 0.007 | 2 | 0.005 | -0.618 | 0.0004 | 1.264 |

Table 4: **Summary data for the integrative analysis with differentially expressed genes from *Sainz et al, 2013*.** Table shows for each individual differentially expressed gene the global $P$ and $\beta$ values for the SNP and CpG association analysis. $P$ and Fold Change (FC) values for gene expression levels are also shown.

• **Integrative analysis with differentially expressed genes from *Sanders et al, 2017*.**

Figure 9 shows a Manhattan plot representation for the 1058 differentially expressed genes found in the study of *Sanders et al, 2017*. Venn diagram in figure 11 shows the overlap of 16 differentially expressed genes with significant 341 genes found in our analysis.

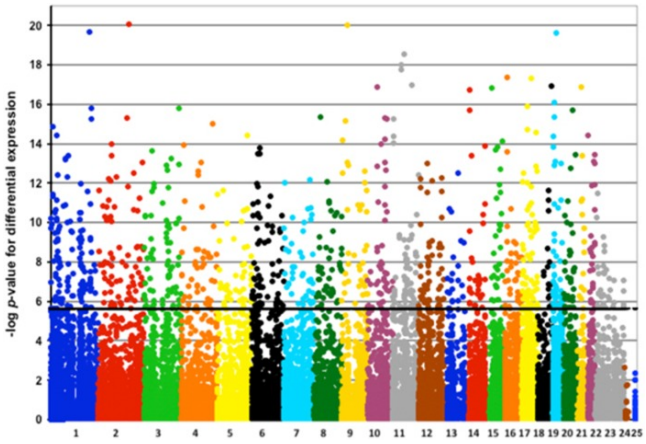


Figure 10: **Manhattan plot of differential expression by schizophrenia status.** The log10 of the $P$ values for the differential expression by SZ status is plotted against the chromosomal location of the analysed genes. The black bar corresponds to Bonferroni $P$ <0.05. Extracted from *Sanders et al, 2017*

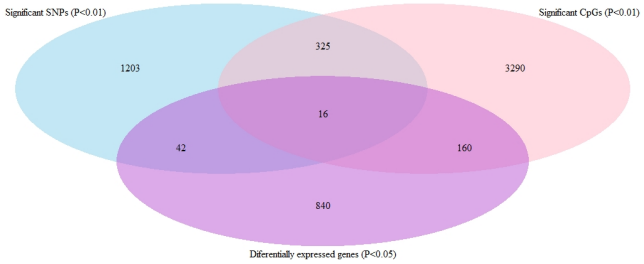


Figure 11: **Venn diagram for significant SNP, CpG sites and differentially expressed genes ($P < 0.01$).** As observed, there is an overlap of 16 differentially expressed genes with the 341 genes associated to SNPs and CpGs found in our study.

Table 5 shows for the 16 differentially expressed genes, global $P$ and $\beta$ values computed as the average values for individual SNPs and CpGs associated to each gene. $P$, $\beta$ and fold change (FC) values for gene expression extracted directly from *Sanders et al, 2017* are also shown.

| Gene | SNP asssociation | | CpG association | | Gene expression | | |
|---|---|---|---|---|---|---|---|
| | $P$ global | $\beta$ global | $P$ global | $\beta$ global | $P$ global | $\beta$ | FC |
| *ANK1* | 0.002 | -1.237 | 0.006 | 7.883 | $4.07 \times 10^{-8}$ | 0.265 | 0.154 |
| *CLNK* | 0,009 | -2.090 | 0.003 | 5.884 | $4.09 \times 10^{-6}$ | 0.190 | 0.122 |
| *COL24A1* | 0,007 | -1.687 | 0.005 | 16.542 | $1.66 \times 10^{-6}$ | -0.121 | -0.060 |
| *GJA3* | 0.005 | 0 | 0.008 | -3.163 | $3.45 \times 10^{-7}$ | 0.098 | 0.161 |
| *GNG7* | 0.005 | 2 | 0.008 | 11.335 | $9.41 \times 10^{-8}$ | -0.213 | -0.057 |
| *IL5RA* | 0.005 | 1.876 | 0.002 | 2.237 | $5.96 \times 10^{-7}$ | -0.058 | -0.212 |
| *IL15* | 0.005 | -2 | 0.0007 | 20.287 | $7.19 \times 10^{-7}$ | 0.149 | 0.053 |
| *KDM2B* | 0.005 | 2.19 | 0.005 | 7.308 | $1.75 \times 10^{-9}$ | 0.212 | 0.032 |
| *NUB1* | 0.007 | 2 | 0.004 | 9.170 | $4.27 \times 10^{-9}$ | 0.336 | 0.041 |
| *RBPMS* | 0.006 | -0.004 | 0.008 | 15.031 | $5.31 \times 10^{-10}$ | 0.452 | 0.210 |
| *SGK1* | 0.005 | 2 | 0.0006 | 3.478 | $7.35 \times 10^{-7}$ | -0.213 | -0.089 |
| *SHANK2* | 0.009 | -2.095 | 0.002 | 26.091 | $7.65 \times 10^{-9}$ | 0.165 | 0.172 |
| *TCN2* | 0.005 | 2 | 0.007 | 14.105 | $8.61 \times 10^{-7}$ | 0.296 | 0.090 |
| *WDFY4* | 0.006 | 2.13 | 0.006 | 3.326 | $1.36 \times 10^{-17}$ | 0.320 | 0.065 |
| *WDR37* | 0.009 | -1.062 | 0.003 | -22.682 | $8.50 \times 10^{-8}$ | -0.047 | -0.020 |
| *ZBTB38* | 0.008 | 1.388 | 0.005 | 2.168 | $1.07 \times 10^{-9}$ | -0.763 | -0.080 |

Table 5: **Summary data from the integrative analysis with differentially expressed genes from *Sanders et al, 2017.*** Table shows for each individual differentially expressed gene the global $P$ and $\beta$ values for the SNP and CpG association analysis. $P$ and Fold Change (FC) values for gene expression levels are also shown.

# 9 Discussion

GWAS have identfied a number of SNPs associated with SZ [18]. Furthermore, increasing body of evidence suggests a complex connection of SNPs and epigenetic regulation of gene expression, which, up to now, is not completely understood. Under these premises, we performed an integrative and exploratory quantitative association analysis that included summary statistics data for SZ severity associated SNPs and methylation sites on a cohort of n= 10 drug-naive SZ subjects. The final purpose was to investigate SNPs and methylation sites associated to SZ severity and to further associate this variants with differentially expressed genes in SZ However, as it has been previous mentioned in section 1.2, this study was considered as merely exploratory and final obtained data must be interpreted only as preliminary results that might be corroborated in the future with larger samples sizes.

The strategy for the analysis was based on previous omic integration studies in SZ [21] and was developed on a 4-step process. Firstly, a SNPs association analysis was performed. This yielded one single significant associated-SZ severity SNP after adjusting for multiple testing. This variant (rs9936526) was located on a non-coding region of chromosome 16 proximate to a lincRNA gene. At a wider significant level of $P < 0.01$, 258 SNPs on 104 genes were identified.

Next, a quantitative association analysis for methylation sites yielded a total of 5720 significant CpG sites on 3791 genes ($P < 0.01$). Later, the number of overlapping genes across both gene data sets was computed resulting in a total of 341 genes. Finally, these genes were used to perform an integrative analysis with differentially expressed genes in SZ from two pusblished studies [39, 41]. Although, 5 and 16 differentially expressed genes were identified for each study respectively, only the 16 genes corresponding to the second study have been selected for further discussion. This choice has been made given the relevance and implication of some of them with previous studies on SZ.

## 9.1 Significantly associated SNP after Bonferroni correction.

The evaluation of statistically significant SNP after adjusting for mutiple testing based on Bonferoni correction ($P < 8{\times}10^{-8}$) yielded only one significant variant: rs9936526 ($P = 3 \times 10^{-8}$) on a non-coding region of chromosome 16:60.604.961. The nearest upstream gene *RP11-354I13.1*, is located at a distance of 47807 bp and corresponds to a lincRNA gene (chr16: 60.486.818 - 60.523.250). LincRNAs (Long intergenic non-coding RNAs) are defined as autonomously transcribed non-coding RNAs longer than 200 nucleotides that do not overlap annotated coding genes. They have an exon-intron-exon structure, similar to protein-coding genes, but do not encompass open-reading frames and do not code for proteins. LincRNAs have been related to a broader lncRNA (long non-coding RNA) family of transcripts, although unlike lincRNAs, many lncRNAs share sequence with coding loci [48]. Many publications, however, do not distinguish between these two sets of transcripts and group them collectively as lncRNAs.

In a similar way to the SNP found in our analysis, GWAS have mapped disease-associated genetic variants to, or in, the vicinity of lincRNA regions [49]. Since molecular functions and mechanisms for lincRNAs are still under debate, it is still not clear how these SNPs may affect the disease. Particularly, some authors have suggested that some lincRNAs represent a novel link between non-coding SNPs and the expression of protein-coding genes, which can be exploited to understand the process of gene-regulation through lincRNAs in more detail [49]. In fact, recent accumulating evidence has revealed that some lncRNAs play a critical role in the regulation of gene expression [50]. Moreover it has been also shown that they participate in the pathogenesis and development of some neurodegenerative diseases such as Alzheimer and Parkinson [51, 52, 53]. Dysfunction of lncRNAs has been also demonstrated to be involved in psychiatric diseases [54]. Particularly, in SZ a growing number of studies show dysregulation of lncRNA in SZ subjects [55]. Other studies have pointed at significant associations of particular lncRNAs to positive SZ symptoms [56] or with SZ early-onset [57]. LncRNA expression profiles have shown differentially expression of specific lncRNAs in SZ subjects compared to healthy controls. Moreover, down-regulation of some lncRNAs was shown to be concurrent with the improvement of symptoms of patients after antipsychotic medication [54],

suggesting that lncRNAs could be considered as novel potential treatment targets.

Overall, no previous association studies were found in literature linking the lincRNA gene *RP11-354I13.1* found in our analysis and SZ. However, only this single tag remained significant after multiple correction. This, suggests the potential usefulness of lncRNA genes to advance in the understanding of specific regulatory pathways for the risk genetic variants to affect SZ.

## 9.2 Differentially expressed associated genes.

The 16 differentially expressed genes found in this study included genes involved in immune response (*IL5*, *ILRA5*, *WDFY4*, *CNK*), genes coding for protein membranes (*ANK1*, *COL24A1*, *WDFY4*), genes with transcriptional activity (*KDMLB*, *RBMPS*, *ZBTB38*), and more remarkably genes previously associated with SZ (*SHANK2* ,*SGK1*, and *TCN2*). These genes have been selected here for further discussion.

### 9.2.1 Genes previously identified in the literature with SZ

***SHANK2* (*SH3 domain and ankyrin repeat containing*)**

SHANK proteins (SHANK1, SHANK2 and SHANK3) are a family of synaptic proteins that function as molecular scaffolds in the postsynaptic density (PSD) of excitatory glutamatergic synapses. By numerous specic protein–protein interactions (figure 12), they are either directly or indirectly linked to other structural proteins, cell adhesion molecules, receptors, ion channels and to actin interacting proteins at the PSD [61].

