Universidad Zaragoza
1542

Máster Universitario en Modelización e Investigación Matemática, Estadística y Computación 2017/2018

Trabajo Fin de Máster

Realización de un estudio de asociación genómica en el repositorio público ADNI

Fátima Belén Marín Fernández

Tutoras
Mónica Hernández Giménez
Elvira Mayordomo

28 de Noviembre de 2018, Zaragoza

Universidad Zaragoza

Máster Universitario en Modelización e Investigación Matemática, Estadística y Computación 2017/2018

Master Thesis

Genome-Wide Association Study Identifying Susceptibility Genes Related with Alzheimer's Disease in ADNI Database

Fátima Belén Marín Fernández

Supervisors

Mónica Hernández Giménez

Elvira Mayordomo

November 28th, 2018. Zaragoza

Contents

List of Figures

Index of Tables

Chapter 1

Abstracts

1.1 Abstract

Genome Wide Association Studies (GWAS) of Alzheimer's disease with quantitative phenotype is an active field of research that has pointed out a number of SNPs potentially related to this disease. Some initiatives such as Alzheimer's Disease Neuroimaging Initiative (ADNI) have been established in order to prove whether anatomical or biological markers (from magnetic resonance imaging (MRI) or positron emission tomography (PET)), genetic information, and clinical and neuropsychological assessments can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). In the last decade, ADNI database has been progressively augmented, including a considerable number of patients yielding the subsequent projects and cohorts ADNI1, ADNIGO, and ADNI2.

Although some GWAS have been carried out with subpopulations of ADNI1, to the best of our knowledge, these works have not been replicated with the whole ADNI1 or the subsequent cohorts. In this thesis, is it intended to study which SNPs are consistently preserved through ADNI cohorts in GWAS, to give support to the reproductivity of the previous study on ADNI (*Hippocampal Atrophy as a Quantitative Trait in a Genome-Wide Association Study Identifying Novel Susceptibility Genes for Alzheimer's Disease*, Potkin et al. 2009), and to discover new risk genes for AD.

This objective has been assessed by performing several GWAS on ADNI1 and ADNI2 cohorts separately, on the merged population of ADNI1 and ADNI2, and also on randomized sub-populations. The results show a positive association for well-known SNPs related to AD: APOE *rs429358* and TOMM40 *rs2075650* SNPs show association with the hippocampal volume in the ADNI1 cohort and in the union of ADNI1 and ADNI2 populations. However, the association for APOE and TOMM40 was not reported in ADNI2. It can be hypothesized that it is the variability of association based on sample size the reason behind that result, and this hypothesis is supported taking into account the results obtained in the randomization of the 50% of the population, but future work is needed to be able to confirm it. In addition, the study reported a weak association of SNPs that are known to be associated to brain-related disorders.

1.2 Resumen

Los estudios de asociación del genoma completo (GWAS) pertenecen a un campo en desarollo muy explotado en los últimos años. En concreto, este método ha intentado asociar la enfermedad de ALzheimer con su base genética, especialmente con ciertos SNPs, mediante el uso fenotipos cuantitativos como el volumen de estructuras cerebrales (que se sabe disminuyen con el progreso de esta infermedad). Algunas iniciativas como ADNI (*Alzheimer's Disease Neuroimaging Initiative*) se establecieron para probar si marcadores anatómicos o biológicos (desde Imágenes de Resonancia Magnética (MRI) hasta Tomografías por emisión de positrones (PET)), la información genética, clínica o análisis neuropsicológicos se pueden combinar para medir la progresión del Alzheimer. En la última década, la base de datos de ADNI ha aumentado considerablemente, dando lugar a las cohortes ADNI1, ADNIGO y ADNI2.

Aunque se han llevado a cabo algunos GWAS con subpoblaciones de ADNI1, que sepamos no se han realizado con todo el proyecto de ADNI1 ni el resto de cohortes. En esta tesis se trata de estudiar qué SNPs están preservados de forma consistente en los diferentes GWAS de las cohortes, de forma que se haga posible la reproducción de los resultados previos del estudio en ADNI *Hippocampal Atrophy as a Quantitative Trait in a Genome-Wide Association Study Identifying Novel Susceptibility Genes for Alzheimer's Disease* (Potkin et al. 2009). También tiene por objetivo descubrir nuevos posibles genes que representen factores de riesgo de padecer Alzheimer.