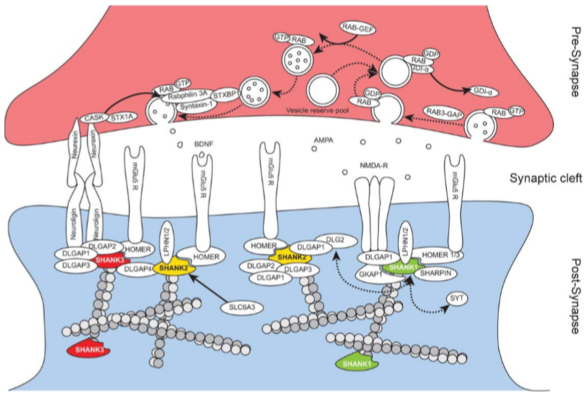


Figure 12: **Interaction network of SHANK proteins at the post-synaptic density of glutamatergic synapses.** Major interaction partners include members of the NMDA receptor complex, members of the metabotropic glutamate receptor complex and actin-associated proteins *Adapted from Guilmatre et al, 2013.*

SHANK proteins seem to have an important role during neurodevelopment [58] and numerous studies have linked a broad spectrum of genetic variations in *SHANK* family members to several neurodevelopmental and neuropsychiatric disorders. Particularly, *SHANK 2* gene mutations have been found in patients with autism sepctrum disorders (ADS) and intelectual disability (ID) [59]. *SHANK2* seems to be upregulated in brains of Alzheimer's disorder cases contrary to *SHANK1* and *SHANK3* that appear downregulated [60]. Moreover, a study sequencing of *SHANK2* in 481 SZ cases and 659 healthy controls identified several non-synonymous variants exclusively in SZ patients [61]. Also, a whole-genome sequencing study in multiplex families with psychoses, reported seven siblings in a family with SCZ spectrum disorders carrying a missense variant in the SHANK2 gene [62]. *SHANK2* mutant mice have shown dysfunction of glutamatergic synapses [63] autistic-like social behaviour [63, 64] and stereotypies ([63]) a common motor symptom observed in SZ patients. According to the study of *Sanders et al SHANK2* seems to be overexpressed in SZ ($\beta = 0{,}165$). However, our study (table 6) identified a highly methylated CpG site (cg0990225 ; $\beta = 26{,}091$) associated with SZ, meaning that this finding is not consistent with the magnitude and direction of the $\beta$ for *SHANK2* expression. Notwithstanding, the high $\beta$ value found for this CpG site might suggest that this methylation mark regulates or favours *SHANK2* expression. This finding would be supported by the idea that *SHANK2* expression is particularly sensitive to DNA methylation pattern suggested in literature [65]. Although, other epigenetic mechanisms such as histone acetylation might expected to regulate the expression of the *SHANK2* gene in an isoform-specific manner [66]. Lastly, 3 different SNPs with a negative effect size ($\beta$ = -2.095) were found in our analysis to be associated with SZ severity (table 6). This would suggest that these variants repress *SHANK2* expression, although to corroborate this hypothesis further and more sophisticated association analyses between SNPs and DNA methylation should be performed in the future.

Table 6: **Association results for *SHANK2***

| Significant SNPs | *P* | $\beta$ | Position |
|---|---|---|---|
| rs4245462 | 0.009 | -2.095 | 70881929 |
| rs4304805 | 0.009 | -2.095 | 70885169 |
| rs4340077 | 0.009 | -2.095 | 70890503 |
| **Significant CpG** | ***P*** | **$\beta$** | **Position** |
| cg09902254 | 0.002 | 26.091 | 70881929 |

Table 6 shows *P* and $\beta$ values for the significant SNP and CpG sites found in our analysis for *SHANK2* gene. Genomic positions for each site are also shown.

Nevertheless, these ndings suggest that *SHANK2* might indeed contribute to the etiology of SZ. Moreover, the fact that alterations of *SHANK2* are common in ASD and SZ might corroborate the hypothesis of the genetic and biological overlap between these two pathologies [61] futher supporting the neurodevelopmental model for SZ. Although, the upregulation found for *SHANK2* in Alzheimer's disorder is also remarkable and might give an insight into the neurodegenerative hypothesis of SZ.

***SGK1* (*serum- and glucocorticoid-inducible-kinase-1*)**

*SGK1* gene encodes a serine/threonine protein kinase that was initially described for its role in the regulation of ion channels in renal cells [67]. However, recent studies have shown the importance of this kinase in the regulation of diverse functions in the brain [68]. Indeed, *SGK1* has been considered to have a key role in long-term memory formation [69] and also in fear conditioning [70]. Moreover, *SGK1* has been also related to the pathophysiology of several neurodegenerative diseases such as Parkinson [71], Alzheimer [70] but also neuropsychiatric disorders like major depresive disorder [72] and SZ [68]. *SGK1* has been even involved in Lafora disease, a severe form of epilepsy among which some of its symptoms are psychosis and dementia [73].

SGK protein is known to upregulate AMPA and kainate receptors and thus they are expected to enhance the excitatory effects of glutamate [74], a neurotransmitter involved in the pathophisiology of SZ [14]. The lack of SGK1 has been proposed to mitigate the glutamate action and at the same time to decrease its clearance from the synaptic cleft [68]. Consistent with this hypothesis, the study of *Sanders et al* found a decrease in *SGK1* expresion in SZ cases ($\beta$ = -0,213). Remarkably, we found a highly methylated CpG site (cg03400131; $\beta$ = 3,478) associated with SZ (table 7), suggesting that this methylation mark might be involved in the inhibition of *SGK1* gene expression. Also, an associated single SNP (rs17063576) with a positive effect size ($\beta$ = 2) was also found in our analysis. This variant migth be associated with downregulation of *SGK1* by promoting DNA methylation of the CpG site.

Overall, the down-regulation of SGK1 in the pathophysiology of SZ might account for an unbalanced SGK1-dependent regulation of AMPA or kainate receptors that would ultimately affect glutamate neurotransmission.

Table 7: **Association results for *SGK1***

| Significant SNPs | *P* | $\beta$ | Position |
|---|---|---|---|
| rs17063576 | 0.005 | 2 | 134578920 |
| **Significant CpG** | ***P*** | $\beta$ | **Position** |
| cg03400131 | 0.0006 | 3.478 | 134497247 |

Table shows *P* and $\beta$ values for the significant SNP and CpG sites found in our analysis for *SGK1* gene. Genomic positions for each site are also shown.

***TCN2* (*Transcobalamin II*)**

*TCN2* encodes a member of the vitamin B12-binding protein family. TCN2 or holotranscobalamin when bound, transports vitamin B12 (cobalamin) to peripheral tissues. Vitamin B12 and other B vitamins like vitamin B6 and folic acid are essential for a correct neuronal function and severe deficiencies of these vitamins have been associated to increased risk for cognitive decline and a variety of neuropsychiatric disorders such as depression, bipolar disorder and SZ [75]. Low blood levels of several B vitamins (also B12) are a relatively consistent finding in SZ and also in drug naive first-episode psychosis patients [76, 77]. Although, B12 vitamin supplementation is sometimes used as an add-on treatment of SZ [75], its admnistration does not always resolve its deficiency. In fact, low levels of B12 are frequently linked to poor absorption and metabolism rather than low consumption, suggesting that B12 intermediates such as transporter TCN2 might be reduced [76]. An epigenetic inhibition of *TCN2* gene expression might account for the decreased levels of TCN2 seen in SZ. Consistent with this hypothesis, we found 3 highly methylated CpG sites ($\beta$ = 17.530; $\beta$ = 12.618; $\beta$ = 12.618) in *TCN2* gene associated with SZ severity (table 8). Two significant associated SNP sites with equal effect sizes ($\beta$ = 2) were also found in *TCN2* gene. This findings might suggest that both variants are associated with downregulation of *TCN2* by promoting DNAm at those CpG sites. These findings are however, contrary to the study of *Sanders et al, 2017*, since these authors described an upregulation of mRNA levels of *TCN2*. This might be explained by the existence of a alternative compensatory mechanism that promotes *TCN2* RNA expression in order to make up for the low blood levels of the transporter.

Table 8: **Association results for *TCN2***

| Significant SNPs | *P* | $\beta$ | Position |
|---|---|---|---|
| rs4820888 | 0.005 | 2 | 31017322 |
| rs5749135 | 0.005 | 2 | 31011906 |
| **Significant CpG** | ***P*** | $\beta$ | **Position** |
| cg00788739 | 0.003 | 17.530 | 31002942 |
| cg17693957 | 0.009 | 12.618 | 31002757 |
| cg22542751 | 0.009 | 12.168 | 31002892 |

Table 8 shows *P* and $\beta$ values for the significant SNP and CpG sites found in our analysis for *TCN2* gene. Genomic positions for each site are also shown.

Overall, although preliminary, our study has been able to identify genes highly described in literature to be associated with SZ. More remarkably, those genes are involved in a wide range of distinct biological processes all of them proposed as affected mechanisms in this pathology. Further analyses might be carry out to confirm our findings.

# 10 Conclusions

SZ constitutes a complex psychiatric disease with multiple aetiological factors involved in its pathology. Omic integration genetic studies might prove as an useful tool to dilucidate the complex genetic architecture of this disease. However, to our knowledge, few studies have investigated so far the conexion between SZ-associated SNPs, methylation marks and gene expression. The methodology presented here, although preliminary has proved to be a novel and useful tool to identify genes previously described in the literature as been associated with SZ.

### 10.0.1 Main Outcomes

In this work, we performed an integrative and exploratory genetic association analysis on a cohort of n= 10 drug-naive SZ subjects. The final purpose was to investigate SNPs and methylation sites associated to SZ severity and to further associate this variants with differentially expressed genes in SZ. For that, a SNP association quantitative analysis was carried out. This yielded a significant variant ($P < 8{\times}10^{-8}$) located on a non-coding region of chromosome 16 proximate to a lincRNA gene. Importantly, specific LncRNA expression profiles have been shown to be differentially expressed in SZ. At a wider significance ($P < 0.01$), 258 SNPs on 104 genes were identified. A quantitative association analysis for methylation sites yielded a total of 5720 significant CpG sites on 3791 genes ($P < 0.01$). We further performed an integrative genetic analysis in order to identify overlapping genes accross both datasets. This computed a total number of 341 matching genes. Finally, an integrative analysis of these genes with with differentially expressed genes in SZ from two previous studies was carried out. From one of this studies, 16 differentially expressed genes were identified. Remarkably, 3 of them: *SHANK2* ,*SGK1*, and *TCN2* had been previously described in literature to be associated to SZ pathology. *SHANK2* gene encodes for a protein with an important role during neurodevelopment. Although, *SHANK2* was upregulated, our study identified a highly methylated CpG site associated with this gene. Thus, probably, other epigenetic mechanisms could be involved in the regulation of *SHANK2* expression. In fact this gene seems to be particular sensitive to DNA methylation pattern as it has been suggested in literature.

*SGK1* gene encodes for a kinase with relevance in the regulation of several functions in the brain [68].*SGK1* expression levels have been shown to be decreased in SZ. Consistently, we found a highly methylated CpG site associated with this gene. Moreover, a single SNP also associated to *SGK1* was identified. This variant showed a positive effect size meaning that it could mediate epigenetic regulation of *SGK1* by promoting DNA methylation of the CpG site and thus, inhibiting the expression of *SGK1*.

Finally, *TCN2* gene encodes for a transporter of the vitamin B12 to peripheal tissues. Low levels of vitamin B12 have been traditionally associated to SZ. Reduced expresion of transporter TCN2 migh account for this deficiency. In line with this hypothesis, We found 3 highly methylated CpG sites in *TCN2* gene associated with SZ. Moreover, two

significant associated SNP sites with equal effect sizes were also associated to *TCN2*. This might suggest that both variants are associated with downregulation of *TCN2* by promoting methylation of those significant CpGs.

### 10.0.2 Reflexion and critical analysis

As already mentioned, given the small size of the sample available, this work was considered initially as a purely exploratory analysis designed to learn useful statistical tools to perform more complex studies in the future on larger sample sizes. Indeed, there were no initial expectations of finding any signicant outcome throughout the analysis process. However, not only this work has served as a comprehensive learning process but also and more remarkably, the methodology presented here has revealed as a useful tool for omic integration analyses. Regarding the process of data analysis, the mentor Helena Brunel provided at all times the information and the proper analysis tools required for an adequate achievement of the proposed objectives. This facilitated the learning process of data analysis and allowed a proper fullment of the deadlines of the project. We also came across to several milestones that were successfully addressed. One was the learning and management of PLINK toolset. As previously mentioned genotyping information was provided in PLINK bedFile-Format, widely considered as standard input format for genotyping array data. In spite of its widespread used, the author of this work had not acquired prior knowledge of this toolset. Therefore, it was required an introductory and basic learning process regarding the dierent format types and management of this tool in order to accomplish the rst proposed objective. The other milestone was the need to perform a prior exploratory analysis regarding the phenotypic characteristics of the subjects consisting of the phenotype assessment and categorization of the sample cohort study.

Regardind the accomplishment of the objectives, it shoudl be mentioned that the initial objective was to perform an integrative case-control association analysis in a study cohort of n = 10 cases and n = 10 controls. However, data were not finally accessible for the control samples and only genotyping and methylation information related to SZ cases (n = 10) were available. Therefore, the general objective was shifted to perform an integrative quantitative association analysis in a cohort of 10 SZ cases. Despite this setback regarding the availability of the data, general and specific objectives were successfully achieved according to the deadlines of the project.

In relation to the metholody employed in the analysis, this was based, as mentioned on a previous omics integration analysis. However, there were several differences regarding the data characteristics of both studies. Firstly, those authors performed a case-control association analysis and secondly, they had a larger sample size. These differences were addressed during the development of the analysis method. For that, we performed an integrative quantitative trait-association analysis as mentioned. Also, given the assumption of non-normality for the phenotype as variable response, a two-predictor logistic regression model for the methylation quantitative analysis was fitted instead a linear model as initially planned.

Finally, this work enables future lines of work such as:

- To perform a methylome quantitative trait locus association analysis in order to test for significant relationships between SNP and CpG loci linked to SZ severity.

- Also, if mRNA data were accesible for SZ cases, an association analysis between mRNA levels and SZ severity could have been also addressed. The outcome of these analyses might offer a clear idea of the genes differentially expressed in the study cohort. These results might be subsequently employed to be correlated with the significant mQTL sites found in our analysis.
- As the initial objective of this work was proposed, a case-control integrative association analysis could also have been performed if data regarding control subjects were available.

Finally, the findings of this study open further and promising lines of research a similar analysis to the one here with a larger sample size could be performed in order to corroborate our results.

# GLOSARY

**AMPAR:** $\alpha$-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid receptor

**BPRS:** Brief Psychiatric Rating Scale

**CGI:** Clinical Global Impression Scale

**CpG:** Cytosine-Phosphate-Guanine

**DNA:** Deoxyribonucleic Acid

**DNAm:** Deoxyribonucleic Acid methylation

**DSM-IV:** Diagnostic and Statistical Manual of Mental Disorders version IV

**FC:** Fold Change

**GWAS:** Genome Wide Association Studies

**ICD-10:** International Classification of Diseases version 10

**lncRNA:** Long non-coding RNA

**lincRNA:** Long intergenic noncoding RNA

**NMDAR:** N-methyl-D-aspartate receptor

**PAFIP:** Programa Atención Fases Iniciales de Psicosis

**RNA:** Ribonucleic acid

**SD:** Satandard Deviation

**SNP:** Single Nucleotide Polymorphism

**SZ:** Schizophrenia

**SPSS:** Statistical Package for Social Sciences

**UCSC ID:** University of California Santa Cruz Identification

# References


[1] Saha S, Chant D, McGrath J. A systematic review of mortality in schizophrenia: is the differential mortality gap worsening over time? Archives of general psychiatry. 2007;64(10):1123–1131.

[2] McGrath J, Saha S, Chant D, Welham J. Schizophrenia: a concise overview of incidence, prevalence, and mortality. Epidemiologic reviews. 2008;30(1):67–76.

[3] Wittchen HU, Jacobi F, Rehm J, Gustavsson A, Svensson M, Jönsson B, et al. The size and burden of mental disorders and other disorders of the brain in Europe 2010. European neuropsychopharmacology. 2011;21(9):655–679.

[4] Tandon R, Nasrallah HA, Keshavan MS. Schizophrenia,"just the facts" 4. Clinical features and conceptualization. Schizophrenia research. 2009;110(1-3):1–23.

[5] Mohr PE, Cheng CM, Claxton K, Conley RR, Feldman JJ, Hargreaves WA, et al. The heterogeneity of schizophrenia in disease states. Schizophrenia research. 2004;71(1):83–95.

[6] Kochunov P, Hong LE. Neurodevelopmental and neurodegenerative models of schizophrenia: white matter at the center stage. Schizophrenia bulletin. 2014;40(4):721–728.

[7] Knoll IV JL, Garver DL, Ramberg JE, Kingsbury SJ, Croissant D, McDermott B. Heterogeneity of the psychoses: is there a neurodegenerative psychosis? Schizophrenia Bulletin. 1998;24(3):365–379.

[8] Woods BT. Is schizophrenia a progressive neurodevelopmental disorder? Toward a unitary pathogenetic mechanism. American Journal of Psychiatry. 1998;155(12):1661–1670.

[9] Gupta S, Kulhara P. What is schizophrenia: A neurodevelopmental or neurodegenerative disorder or a combination of both? A critical analysis. Indian journal of psychiatry. 2010;52(1):21.

[10] Howes O, McCutcheon R, Stone J. Glutamate and dopamine in schizophrenia: an update for the 21st century. Journal of psychopharmacology. 2015;29(2):97–115.

[11] Lieberman J, Kane J, Alvir J. Provocative tests with psychostimulant drugs in schizophrenia. Psychopharmacology. 1987;91(4):415–433.

[12] Toda M, Abi-Dargham A. Dopamine hypothesis of schizophrenia: making sense of it all. Current psychiatry reports. 2007;9(4):329–336.

[13] Hu W, MacDonald ML, Elswick DE, Sweet RA. The glutamate hypothesis of schizophrenia: evidence from human brain tissue studies. Annals of the New York Academy of Sciences. 2015;1338(1):38–57.

[14] Stone JM, Morrison PD, Pilowsky LS. Glutamate and dopamine dysregulation in schizophrenia—a synthesis and selective review. Journal of psychopharmacology. 2007;21(4):440–452.

[15] Lichtenstein P, Yip BH, Björk C, Pawitan Y, Cannon TD, Sullivan PF, et al. Common genetic determinants of schizophrenia and bipolar disorder in Swedish families: a population-based study. The Lancet. 2009;373(9659):234–239.

[16] Sullivan PF, Kendler KS, Neale MC. Schizophrenia as a complex trait: evidence from a meta-analysis of twin studies. Archives of general psychiatry. 2003;60(12):1187–1192.

[17] Collins AL, Kim Y, Sklar P, O'Donovan MC, Sullivan PF, Consortium IS, et al. Hypothesis-driven candidate genes for schizophrenia compared to genome-wide association results. Psychological medicine. 2012;42(3):607–616.

[18] Sullivan PF, Daly MJ, O'Donovan M. Genetic architectures of psychiatric disorders: the emerging picture and its implications. Nature Reviews Genetics. 2012 Jul;13(8):537–551.

[19] Voisin S, Almén MS, Zheleznyakova GY, Lundberg L, Zarei S, Castillo S, et al. Many obesity-associated SNPs strongly associate with DNA methylation changes at proximal promoters and enhancers. Genome medicine. 2015;7(1):103.

[20] Moore LD, Le T, Fan G. DNA methylation and its basic function. Neuropsychopharmacology. 2013;38(1):23.

[21] Montano C, Taub MA, Jaffe A, Briem E, Feinberg JI, Trygvadottir R, et al. Association of DNA Methylation Differences With Schizophrenia in an Epigenome-Wide Association Study. JAMA psychiatry. 2016;73(5):506–514.

[22] Voisin S, Almén MS, Zheleznyakova GY, Lundberg L, Zarei S, Castillo S, et al. Many obesity-associated SNPs strongly associate with DNA methylation changes at proximal promoters and enhancers. Genome Medicine. 2015 Oct;7:103.

[23] Ciuculete DM, Boström AE, Voisin S, Philipps H, Titova OE, Bandstein M, et al. A methylome-wide mQTL analysis reveals associations of methylation sites with GAD1 and HDAC3 SNPs and a general psychiatric risk score. Translational Psychiatry. 2017;7(1):e1002.

[24] Overall JE, Gorham DR. The brief psychiatric rating scale. Psychological reports. 1962;10(3):799–812.

[25] Andersen J, Larsen J, Schultz V, Nielsen B, Kørner A, Behnke K, et al. The brief psychiatric rating scale. Psychopathology. 1989;22(2-3):168–176.

[26] Leucht S, Kane JM, Kissling W, Hamann J, Etschel E, Engel R. Clinical implications of Brief Psychiatric Rating Scale scores. The British Journal of Psychiatry: The Journal of Mental Science. 2005 Oct;187:366–371.

[27] Guy W. ECDEU assessment manual for psychopharmacology. US Department of Health, and Welfare. 1976;p. 534–537.

[28] Psychiatric Genomics Consortium;. `www.med.unc.edu/pgc`.

[29] Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. The American journal of human genetics. 2007;81(3):559–575.

[30] PLINK;. Available from: `http://pngu.mgh.harvard.edu/purcell/plink/`.