Estos objetivos han sido abordados mediante la realización de GWAS para las cohortes ADNI1 y ADNI2 por separado, para la población conjunta de ADNI1 y ADNI2, así como para subpoblaciones randomizadas. Los resultados muestran en varios SNPs una asociación positiva con el volumen de hipocampo en el análisis de ADNI1 y en la población total. Se trata de dos SNPs ya relacionados con ALzheimer: *rs429358* del gen APOE y *rs2075650* del gen TOMM40. Sin embargo, esta relación no se encontró en ADNI2. Se podría hipotetizar que la razón de este resultado reside en la variabilidad encontrada en la asociación dependiendo del tamaño muestral. Esta hipótesis estaría respaldada teniendo en cuenta los resultados que se obtienen para el 50% de la población randomizada (ADNI2 representa aproximadamente el 50% de la población total). Sin embargo, se necesitaría profundizar en este aspecto para poder confirmar la hipótesis. Además, en este estudio se encontró una leve asociación en dos SNPs ya relacionados con otros transtornos mentales.

Chapter 2

Introduction

The increase of life expectancy and the growing number of elderly people is causing dementia to emerge as a major health problem [1]. Among them, Alzheimer's disease (AD), a degenerative brain disease, stands as the most common expression of dementia [2]. The characteristic symptoms of dementia are difficulties with memory, language, problem-solving and other cognitive skills that affect a person's daily life. These difficulties derive from profound functional and structural changes observed in neurons, their processes and synapses, and the microgliosis and astrocytosis which accompany these changes [3].

In particular, in Alzheimer's disease, neurons in other parts of the brain are eventually damaged or destroyed as well, including those that enable a person to carry out basic bodily functions such as walking and swallowing. People in the final stages of the disease are bed-bound and require around-the-clock care. Alzheimer's disease is ultimately fatal [1] and together with unintentional injuries, homicide, chronic lower respiratory diseases and suicide, was an important cause of the decrease in life expectancy at birth for the total population of USA in 2015, where the 7 leading causes of death were [4]:

1. Heart disease.
2. Malignant neoplasms (cancer).
3. Chronic lower respiratory diseases.
4. Accidents.
5. Cerebrovascular disease.
6. Alzheimer's disease.
7. Diabetes mellitus.

Even though it is difficult to stipulate whether a death is due to Alzheimer's, according to data from the National Center for Health Statistics of the Centers for Disease Control and Prevention (CDC), 93.541 people died from Alzheimer's disease in 2014 [5]. The CDC considers a person to have died from Alzheimer's if the death certificate lists Alzheimer's as the underlying cause of death, defined by