[31] Infinium Human Methylation 450 BeadChip;. `support.illumina.com/content/dam/illumina-marketing/documents/products/product_information_sheets/product_info_hm450.pdf`, last accessed on 20/05/19.

[32] Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, Le JM, et al. High density DNA methylation array with single CpG site resolution. Genomics. 2011;98(4):288–295.

[33] Wang Z, Wu X, Wang Y. A framework for analyzing DNA methylation data from Illumina Infinium HumanMethylation450 BeadChip. BMC bioinformatics. 2018;19(5):115.

[34] Du P, Zhang X, Huang CC, Jafari N, Kibbe WA, Hou L, et al. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC bioinformatics. 2010;11(1):587.

[35] Irizarry RA, Ladd-Acosta C, Carvalho B, Wu H, Brandenburg SA, Jeddeloh JA, et al. Comprehensive high-throughput arrays for relative methylation (CHARM). Genome research. 2008;18(5):780–790.

[36] Bibikova M, Fan JB. GoldenGate® assay for DNA methylation profiling. In: DNA Methylation. Springer; 2009. p. 149–163.

[37] Bibikova M, Lin Z, Zhou L, Chudin E, Garcia EW, Wu B, et al. High-throughput DNA methylation profiling using universal bead arrays. Genome research. 2006;16(3):383–393.

[38] Xu J, Sun J, Chen J, Wang L, Li A, Helm M, et al. RNA-Seq analysis implicates dysregulation of the immune system in schizophrenia. BMC genomics. 2012;13(8):S2.

[39] Sainz J, Mata I, Barrera J, Perez-Iglesias R, Varela I, Arranz MJ, et al. Inflammatory and immune response genes have significantly altered expression in schizophrenia. Molecular psychiatry. 2013;18(10):1056.

[40] Crespo-Facorro B, Prieto C, Sainz J. Schizophrenia gene expression profile reverted to normal levels by antipsychotics. International Journal of Neuropsychopharmacology. 2015;18(4).

[41] Sanders A, Drigalenko E, Duan J, Moy W, Freda J, Göring H, et al. Transcriptome sequencing study implicates immune-related genes differentially expressed in schizophrenia: new data and a meta-analysis. Translational psychiatry. 2017;7(4):e1093.

[42] Li X, Teng S. RNA sequencing in schizophrenia. Bioinformatics and biology insights. 2015;9:BBI–S28992.

[43] Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, et al. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics. 2005;21(16):3439–3440.

[44] Infinium PsychArray-24 Kit;. `http://emea.support.illumina.com/array/array_kits/infinium-psycharray-beadchip-kit/downloads.html?langsel=/es/`, last accessed on 18/05/19.

[45] HumanOmniExpress-24 v1.1 BeadChip;. `http://emea.support.illumina.com/downloads/humanomniexpress-24-v1-1-support-files.html?langsel=/es/`, last accessed on 18/05/19.

[46] SNP Nexus tool;. `https://www.snp-nexus.org/`, last accessed on 20/05/19.

[47] R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2014. Available from: `http://www.R-project.org/`.

[48] Ransohoff JD, Wei Y, Khavari PA. The functions and unique features of long intergenic non-coding RNA. Nature reviews Molecular cell biology. 2018;19(3):143.

[49] Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, et al. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes & development. 2011;25(18):1915–1927.

[50] Guttman M, Rinn JL. Modular regulatory principles of large non-coding RNAs. Nature. 2012;482(7385):339.

[51] Wu P, Zuo X, Deng H, Liu X, Liu L, Ji A. Roles of long noncoding RNAs in brain development, functional diversification and neurodegenerative diseases. Brain research bulletin. 2013;97:69–80.

[52] Arisi I, D'Onofrio M, Brandi R, Felsani A, Capsoni S, Drovandi G, et al. Gene expression biomarkers in the brain of a mouse model for Alzheimer's disease: mining of microarray data by logic classification and feature selection. Journal of Alzheimer's Disease. 2011;24(4):721–738.

[53] Sai Y, Zou Z, Peng K, Dong Z. The Parkinson's disease-related genes act in mitochondrial homeostasis. Neuroscience & Biobehavioral Reviews. 2012;36(9):2034–2043.

[54] Chen S, Sun X, Niu W, Kong L, He M, Li W, et al. Aberrant expression of long non-coding RNAs in schizophrenia patients. Medical science monitor: international medical journal of experimental and clinical research. 2016;22:3340.

[55] Barry G, Briggs J, Vanichkina D, Poth E, Beveridge N, Ratnu V, et al. The long non-coding RNA Gomafu is acutely regulated in response to neuronal activation and involved in schizophrenia-associated alternative splicing. Molecular psychiatry. 2014;19(4):486.

[56] Rao SQ, Hu HL, Ye N, Shen Y, Xu Q. Genetic variants in long non-coding RNA MIAT contribute to risk of paranoid schizophrenia in a Chinese Han population. Schizophrenia research. 2015;166(1-3):125–130.

[57] Ren Y, Cui Y, Li X, Wang B, Na L, Shi J, et al. A co-expression network analysis reveals lncRNA abnormalities in peripheral blood in early-onset schizophrenia. Progress in Neuro-Psychopharmacology and Biological Psychiatry. 2015;63:1–5.

[58] Sheng M, Kim E. The Shank family of scaffold proteins. J Cell Sci. 2000;113(11):1851–1856.

[59] Berkel S, Marshall CR, Weiss B, Howe J, Roeth R, Moog U, et al. Mutations in the SHANK2 synaptic scaffolding gene in autism spectrum disorder and mental retardation. Nature genetics. 2010;42(6):489.

[60] Guilmatre A, Huguet G, Delorme R, Bourgeron T. The emerging role of SHANK genes in neuropsychiatric disorders. Developmental neurobiology. 2014;74(2):113–122.

[61] Peykov S, Berkel S, Schoen M, Weiss K, Degenhardt F, Strohmaier J, et al. Identification and functional characterization of rare SHANK2 variants in schizophrenia. Molecular psychiatry. 2015;20(12):1489.

[62] Homann OR, Misura K, Lamas E, Sandrock RW, Nelson P, McDonough SI, et al. Whole-genome sequencing in multiplex families with psychoses reveals mutations in the SHANK2 and SMARCA1 genes segregating with illness. Molecular psychiatry. 2016;21(12):1690.

[63] Schmeisser MJ. Translational neurobiology in Shank mutant mice-Model systems for neuropsychiatric disorders. Annals of Anatomy-Anatomischer Anzeiger. 2015;200:115–117.

[64] Won H, Lee HR, Gee HY, Mah W, Kim JI, Lee J, et al. Autistic-like social behaviour in Shank2-mutant mice improved by restoring NMDA receptor function. Nature. 2012;486(7402):261.

[65] Kolarova J, Tangen I, Bens S, Gillessen-Kaesbach G, Gutwein J, Kautza M, et al. Array-based DNA methylation analysis in individuals with developmental delay/intellectual disability and normal molecular karyotype. European journal of medical genetics. 2015;58(8):419–425.

[66] Eltokhi A, Rappold G, Sprengel R. Distinct phenotypes of shank2 mouse models reflect neuropsychiatric spectrum disorders of human patients with shank2 variants. Frontiers in molecular neuroscience. 2018;11.

[67] Lang F, Huang DY, Vallon V. SGK, renal function and hypertension. Journal of nephrology. 2010;23(0 16):S124.

[68] Lang F, Strutz-Seebohm N, Seebohm G, Lang UE. Significance of SGK1 in the regulation of neuronal function. The Journal of physiology. 2010;588(18):3349–3354.

[69] Ma YL, Tsai MC, Hsu WL, Lee EH. SGK protein kinase facilitates the expression of long-term potentiation in hippocampal neurons. Learning & Memory. 2006;13(2):114–118.

[70] Lang F, Bohmer C, Palmada M, Seebohm G, Strutz-Seebohm N, Vallon V. (Patho) physiological significance of the serum-and glucocorticoid-inducible kinase isoforms. Physiological reviews. 2006;86(4):1151–1178.

[71] Sakai R, Irie Y, Murata T, Ishige A, Anjiki N, Watanabe K. Toki-to protects dopaminergic neurons in the substantia nigra from neurotoxicity of MPTP in mice. Phytotherapy Research: An International Journal Devoted to Pharmacological and Toxicological Evaluation of Natural Product Derivatives. 2007;21(9):868–873.

[72] Frodl T, Carballedo A, Hughes M, Saleh K, Fagan A, Skokauskas N, et al. Reduced expression of glucocorticoid-inducible genes GILZ and SGK-1: high IL-6 levels are associated with reduced hippocampal volumes in major depressive disorder. Translational psychiatry. 2012;2(3):e88.

[73] Singh PK, Singh S, Ganesh S. Activation of serum/glucocorticoid-induced kinase 1 (SGK1) underlies increased glycogen levels, mTOR activation, and autophagy defects in Lafora disease. Molecular biology of the cell. 2013;24(24):3776–3786.

[74] Lang F, Artunc F, Vallon V. The physiological impact of the serum-and glucocorticoid-inducible kinase SGK1. Current opinion in nephrology and hypertension. 2009;18(5):439.

[75] Mitchell ES, Conus N, Kaput J. B vitamin polymorphisms and behavior: Evidence of associations with neurodevelopment, depression, schizophrenia, bipolar disorder and cognitive decline. Neuroscience & Biobehavioral Reviews. 2014;47:307–320.

[76] Brown HE, Roffman JL. Vitamin supplementation in the treatment of schizophrenia. CNS drugs. 2014;28(7):611–622.

[77] Kemperman RF, Veurink M, van der Wal T, Knegtering H, Bruggeman R, Fokkema MR, et al. Low essential fatty acid and B-vitamin status in a subgroup of patients with schizophrenia and its response to dietary supplementation. Prostaglandins, leukotrienes and essential fatty acids. 2006;74(2):75–85.

# Appendices

**A.** R CODE FOR SNPs ASSOCIATION ANALYSIS

```
# Read plink files from SNP association analysis
-----------------------------------------------------------
dt <-read.table("SNPanalysis.qassoc", header=T)
SNPassoc<-write.table(dt,"SNPanalysisqassoc.txt")

# Remove NA values
-----------------------------------------------------------
dt2<- na.omit(dt)

# Remove rows with NAs in BETA column & Sort table by ascending p-values
-----------------------------------------------------------
pval_ord<-(dt[!is.na(dt$BETA),])[order((dt[!is.na(dt$BETA),])$P),]

# Selecting top 20 p-value positions

top20<- pval_ord[1:20,]

#Select SNPs with p-values < 0.05
-----------------------------------------------------------
# Selecting all cases with p-values < 0.05

allpval<-subset(pval_ord, pval_ord$P <=0.05)
tail(allpval)
dim(allpval) # 25950 SNPs with p-values < 0.05

mergedSNPsbimALL <- merge(allpval,bimPositions, by.x = 2,
by.y = 1, all.x = TRUE, all.y = TRUE)
snpListALLpval <- allpval$SNP

write.table(snpListALLpval,"snpListALLpval.txt")
write.xlsx(snpListALLpval,"snpListALLpval.xlsx")

mergedSNPsbimALL[mergedSNPsbimALL$SNP %in% snpListALLpval,]
selALLSNPsbim<-mergedSNPsbimALL[mergedSNPsbimALL$SNP
%in% snpListALLpval,]

write.table(selALLSNPsbim,"selALLSNPsbim.txt")

write.xlsx(selALLSNPsbim,"selALLSNPsbim.xlsx")

# Apply Bonferroni correction
-----------------------------------------------------------
dim(dt2)
pvalue=0.05
```

```
adjpval<-pvalue/nrow(dt2)

# Select SNPs with pvalue<pbonfer
----------------------------------------------------------
SNpsA<-subset(dt2, dt2$P <=adjpval)
write.xlsx(SNpsA, "SNpsA.xlsx")

#CHR       SNP       BP NMISS BETA        SE R2         T
P
#782568  16 rs9936526 60604961     3   -2 9.424e-08  1 -21220000 3e-08

# Selecting SNPs with pvalue<=0.001)
----------------------------------------------------------
pvalue=0.001
valSNP<-as.numeric(as.character(dt2$P))
SNP0.001<-subset(dt2, valSNP <=0.001)
write.xlsx(SNP0.001,"SNP0.001.xlsx")
dim(SNP0.001) # 258

# Selecting SNPs with pvalue<=0.01)
----------------------------------------------------------
valSNP<-as.numeric(as.character(dt2$P))
SNP0.01<-subset(dt2, valSNP <=0.01)
dim(SNP0.01) # 6437

#Basic Manhattan Plot
----------------------------------------------------------
library(qqman)
SNpsA<-subset(dt2, dt2$P <=adjpval)
manhattan(dt2, chr = "CHR", bp = "BP", p = "P", snp = "SNP",
col = c("gray10", "gray60"), chrlabs = c(1:22), suggestiveline =
-log10(1e-05), genomewideline = -log10(adjpval),
highlight = SNpsA, logp = TRUE)


#1. Annotation Strategy 1 based on Ensembl
----------------------------------------------------------
library(biomaRt)
snp.ensembl <- useEnsembl(biomart = "snp", dataset = "hsapiens_snp")
class(snp.ensembl)

## Annotation for  SNPs (pvalue < 0,05)

allpval<-subset(pval_ord, pval_ord$P <=0.05)
snp.ensembl <- useEnsembl(biomart = "snp", dataset = "hsapiens_snp")
class(snp.ensembl)

snpListALL <- allpval$SNP

out.ALL <- getBM(
```

```
attributes = c('ensembl_gene_stable_id', 'refsnp_id',
'chr_name', 'chrom_start', 'chrom_end', 'minor_allele', 'minor_allele_freq'),
filters = 'snp_filter',
values = snpListALL,
mart = snp.ensembl.grch37
)
out.ALL


selALL<-out.ALL[,c("ensembl_gene_stable_id","refsnp_id","chr_name")]

selALL<-allpval[,c("SNP", "BETA", "P")]

mergedALL <- merge(selALL,selALL, by.x = 2, by.y = 1, all.x = TRUE,
all.y = TRUE)


# 2. Annotation Strategy 2 based on PsychArrays
-------------------------------------------------------

#ARRAY PsychArray_A_ (Not the proper Psycharray)
-------------------------------------------------------
annotationA<-read.table("PsychArray_A_annotated.txt",
header=F, fill=T)
head(annotationA)

annotA<-annotationA[-1,]
head(annotA)

colnames(annotA) <- c("SNP", "Chr", "MapInfo",
"Alleles", "Transcript(s)","Gene(s)","In-exon","Mutation(s)")

mergedSNPsA <- merge(top20,annotA, by.x = 2, by.y = 1, all.x = TRUE,
all.y = TRUE)
snpList20 <- top20$SNP

mergedSNPsA[mergedSNPsA$SNP %in% snpList20,]
selA<-mergedSNPs1[mergedSNPsA$SNP %in% snpList20,]

subset(annotA, SNP=="rs4970383") #SNP on ArraY
subset(annotA, SNP=="rs10505477") # SNP NOT on ArraY
subset(annotA, SNP=="rs42905") #SNP on ArraY
subset(annotA, SNP=="rs5760918") # SNP NOT on ArraY
subset(annotA, SNP=="rs7730928") # SNP NOT on ArraY

#ARRAY 1.2 (Not the proper Psycharray)
-------------------------------------------------------
##WORKS BUT IT IS NOT THE CORRRECT PSYCHARRAY

annotation<-read.table("InfiniumPsychArray-24v1-2_A1.
```

```
annotated.txt", header=F, fill=T)
head(annotation)

annot<-annotation[-1,]
colnames(annot) <- c("SNP", "Chr", "MapInfo",
"Alleles", "Transcript(s)","Gene(s)","In-exon","Mutation(s)")

mergedSNPs <- merge(top20,annot, by.x = 2, by.y = 1, all.x = TRUE,
all.y = TRUE)
snpList20 <- top20$SNP


mergedSNPs[mergedSNPs$SNP %in% snpList20,]
sel2<-mergedSNPs[mergedSNPs$SNP %in% snpList20,]

subset(annot, SNP=="rs4970383") #SNP on ArraY 1.2
subset(annot, SNP=="rs10505477") # SNP NOT on ArraY 1.2

#ARRAY 1.1 (Not the proper Psycharray)
-----------------------------------------------------

annotation1<-read.table("InfiniumPsychArray-24v1-1_A1.
annotated.txt", header=F, fill=T)
head(annotation1)

annot1<-annotation1[-1,]
colnames(annot1) <- c("SNP", "Chr", "MapInfo",
"Alleles", "Transcript(s)","Gene(s)","In-exon","Mutation(s)")

mergedSNPs1 <- merge(top20,annot1, by.x = 2, by.y = 1, all.x = TRUE,
 all.y = TRUE)
snpList20 <- top20$SNP

mergedSNPs1[mergedSNPs1$SNP %in% snpList20,]
sel1<-mergedSNPs1[mergedSNPs1$SNP %in% snpList20,]

subset(annot1, SNP=="rs4970383") #SNP on ArraY
subset(annot1, SNP=="rs10505477") # SNP NOT on ArraY
subset(annot1, SNP=="rs42905") #SNP on ArraY
subset(annot1, SNP=="rs5760918") # SNP NOT on ArraY


#ARRAY 1.3 (Not the proper Psycharray)
-----------------------------------------------------

annotation3<-read.table("InfiniumPsychArray-24v1-3_A1.hg19.
annotated.txt", header=F, fill=T)
head(annotation3)

annot3<-annotation3[-1,]
```

```
colnames(annot3) <- c("SNP", "Chr", "MapInfo",
"Alleles", "Transcript(s)","Gene(s)","In-exon","Mutation(s)")

mergedSNPs3 <- merge(top20,annot3, by.x = 2, by.y = 1, all.x = TRUE,
all.y = TRUE)
snpList20 <- top20$SNP

mergedSNPs3[mergedSNPs3$SNP %in% snpList20,]
sel3<-mergedSNPs3[mergedSNPs3$SNP %in% snpList20,]

subset(annot3, SNP=="rs4970383") #SNP on ArraY
subset(annot3, SNP=="rs10505477") # SNP NOT on ArraY
subset(annot3, SNP=="rs42905") #SNP on ArraY
subset(annot3, SNP=="rs5760918") # SNP NOT on ArraY
subset(annot3, SNP=="rs7730928") # SNP NOT on ArraY


# 3. Annotation Strategy 3 based on HumanOmni

#ARRAY HumanOmniExpress-24 v1.1 (Not the proper Psycharray)
-------------------------------------------------------------

annotationOmni1<-read.table("HumanOmniExpress-24v1-1_A.
annotated.txt", header=F, fill=T)

head(annotationOmni1)

Omni1<-annotationOmni1[-1,]
head(Omni1)
colnames(Omni1) <- c("SNP", "Chr", "MapInfo",
"Alleles", "Transcript(s)","Gene(s)","In-exon","Mutation(s)")

mergedSNPsOmni1 <- merge(top20,Omni1, by.x = 2,
by.y = 1, all.x = TRUE, all.y = TRUE)
snpList20 <- top20$SNP

mergedSNPsOmni1[mergedSNPsOmni1$SNP %in% snpList20,]
selOmni1<-mergedSNPsOmni1[mergedSNPsOmni1$SNP %in% snpList20,]

subset(Omni1, SNP=="rs4970383") #SNP NOT ON ArraY
subset(Omni1, SNP=="rs10505477") # SNP  on ArraY
subset(Omni1, SNP=="rs42905") #SNP NOT  ArraY
subset(Omni1, SNP=="rs5760918") # SNP  on ArraY
subset(Omni1, SNP=="rs7730928") # SNP  on ArraY

View(selOmni1)


# 4. Annotation Strategy 4 based on bim coordinates
-------------------------------------------------------------------
```

```
##THIS STRATEGY GIVES POSITION FOR ALL SNPS IN THE ANALYSIS

bim <-read.table("metilationNY.bim", header=T) #files from 23.04.19
head(bim)
colnames(bim)

bimsel<-bim[,c(2,4)]
colnames(bimsel)<-c("SNP", "Position")

frow = data.frame(SNP='rs4477212', Position='82154',
stringsAsFactors = FALSE)

bimPositions<-rbind(frow,bimsel)

write.table(bimPositions,"bimPositions.txt")
bimPositions=read.table("bimPositions.txt",dec=",",header=TRUE)

mergedSNPsbim <- merge(top20,bimPositions, by.x = 2,
by.y = 1, all.x = TRUE, all.y = TRUE)
snpList20 <- top20$SNP

mergedSNPsbim[mergedSNPsbim$SNP %in% snpList20,]
sel20SNPsbim<-mergedSNPsbim[mergedSNPsbim$SNP %in% snpList20,]

write.table(sel20SNPsbim,"sel20SNPsbim.txt")

write.xlsx(sel20SNPsbim,"sel20SNPsbim.xlsx")

View(sel20SNPsbim)

subset(mergedSNPsbim , SNP=="rs10505477") # SNP  on ArraY
subset(mergedSNPsbim , SNP=="rs42905") #SNP on  ArraY
```