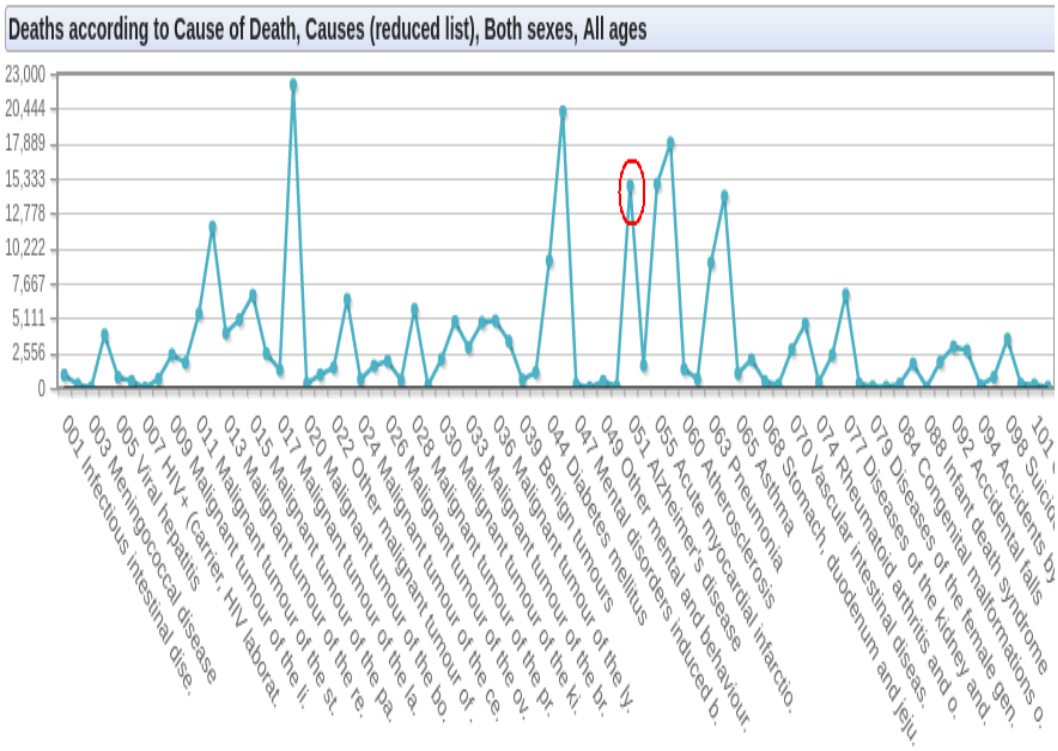


Figure 2.1: Causes of death, 2016: Number 051 represents Alzheimer's disease (Circled in red). Image generated from the Spanish National Institute of Statistics

the World Health Organization as "the disease or injury which initiated the train of events leading directly to death [6]" .

In Spain, Alzheimer's also represents one of the most common causes of death (See Figure 2.1).

These already remarkable numbers are not dropping but increasing: between 2017 and 2025 the number of people with Alzheimer's in USA is going to rise at least a 14% [1] and the annual number of new cases of Alzheimer's and other dementias is projected to double by 2050 [7]. This means that, while today every 66 seconds someone in the United States develops Alzheimer's dementia, by 2050 a person in the United States will develop Alzheimer's dementia every 33 seconds [1].

This alarming situation is making Alzheimer's disease be the focus of a large number of researches. However, there is no available medication by the date for stopping or slowing down the damage and destruction of the neural system caused by this disease [1]. Thus, it is interesting to know which groups are under special risk of developing the disease, for example, in order to consider certain group characteristics or similarities as risks factors susceptible for treatment or having a better understanding of the disease.

2.1 Risk Factors

The most common risk factors are:

- *Age*

  Age is the greatest of the risk factors mentioned here, with the majority of people with Alzheimer's dementia being age 65 or older. The percentage of people with Alzheimer's increases dramatically with age: 3% of people age between 65 and 74, 17% of people age 75-84, and 32% of people age 85 or older have Alzheimer's dementia [9]. It is important to note that Alzheimer's dementia is not a normal part of aging process, and older age alone is not sufficient to cause Alzheimer's dementia [1].

- *Gender*

  As compared to men, 70% of the patients are females, and the reason is not clear: while some believe that could be due to the higher life expectancy of women [10], others think that could lay on biological or genetic variations or differences in life experiences [11].

- *Down syndrome*

  According to the National Down Syndrome Society, about 30% of people with Down syndrome at the age of 50 have Alzheimer's dementia. Around 50% of people with Down syndrome will develop Alzheimer's dementia as they age [1].

- *APOE ϵ4 gene*

  The best known genetic risk factor for Alzheimer's dementia is apolipoprotein E (APOE) ϵ4 gene form. The APOE gene encodes a protein that transports cholesterol in the bloodstream. There are three main forms of the APOE gene: ϵ2, ϵ3 and ϵ4, and everyone inherits one of the three forms from each parent. As a result, everyone has two copies of the gene, and depending on the heritage, this pair could be any combination of two from the three forms mentioned above [1]. The ϵ3 form is the most common, with between 50% and 90% of individuals having one or two copies [12]. The ϵ4 form is the next most common, with 5% to 35% having one or two copies, followed by ϵ2 which is the least common, with up to 5% having one or two copies [12].

alleles	ϵ_2	ϵ_3	ϵ_4
ϵ_2	0.5	11.0	2.0
ϵ_3	11.0	61.0	23.0
ϵ_4	2.0	23.0	2.0

Table 2.1: Distribution of the six combinations of apolipoprotein E (APOE) main alleles [2] ϵ2, ϵ3 and ϵ4 of the U.S. population (in percentages).

The risk of developing Alzheimer's differs depending the number of copies and form of the APOE gene: having the ϵ4 form increases one's risk of

developing Alzheimer's compared with having the $\epsilon 3$ form, while the $\epsilon 2$ form may decrease the risk compared with having the $\epsilon 3$ form. More specifically, those who inherit one copy of the $\epsilon 4$ form have three times the risk of developing Alzheimer's compared with those having the $\epsilon 3$ form. Also, inheriting two copies of the $\epsilon 4$ form can lead an 8- to 12-fold risk of developing the disease [13], [14], [15]. A meta-analysis of 20 publications has described the frequency of the $\epsilon 4$ form among people in the United States who had been diagnosed with Alzheimer's, and it was found that 56% of those had one copy of the APOE $\epsilon 4$ gene, and 11% had two copies of the APOE $\epsilon 4$ gene [16].