**B.** R CODE FOR CpGs ASSOCIATION ANALYSIS

```
# Read methylation data
-----------------------------------------------------------------
#methylation data
a <-read.table("bvalues.tsv",header=T)
met <- as.data.frame(t(a))
met$sample_id <- row.names(met)
#demographic data
dem<- read.table("id_age_severity.txt",header=T,sep =""  )

#change rownames in dem file
```

```
id<-dem$sample_id
row.names(dem)<-id
dem2<-dem[,2:3]

#merge methylation and demographic data
dt_merged <- cbind(dem2,met)

#convert CGI score from categorical to a numerical variable
dt_merged$CGI_score<-as.numeric(dt_merged$CGI_score)

#remove column sample_id from the merged dataframe
subset(dt_merged, select=-c(sample_id))->dt_mergedrem


# Logistic Regression Model
-----------------------------------------------------------------
dt_test <- dt_mergedrem[ , ! apply( dt_mergedrem , 2 ,
function(x) any(is.na(x)) ) ]
#p values

m2.pval<- apply(dt_test, 2, function(x) summary(lm(log(dt_test$CGI_score)
~ x + dt_test$edad))$coefficients[2,4])
write.foreign(m2.pval, "m2.pval.sps","m2.pval.txt", package="SPSS")
#Export data to .txt and SPSS files

#Betas
m2.beta <- apply(dt_test, 2, function(x) summary(lm(log
(dt_test$CGI_score) ~ x + dt_test$edad))$coefficients[2,1])
write.foreign(m2.beta, "m2.beta.sps","m2.beta.txt", package="SPSS")
#Export data to .txt and SPSS files


#Select only pvalues and beta values and merge them in a dataframe
-----------------------------------------------------------------
pvalmethyl<-as.data.frame(m2.pval)
betamethyl<-as.data.frame(m2.beta)
cpgsALL<-cbind(pvalmethyl,betamethyl)
cpgs<-cpgsALL[-c(1, 2), ]

cpgsorder<-cpgs[order(cpgs$m2.pval, decreasing = FALSE),]

namesCpGs<-rownames(cpgsorder)
as.data.frame(namesCpGs)
cpgs2<-cbind(namesCpGs,cpgsorder)
rownames(cpgs2) <- c()
head(cpgs2)

## Annotation for ALL CpGs
-----------------------------------------------------------------
install.packages("IlluminaHumanMethylation450kanno.ilmn12.
```

```
hg19")
library("IlluminaHumanMethylation450kanno.ilmn12.
hg19",
lib.loc="~/R/win-library/3.5")
ann450k = getAnnotation(IlluminaHumanMethylation450kanno.ilmn12.hg19)
head(ann450k)
position<-ann450k[, 1:2]
select450k<-ann450k[, c("chr","pos",
"Name", "UCSC_RefGene_Name","UCSC_RefGene_Accession")]
cpgsMergedALL <- merge(select450k,cpgs2, by.x = 3, by.y = 1, all.x
```

**C.** R CODE for integration analysis of genetic and methylation data

```
# pvalue <0.001
-----------------------------------------------------------------
-----------------------------------------------------------------
# Selecting SNPs for Venn Diagram-> pvalue <0.001
-----------------------------------------------------------------
pvalue=0.001
valSNP<-as.numeric(as.character(dt2$P))
SNP0.001<-subset(dt2, valSNP <=0.001)
write.xlsx(SNP0.001,"SNP0.001.xlsx")
dim(SNP0.001) # 258
SNPvenn<-SNP0.001$SNP
library("openxlsx", lib.loc="~/R/win-library/3.5")
write.table(SNPvenn,"SNPvenn.txt") -> #SNPs for NEXUS
write.xlsx(SNPvenn, "SNPvenn.xlsx")
dim(SNPvenn)

VennmergedSNP <- merge(SNP0.001,bimPositions, by.x = 2,
by.y = 1, all.x = TRUE, all.y = TRUE)

VennmergedSNP[VennmergedSNP$SNP %in% SNPvenn,]
selSNPVenn<-VennmergedSNP[VennmergedSNP$SNP %in% SNPvenn,]

write.table(selSNPVenn,"selSNPVenn.txt")
write.xlsx(selSNPVenn, "selSNPVenn0.001.xlsx")

### FINAL LIST FOR VENN PLOT
SNPVENNcleared=read.xlsx("SNPVENNcleared.xlsx")

# Selecting CpGs for Venn Diagram-> pvalue <0.001
-----------------------------------------------------------------
pvalue=0.001
val<-as.numeric(as.character(cpgsMergedALLtable$P))
cpgs0.001<-subset(cpgsMergedALLtable, val <=pvalue)
```

```
write.xlsx(cpgs0.001,"cpgs0.001.xlsx")
dim(cpgs0.001) # 499
head(cpgs0.001)
selCPGsVenn<-cpgs0.001[,4]
write.xlsx(selCPGsVenn,"selCPGsVenn.xlsx")

### FINAL LIST FOR VENN PLOT
CPGsVENNcleared=read.xlsx("CPGsVENNcleared.xlsx")

# Find intersection SNP and CpGs
-----------------------------------------------------------------
library(data.table)
dim(SNPVENNcleared) #103
dim(CPGsVENNcleared)#408
intersect0.001=fintersect(setDT(SNPVENNcleared),
setDT(CPGsVENNcleared)) #dim=4

##VENN DIAGRAM for intersected genes
-----------------------------------------------------------------
dim(SNPVENNcleared) #104
dim(CPGsVENNcleared)#408
grid.newpage()
draw.pairwise.venn(area1 = 104, area2 = 408,
cross.area = 4, category = c("Significant␣SNPs␣(P<0.001)",
"Significant␣CpGs␣(P<0.001)"))

# pvalue <0.01
-----------------------------------------------------------------
-----------------------------------------------------------------

# Selecting SNPs for Venn Diagram-> pvalue <0.01
-----------------------------------------------------------------

valSNP<-as.numeric(as.character(dt2$P))
SNP0.01<-subset(dt2, valSNP <=0.01)
dim(SNP0.01) # 6437
SNPvenn0.01<-SNP0.01$SNP
write.table(SNPvenn0.01,"SNPvenn0.01.txt")  #SNPs for NEXUS
write.xlsx(SNPvenn0.01, 'SNPvenn0.01.xlsx')
bimPositions=read.table("bimPositions.txt",dec=",",header=TRUE)
VennmergedSNP0.01 <- merge(SNPvenn0.01,bimPositions, by.x = 2,
by.y = 1, all.x = TRUE, all.y = TRUE)
VennmergedSNP0.01[VennmergedSNP0.01$SNP %in% SNPvenn0.01,]
selSNPVenn0.01<-VennmergedSNP0.01[VennmergedSNP0.01$SNP %in%
SNPvenn0.01,]
write.table(selSNPVenn0.01,"selSNPVenn0.01.txt")
selSNPVenn0.01=read.table("selSNPVenn0.01.txt",dec=",",header=TRUE)
selSNPVenn0.01modif= write.xlsx(selSNPVenn0.01, "selSNPVenn0.01.xlsx")
dim(selSNPVenn0.01)
ucscnameALLgenesNEXUS=read.xlsx("ucscnameALLgenesNEXUS.xlsx")
```

```
### LIST FROM NEXUS


# Selecting CpGs for Venn Diagram-> pvalue <0.01
-----------------------------------------------------------------
pvalue=0.01
val<-as.numeric(as.character(cpgsMergedALLtable$P))
cpgs0.01<-subset(cpgsMergedALLtable, val <=pvalue)
write.xlsx(cpgs0.01,"cpgs0.01.xlsx")
selCPGsVenn001<-cpgs0.01[,4]
write.xlsx(selCPGsVenn001,"selCPGsVenn001.xlsx")


### FINAL LIST FOR VENN PLOT
CPGsVENNcleared0.01=read.xlsx("selCPGsVenn0.01Cleared.xlsx")
 ### FINAL LIST FOR VENN PLOT

# Find intersection SNP and CpGs
-----------------------------------------------------------------
dim(ucscnameALLgenesNEXUS) #1586
dim(CPGsVENNcleared0.01) #3791
library(data.table)
dim(ucscnameALLgenesNEXUS) #1586
colnames(ucscnameALLgenesNEXUS) <- c("GENE")

intersect001<-fintersect(setDT(ucscnameALLgenesNEXUS), setDT(CPGsVENNcleared0.01))
dim(intersect001) #341
write.xlsx(intersect001,"intersect001.xlsx")


##VENN DIAGRAM for intersected genes
-----------------------------------------------------------------
library(VennDiagram)
grid.newpage()
draw.pairwise.venn(1586, 3791, 341, category = c("Significant
SNPs␣(P<0.01)", "Significant␣CpGs␣(P<0.01)"), lty =
rep("blank", 2), fill = c("light␣blue", "pink"), alpha =
rep(0.5, 2), cat.pos = c(2,2), cat.dist = rep(0.025, 2))
```

**D.** R CODE for integrative analysis with differentially expressed genes

```
# Find intersect CGPs,SNP, and DE genes (Sanders et al, 2017)
-----------------------------------------------------------------
library("openxlsx", lib.loc="~/R/win-library/3.5")
DEgenesSanders<- read.xlsx("DEgenesSanders.xlsx", 1) #1058
colnames(DEgenesSanders) <- c("GENE")
```

```
dim(DEgenesSanders) #1058

library(data.table)

intersect001Sanders<-fintersect(setDT(intersect001), setDT(DEgenesSanders))
intersect001Sanders
dim(intersect001Sanders) #16
write.table(intersect001Sanders,"intersect001Sanders.txt")
read.table("intersect001Sanders.txt", header=T,sep ="")

GENE
1  COL24A1
2    IL5RA
3   ZBTB38
4     CLNK
5     IL15
6     SGK1
7     NUB1
8    RBPMS
9     ANK1
10   WDR37
11   WDFY4
12  SHANK2
13   KDM2B
14    GJA3
15    GNG7
16    TCN2

# SNP related information

#1   COL24A1
COL24A1_SNP<-subset(genesSNP0.01MERGED, Symbol =="COL24A1" )
write.xlsx(COL24A1_SNP,"COL24A1_SNP.xlsx", sheetName="Sheet1")
#2    IL5RA
IL5RA_SNP<-subset(genesSNP0.01MERGED, Symbol =="IL5RA" )
write.xlsx(IL5RA_SNP,"IL5RA_SNP.xlsx", sheetName="Sheet1")
#3   ZBTB38
ZBTB38_SNP<-subset(genesSNP0.01MERGED, Symbol =="ZBTB38" )
write.xlsx(ZBTB38_SNP,"ZBTB38_SNP.xlsx", sheetName="Sheet1")
#4     CLNK
CLNK_SNP<-subset(genesSNP0.01MERGED, Symbol =="CLNK" )
write.xlsx(CLNK_SNP,"CLNK_SNP.xlsx", sheetName="Sheet1")
#5     IL15
IL15_SNP<-subset(genesSNP0.01MERGED, Symbol =="IL15" )
write.xlsx(IL15_SNP,"IL15_SNP.xlsx", sheetName="Sheet1")
#6     SGK1
SGK1_SNP<-subset(genesSNP0.01MERGED, Symbol =="SGK1" )
write.xlsx(SGK1_SNP,"SGK1_SNP.xlsx", sheetName="Sheet1")
#7     NUB1
NUB1_SNP<-subset(genesSNP0.01MERGED, Symbol =="NUB1" )
```

```
write.xlsx(NUB1_SNP,"NUB1_SNP.xlsx", sheetName="Sheet1")
#8    RBPMS
RBPMS_SNP<-subset(genesSNP0.01MERGED, Symbol =="RBPMS" )
write.xlsx(RBPMS_SNP,"RBPMS_SNP.xlsx", sheetName="Sheet1")
#9     ANK1
ANK1_SNP<-subset(genesSNP0.01MERGED, Symbol =="ANK1" )
write.xlsx(ANK1_SNP,"ANK1_SNP.xlsx", sheetName="Sheet1")
#10   WDR37
WDR37_SNP<-subset(genesSNP0.01MERGED, Symbol =="WDR37" )
write.xlsx(WDR37_SNP,"WDR37_SNP.xlsx", sheetName="Sheet1")
#11   WDFY4
WDFY4_SNP<-subset(genesSNP0.01MERGED, Symbol =="WDFY4" )
write.xlsx(WDFY4_SNP,"␣WDFY4_SNP.xlsx", sheetName="Sheet1")
#12  SHANK2
SHANK2_SNP<-subset(genesSNP0.01MERGED, Symbol =="SHANK2" )
write.xlsx(SHANK2_SNP,"SHANK2_SNP.xlsx", sheetName="Sheet1")
#13   KDM2B
KDM2B_SNP<-subset(genesSNP0.01MERGED, Symbol =="KDM2B" )
write.xlsx(KDM2B_SNP,"KDM2B_SNP.xlsx", sheetName="Sheet1")
#14    GJA3
GJA3_SNP<-subset(genesSNP0.01MERGED, Symbol =="GJA3")
write.xlsx(GJA3_SNP,"GJA3_SNP.xlsx", sheetName="Sheet1")
#15    GNG7
GNG7_SNP<-subset(genesSNP0.01MERGED, Symbol =="GNG7")
write.xlsx(GNG7_SNP,"GNG7_SNP.xlsx", sheetName="Sheet1")
#16    TCN2
TCN2_SNP<-subset(genesSNP0.01MERGED, Symbol =="TCN2" )
write.xlsx(TCN2_SNP,"TCN2_SNP.xlsx", sheetName="Sheet1")

# CpG related information

#1   COL24A1
COL24A1_CpG<-subset(cpgs001, UCSC_RefGene_Name  =="COL24A1" )
write.xlsx(COL24A1_CpG,"COL24A1_CpG.xlsx", sheetName="Sheet1")
#2    IL5RA
IL5RA_CpG<-subset(cpgs001, UCSC_RefGene_Name =="IL5RA" )
write.xlsx(IL5RA_CpG,"IL5RA_CpG.xlsx", sheetName="Sheet1")
#3   ZBTB38
ZBTB38_CpG<-subset(cpgs001, UCSC_RefGene_Name =="ZBTB38" )
write.xlsx(ZBTB38_CpG,"ZBTB38_CpG.xlsx", sheetName="Sheet1")
#4     CLNK
CLNK_CpG<-subset(cpgs001, UCSC_RefGene_Name =="CLNK" )
write.xlsx(CLNK_CpG,"CLNK_CpG.xlsx", sheetName="Sheet1")
#5     IL15
IL15_CpG<-subset(cpgs001, UCSC_RefGene_Name=="IL15" )
write.xlsx(IL15_CpG,"IL15_CpG.xlsx", sheetName="Sheet1")
#6     SGK1
SGK1_CpG<-subset(cpgs001, UCSC_RefGene_Name =="SGK1" )
write.xlsx(SGK1_CpG,"SGK1_CpG.xlsx", sheetName="Sheet1")
#7     NUB1
```

```
NUB1_SNP<-subset(cpgs001, UCSC_RefGene_Name =="NUB1" )
write.xlsx(NUB1_SNP,"NUB1_CpG.xlsx", sheetName="Sheet1")
#8    RBPMS
RBPMS_CpG<-subset(cpgs001, UCSC_RefGene_Name =="RBPMS" )
write.xlsx(RBPMS_CpG,"RBPMS_CpG.xlsx", sheetName="Sheet1")
#9     ANK1
ANK1_CpG<-subset(cpgs001, UCSC_RefGene_Name =="ANK1")
write.xlsx(ANK1_CpG,"ANK1_CpG.xlsx", sheetName="Sheet1")
#10   WDR37
WDR37_CpG<-subset(cpgs001, UCSC_RefGene_Name =="WDR37" )
write.xlsx(WDR37_CpG,"WDR37_CpG.xlsx", sheetName="Sheet1")
#11   WDFY4
WDFY4_CpG<-subset(cpgs001, UCSC_RefGene_Name =="WDFY4" )
write.xlsx(WDFY4_CpG,"WDFY4_CpG.xlsx", sheetName="Sheet1")
#12  SHANK2
SHANK2_CpG<-subset(cpgs001, UCSC_RefGene_Name =="SHANK2")
write.xlsx(SHANK2_CpG,"SHANK2_CpG.xlsx", sheetName="Sheet1")
#13   KDM2B
KDM2B_CpG<-subset(cpgs001, UCSC_RefGene_Name =="KDM2B" )
write.xlsx(KDM2B_CpG,"KDM2B_CpG.xlsx", sheetName="Sheet1")
#14    GJA3
GJA3_CpG<-subset(cpgs001, UCSC_RefGene_Name=="GJA3")
write.xlsx(GJA3_CpG,"GJA3_CpG.xlsx", sheetName="Sheet1")
#15    GNG7
GNG7_CpG<-subset(cpgs001, UCSC_RefGene_Name=="GNG7")
write.xlsx(GNG7_CpG,"GNG7_CpG.xlsx", sheetName="Sheet1")
#16    TCN2
TCN2_CpG<-subset(cpgs001, UCSC_RefGene_Name =="TCN2" )
write.xlsx(TCN2_CpG,"TCN2_CpG.xlsx", sheetName="Sheet1")



# FIND INTERSECT CGPs,SNP, and 200 DE genes (Sainz et al, 2013)
-----------------------------------------------------------------
DEgenesSainz=read.xlsx("DEgenesSainz.xlsx")### LIST FROM BENE PAPER
head(DEgenesSainz)
dim(DEgenesSainz) #199

read.table("intersect001.txt",dec=",",header=TRUE) #-> dim 341

colnames(DEgenesSainz) <- c("GENE")
head(DEgenesSainz)


library(data.table)
intersect001Sanders<-fintersect(setDT(intersect001), setDT(DEgenesSainz))
intersect001Sanders
GENE
1:  CSMD1
2: ABCC13
```

```
3: RIMBP2
4:   TNS1
5:  SGIP1

dim(intersect001Sanders) #5
write.table(intersect001Sanders,"intersect001Sanders.txt")
read.table("intersect001Sanders.txt", header=T,sep ="")


# SNP related information

CSMD1_SNP<-subset(genesSNP0.01MERGED, Symbol =="CSMD1" )
write.xlsx(CSMD1_SNP,"CSMD1_SNP.xlsx", sheetName="Sheet1")
ABCC13_SNP<-subset(genesSNP0.01MERGED, Symbol =="ABCC13" )
write.xlsx(ABCC13_SNP,"ABCC13_SNP.xlsx")
RIMBP2_SNP<-subset(genesSNP0.01MERGED, Symbol =="RIMBP2" )
write.xlsx(RIMBP2_SNP,"RIMBP2_SNP.xlsx")
TNS1_SNP<-subset(genesSNP0.01MERGED, Symbol =="TNS1" )
write.xlsx(TNS1_SNP,"TNS1_SNP.xlsx")
SGIP1_SNP<-subset(genesSNP0.01MERGED, Symbol =="SGIP1" )
write.xlsx(SGIP1_SNP,"SGIP1_SNP.xlsx")


# CpG related information

cpgs001=read.xlsx("cpgs001.xlsx",1)
dim(cpgs001) # 5720
head(cpgs001)

CSMD1_CpG<-subset(cpgs001, UCSC_RefGene_Name  =="CSMD1" )
write.xlsx(CSMD1_CpG,"CSMD1_CpG.xlsx", sheetName="Sheet2")
ABCC13_CpG<-subset(cpgs001, UCSC_RefGene_Name  =="ABCC13" )
write.xlsx(ABCC13_CpG,"ABCC13_CpG.xlsx", sheetName="Sheet2")
RIMBP2_CpG<-subset(cpgs001, UCSC_RefGene_Name =="RIMBP2" )
write.xlsx(RIMBP2_CpG,"RIMBP2_CpG.xlsx")
TNS1_CpG<-subset(cpgs001, UCSC_RefGene_Name =="TNS1" )
write.xlsx(TNS1_CpG,"TNS1_CpG.xlsx")
SGIP1_CpG<-subset(cpgs001, UCSC_RefGene_Name =="SGIP1" )
write.xlsx(SGIP1_CpG,"SGIP1_CpG.xlsx")

# Venn diagram SNPs, cpgs and genes (Sanders et al, 2017)
-------------------------------------------------------
intersect001Sanders<-fintersect(setDT(intersect001), setDT(DEgenesSanders))
intersect001Sanders
dim(intersect001Sanders) #16

dim(ucscnameALLgenesNEXUS)# 1586
dim(CPGsVENNcleared0.01) #3791
dim(DEgenesSanders) #1058
```

```
# venn 3 (snp, cpgs and genes)
grid.newpage()
draw.triple.venn(area1 = 1586, area2 = 3791, area3 = 1058,
n12 = 341, n23 = 176, n13 = 58, cat.cex = rep(1, 1),
n123 = 16, category = c("Significant␣SNPs␣(P<0.01)",
"Significant␣CpGs␣(P<0.01)", "Diferentially␣expressed␣genes␣(P<0.05)"),
 lty = "blank",fill = c("skyblue", "pink1", "mediumorchid"))

 # Venn diagram SNPs, cpgs and genes (Sainz et al, 2013)
 ----------------------------------------------------------
 dim(ucscnameALLgenesNEXUS) #1586
 dim(CPGsVENNcleared0.01) #3791

 #Intersect SNPs and GENES
 SNPandGenes<-fintersect(setDT(ucscnameALLgenesNEXUS),
 setDT(DEgenesSainz))
 dim(SNPandGenes) #12

 #Intersect CpGs and genes
 CPGandGenes<-fintersect(setDT(CPGsVENNcleared0.01),
 setDT(DEgenesSainz))
 dim(CPGandGenes) #35

 #Intersect SNPs and cpgS

 dim(intersect0.01) #341

 # venn 3 (snp, cpgs and genes)
 library(VennDiagram)
 grid.newpage()
 draw.triple.venn(area1 = 1586, area2 = 3791, area3 = 199,
 n12 = 341, n23 = 35, n13 = 12, cat.cex = rep(1, 1),
 n123 = 5, category = c("Significant␣SNPs␣(P<0.01)",
 "Significant␣CpGs␣(P<0.01)",
 "Diferentially␣expressed␣genes␣(P<0.05)"),
 lty = "blank",fill = c("skyblue", "pink1", "mediumorchid"))
```