Nevertheless, inheriting the APOE $\epsilon 4$ gene does not guarantee that an individual will develop Alzheimer's. This is also true for more than 20 recently identified genes that appear to affect the risk of Alzheimer's. These genes are believed to have a limited effect on the overall prevalence of Alzheimer's because they are rare or only slightly increase risk [17].

In addition, those with the $\epsilon 4$ form are more likely to develop Alzheimer's at a younger age than those with the $\epsilon 2$ or $\epsilon 3$ forms of the APOE gene [18].

Genetically, the apoE main protein isoforms, $\epsilon 2$, $\epsilon 3$ and $\epsilon 4$ are the result of two non-synonymous Single Nucleotide Polymorphisms (SNPs) [3]. SNPs *rs429358* and *rs7412*, located in exon 4 of the APOE gene (See table 2.2). Structural consequences of the exon 4 APOE haplotype appear to be that the apoE $\epsilon 4$ protein binds preferentially to plasma very low density lipids (VLDLs) whereas apoE $\epsilon 2$ and $\epsilon 3$ bind preferentially to plasma high density lipoproteins (HDLs). In addition, apoE isoforms appear to influence plasma cholesterol levels, neuronal growth and amyloid deposition [19].

rs429358	*rs7412*	*Name*
C	T	$\epsilon 1$
T	T	$\epsilon 2$
T	C	$\epsilon 3$
C	C	$\epsilon 4$

Table 2.2: Codification of APOE gene forms $\epsilon 2$, $\epsilon 3$ and $\epsilon 4$

Apart from these risk factors (age, gender, Down Syndrome and genetics), having a family history of Alzheimer's [23] also plays an important role, even though it is not necessary for an individual to develop the disease. However, individuals who have a parent, brother or sister with Alzheimer's are more likely to develop the disease than those who do not have a first-degree relative with Alzheimer's [23]. Those who have more than one first-degree relative with Alzheimer's are at even higher risk. The increased risk associated with having a family history of Alzheimer's is not entirely explained by whether the individual has inherited the APOE $\epsilon 4$ risk gene [1].

[3]Single nucleotide polymorphisms (SNPs) are an abundant form of genome variation, distinguished from rare variations by a requirement for the least abundant allele to have a frequency of 1% or more. [25]

With the exception of cases of Alzheimer's caused by genetic abnormalities, experts believe that Alzheimer's, like other common chronic diseases, develops as a result of multiple factors rather than a single cause [1].

2.2 Biomarkers and SNPs

Together with the groups of risk, diagnosis is crucial. Thus, due to the importance of diagnosis, several investigations have been carried out and nowadays it is possible to distinguish between Alzheimer's and other causes of dementia using certain Biomarkers [4]. These biomarkers are, among other studied factors: changes in brain volume, the amount of beta-amyloid in the brain as shown on positron emission tomography (PET) imaging and levels of certain proteins in fluid: for example, the accumulation of the protein fragment beta-amyloid (beta-amyloid plaques) outside neurons and the accumulation of an abnormal form of the protein tau (tau tangles) inside neurons are brain changes associated with Alzheimer's. Beta-amyloid plaques are believed to contribute to cell death by interfering with neuron-to-neuron communication at synapses, while tau tangles block the transport of nutrients and other essential molecules inside neurons [1].

It would be ideal to find a fast and inexpensive Biomarker-based test, such as a blood test, to diagnose Alzheimer's. Research is underway to develop such a test, but presently there is no test reliable nor accurate enough to diagnose Alzheimer's [1]. In this direction, more research is needed and this thesis aims to help to clarify the Alzheimer's diagnosis via genetic biomarkers, in particular SNPs.

2.3 Genome-wide Association Studies (GWAS)

Association tests between biomarkers and a phenotype (state of an organism resulting from interactions between genes, environment, disease, molecular mechanisms, and chance [26]) of interest are used to identify SNPs related with diseases. There are association studies using the whole genome, called Genome-wide association studies (GWAS) in which the phenotype could be dichotomous (affected, unaffected) or quantitative (biomarker levels, imaging metrics, etc.). This approach has identified susceptibility loci [5] in several diseases. GWAS of Alzheimer's, have confirmed the strong influence of APOE [21] [22], but there is no convincing evidence implicating other genes, despite many biologically plausible and interesting candidates. By design, GWAS' content and meta-analysis results are dynamically changing and reflect the continuing evolution of leading candidate genes for AD and the biological pathways they may represent [20]. The most

[4] A Biomarker is a measurable indicator of some biological state or condition in the human body. Biomarkers are used to diagnose the presence or absence of disease, assess the risk of developing a disease, or understand how a patient has responded to a treatment. [8] For example, the level of glucose is a biomarker for diabetes.