**E. Supplementary tables**

Summary data for the 16 differentially expressed genes

**Supplementary table.** Summary data for the 16 differentially expressed genes.

| ***ANK1*** | | | |
|---|---|---|---|
| ***SNP*** | ***Position*** | ***Beta coefficient*** | ***P-value*** |
| rs10958698 | 41641075 | -1,237 | 0,001 |
| rs12156072 | 41640657 | -1,237 | 0,001 |
| rs13272350 | 41661716 | -1,321 | 0,006 |
| rs13273224 | 41641624 | -1,237 | 0,001 |
| rs9650332 | 41640125 | -1,237 | 0,001 |
| ***CpG*** | ***Position*** | ***Beta coefficient*** | ***P-value*** |
| cg12292531 | 41685719 | 16,472 | 0,008 |
| cg19844326 | 41755409 | 7,078 | 0,006 |
| cg22845790 | 41694003 | 4,202 | 0,006 |
| ***ensGene*** | ***Fold Change*** | ***Beta coefficient*** | ***Bonferroni*** |
| ENSG00000029534.16 | 0.154 | 0.265 | 8.60E-04 |

| ***CLNK*** | | | |
|---|---|---|---|
| ***SNP*** | ***Position*** | ***Beta coefficient*** | ***P-value*** |
| rs13112750 | 10547343 | -2,095 | 0,009 |
| ***CpG*** | ***Position*** | ***Beta coefficient*** | ***P-value*** |
| cg03254067 | 10591571 | 10,036 | 0,005 |
| cg06628679 | 10609880 | 3,450 | 0,002 |
| ***ensGene*** | ***Fold Change*** | ***Beta coefficient*** | ***Bonferroni*** |
| ENSG00000109684.11 | 0.122 | 0,190 | 4.09E-06 |

| ***COL24A1*** | | | |
|---|---|---|---|
| ***SNP*** | ***Position*** | ***BETA*** | ***P-value*** |
| rs11589722 | 86352664 | -2 | 0,005 |
| rs12755267 | 86350554 | -2 | 0,005 |
| rs1360903 | 86572463 | -1,062 | 0,010 |
| ***CpG*** | ***Position*** | ***Beta coefficient*** | ***P-value*** |
| cg04622601 | 86621584 | 16,542 | 0,005 |
| ***ensGene*** | ***Fold Change*** | ***Beta coefficient*** | ***Bonferroni*** |
| ENSG00000171502.11 | -0.060 | -0.121 | 1.66E-06 |

| ***GJA3*** | | | |
|---|---|---|---|
| ***SNP*** | ***Position*** | ***BETA*** | ***P-value*** |
| rs1886176 | 20715801 | -2.19 | 0.005 |
| rs9509058 | 20724285 | 2.19 | 0.005 |
| ***CpG*** | ***Position*** | ***Beta coefficient*** | ***P-value*** |
| cg26133081 | 20736342 | -3163,000 | 0.008 |
| ***ensGene*** | ***Fold Change*** | ***Beta coefficient*** | ***Bonferroni*** |
| ENSG00000121743.3 | 0.161 | 0.098 | 7.30E-03 |

| ***GNG7*** | | | |
|---|---|---|---|
| ***SNP*** | ***Position*** | ***BETA*** | ***P-value*** |
| rs917415 | 2514373 | 19,000 | 0.005 |
| ***CpG*** | ***Position*** | ***Beta coefficient*** | ***P-value*** |
| cg04603130 | 2550027 | 1.747 | 0.009 |
| cg08461840 | 2620967 | 25.841 | 0.0039 |

| | | | |
|---|---|---|---|
| cg21340148 | 2702986 | 13.113 | 0.009 |
| cg26863600 | 2616921 | 4.649 | 0.008 |
| ***ensGene*** | ***Fold Change*** | ***Beta coefficient*** | ***Bonferroni*** |
| ENSG00000176533.9 | -0.057 | -0.213 | 1.99E-03 |

***IL5RA***

| ***SNP*** | ***Position*** | ***BETA*** | ***P-value*** |
|---|---|---|---|
| rs334788 | 3153069 | 2.19 | 0.005 |
| rs340831 | 3109444 | 1.561 | 0.005 |
| ***CpG*** | ***Position*** | ***Beta coefficient*** | ***P-value*** |
| cg25381017 | 3151795 | 2237,000 | 0.002 |
| ***ensGene*** | ***Fold Change*** | ***Beta coefficient*** | ***Bonferroni*** |
| ENSG00000091181.16 | -0.212 | -0.058 | 1.26E-02 |

***IL15***

| ***SNP*** | ***Position*** | ***BETA*** | ***P-value*** |
|---|---|---|---|
| rs10519610 | 142637101 | -2 | 0.005 |
| rs12504148 | 142590103 | -2 | 0.005 |
| rs12510514 | 142601496 | -2 | 0.005 |
| ***CpG*** | ***Position*** | ***Beta coefficient*** | ***P-value*** |
| cg25546588 | 142557391 | 20869,000 | 0.000725640950505874 |
| ***ensGene*** | ***Fold Change*** | ***Beta coefficient*** | ***Bonferroni*** |
| ENSG00000164136.13 | 0.053 | 1.52E-02 | 0.149 |

***KDM2B***

| ***SNP*** | ***Position*** | ***BETA*** | ***P-value*** |
|---|---|---|---|
| rs13754 | 121867257 | 2.19 | 0.005 |
| rs7307400 | 121925468 | 2.19 | 0.005 |
| rs7316418 | 121927603 | 2.19 | 0.005 |
| ***CpG*** | ***Position*** | ***Beta coefficient*** | ***P-value*** |
| cg23972735 | 121890311 | 7.308 | 0.005 |
| ***ensGene*** | ***Fold Change*** | ***Beta coefficient*** | ***Bonferroni*** |
| ENSG00000089094.13 | 0.032 | 0.212 | 3.70E-05 |

***NUB1***

| ***SNP*** | ***Position*** | ***BETA*** | ***P-value*** |
|---|---|---|---|
| rs4430016 | 151038255 | 2 | 0.007 |
| ***CpG*** | ***Position*** | ***Beta coefficient*** | ***P-value*** |
| cg02217041 | 151037549 | 9.170 | 0.004 |
| ***ensGene*** | ***Fold Change*** | ***Beta coefficient*** | ***Bonferroni*** |
| ENSG00000013374.12 | 0.041 | 0.336 | 9.03E-05 |

***RBPMS***

| ***SNP*** | ***Position*** | ***BETA*** | ***P-value*** |
|---|---|---|---|
| rs17554116 | 30419334 | 1.388 | 0.008081 |
| rs17554408 | 30427531 | 1.388 | 0.008081 |
| rs2979531 | 30383013 | -1.469 | 0.00341 |
| rs7812836 | 30352258 | -1.321 | 0.006348 |
| ***CpG*** | ***Position*** | ***Beta coefficient*** | ***P-value*** |

| | | | |
|---|---|---|---|
| cg13490635 | 30242021 | 15.030 | 0.007 |
| ***ensGene*** | ***Fold Change*** | ***Beta coefficient*** | ***Bonferroni*** |
| ENSG00000157110.12 | 0.210 | 0.452 | 1.12E-05 |

***SGK1***

| ***SNP*** | ***Position*** | ***BETA*** | ***P-value*** |
|---|---|---|---|
| rs17063576 | 134578920,00 | 2 | 0.005391 |
| ***CpG*** | ***Position*** | ***Beta coefficient*** | ***P-value*** |
| cg03400131 | 134497247 | 3.478 | 0.0005 |
| ***ensGene*** | ***Fold Change*** | ***Beta coefficient*** | ***Bonferroni*** |
| ENSG00000118515.8 | -0.089 | -0.213 | 1.55E-02 |

***SHANK2***

| ***SNP*** | ***Position*** | ***BETA*** | ***P-value*** |
|---|---|---|---|
| rs4245462 | 70881929 | -2.095 | 0.009364 |
| rs4304805 | 70885169 | -2.095 | 0.009364 |
| rs4340077 | 70890503 | -2.095 | 0.009364 |
| ***CpG*** | ***Position*** | ***Beta coefficient*** | ***P-value*** |
| cg09902254 | 70858237 | 26.091 | 0.002 |
| ***ensGene*** | ***Fold Change*** | ***Beta coefficient*** | ***Bonferroni*** |
| ENSG00000162105.13 | 0.172 | 0.165 | 1.62E-04 |

***TCN2***

| ***SNP*** | ***Position*** | ***BETA*** | ***P-value*** |
|---|---|---|---|
| rs4820888 | 31017322 | 2 | 0.005391 |
| rs5749135 | 31011906 | 2 | 0.005391 |
| ***CpG*** | ***Position*** | ***Beta coefficient*** | ***P-value*** |
| cg00788739 | 31002942 | 17.530 | 0.003 |
| cg17693957 | 31002757 | 12.618 | 0.008 |
| cg22542751 | 31002892 | 12.167 | 0.009 |
| ***ensGene*** | ***Fold Change*** | ***Beta coefficient*** | ***Bonferroni*** |
| ENSG00000185339.5 | 0.090 | 0.296 | 1.82E-02 |

***WDFY4***

| ***SNP*** | ***Position*** | ***BETA*** | ***P-value*** |
|---|---|---|---|
| rs2448544 | 50008287 | 2.19 | 0.005 |
| rs2928391 | 49997272 | 2.19 | 0.005 |
| rs7895907 | 50177420 | 2 | 0.007 |
| ***CpG*** | ***Position*** | ***Beta coefficient*** | ***P-value*** |
| cg04749316 | 49893346 | 2.903 | 0.003 |
| cg15164194 | 49892930 | 2.812 | 0.008 |
| cg17967780 | 49893445 | 3.546 | 0.009 |
| cg20504007 | 49892954 | 4.89 | 0.004 |
| cg26246740 | 49893026 | 2.763 | 0.005 |
| cg27459529 | 49892943 | 3.039 | 0.007 |
| ***ensGene*** | ***Fold Change*** | ***Beta coefficient*** | ***Bonferroni*** |
| ENSG00000128815.14 | 0.065 | 0.320 | 2.88E-13 |

***WDR37***

| *SNP* | *Position* | *BETA* | *P-value* |
|---|---|---|---|
| rs12359250 | 1096538 | -1.062 | 0.009 |
| ***CpG*** | ***Position*** | ***Beta coefficient*** | ***P-value*** |
| cg17833322 | 1102835 | -22.682 | 0.003 |
| ***ensGene*** | ***Fold Change*** | ***Beta coefficient*** | ***Bonferroni*** |
| ENSG00000047056.11 | -0.020 | -0.047 | 1.80E-03 |

***ZBTB38***

| *SNP* | *Position* | *BETA* | *P-value* |
|---|---|---|---|
| rs7612543 | 141158212 | 1.388 | 0.008 |
| rs9846396 | 141140968 | 1.388 | 0.008 |
| ***CpG*** | ***Position*** | ***Beta coefficient*** | ***P-value*** |
| cg17495555 | 141042988 | 2.168 | 0.005 |
| ***ensGene*** | ***Fold Change*** | ***Beta coefficient*** | ***Bonferroni*** |
| ENSG00000177311.7 | -0.080 | -0.763 | 2.26E-05 |