[5] A locus (pl. *loci*) can be defined either as a segment of DNA with alternate nucleotide sequences as alleles, or as a nucleotide site with alternate nucleotides as alleles. [27]

robust findings from case-control GWAS and other types of genetic association studies, can provide targets for examining quantitative phenotypes derived from biomarkers' databases. Between them, Alzheimer's Disease Neuroimaging Initiative (ADNI) plays an important role on GWAS related with Alzheimer's disease [28]. One of the phenotypes used for Alzheimer's is the concentration of grey matter in the hippocampus, because it is affected early in AD (hippocampal atrophy is a well-known feature of Alzheimer's disease [30], see 2.2), is implicated in the conversion to AD, its progress and is associated with many of the main symptoms, and it can be reliably measured in vivo [46].

A big challenge in case-control GWAS designs in which allele and genotype frequencies are compared between AD and control patients is achieving sufficient statistical power. Such categorical approaches require approximately 6000 cases and controls to obtain 85% power to detect a 30% difference (odds-ratio of 1.3) with a minor allele frequency [6] of 0.15. Quantitative trait association studies offer several advantages over case-control studies, since the dependent measures are quantitative and more objective than diagnostic categorization, and can increase the statistical power four to eight fold, thus greatly decreasing the required sample size to achieve sufficient statistical power [28].

In this direction, in 2009 Potkin et al. [46] published the study *Hippocampal Atrophy as a Quantitative Trait in a Genome-Wide Association Study Identifying Novel Susceptibility Genes for Alzheimer's Disease* in which they used hippocampal atrophy as a quantitative phenotype in a GWAS study with data from ADNI database. They identified candidate risk genes for sporadic Alzheimer's disease. The most representative ones are shown in Table 2.3 :

Gene	***SNP***	**p-value**
TOMM40	rs2075650	7.48×10^{-7}
APOE	rs429358	2.30×10^{-6}
EFNA5	rs10074258	2.5×10^{-7}
EFNA5	rs12654281	3.72×10^{-7}
PRUNE2	rs10781380	10×10^{-6}
I509T	rs8115854	10×10^{-6}
RPN2	rs6031882	6.2×10^{-6}
E88K	rs2073145	2.13×10^{-6}
RP11-232A1.1	rs10867752	3.08×10^{-6}
CAND1	rs1082714	4.93×10^{-6}
AL079307.7	rs11626056	1.18×10^{-6}

Table 2.3: Extract of the SNPs related to AD found by *Potkin et al.*

In the study, a subset of ADNI1 cohort was used as source data. In the last decade, ADNI database has been progressively augmented, including a considerable number of patients yielding to the subsequent projects ADNI1, ADNIGO, and ADNI2. To the best of our knowledge, the work of Potkin et al. [46] has not been replicated with the whole ADNI1 or the subsequent databases,

[6] Allele frequency (also called gene frequency) is the term used to describe the fraction of gene copies that are of a particular allele in a defined population. [29]

perhaps because the arise of meta-studies has opened the opportunity to deal directly with augmented cohort sizes [33]. However, it is also important to obtain GWAS results on more homogenous data before attempting more sophisticated meta-analysis.

2.4 Scope of the study

In this work, it is intended to study which SNPs are consistently preserved through ADNI cohorts in GWAS. The association studies have been conducted separately in the different ADNI cohorts and in the whole ADNI database. Also, an exhaustive randomized analysis for the assessment of the consistency of the association results with the sample selection was performed.

In addition, replication of results in independent samples is an important strategy, therefore this thesis aims to give support to the reproducibility of this previous study (*Hippocampal Atrophy as a Quantitative Trait in a Genome-Wide Association Study Identifying Novel Susceptibility Genes for Alzheimer's Disease*, Potkin et al. 2009 [46]), and to discover new risk genes for AD. Consensus criteria for replication have recently been published. They include study of the same or very similar phenotypes and populations, together with the demonstration of a similar magnitude of effect and significance for the same SNPs and alleles described in the initial report. Replication is usually first attempted in studies as similar as possible to the initial report [21] , so that is what is attempted here. At present, the best way of resolving inconsistencies between studies seems to be additional replication studies with larger sample sizes, and in this project more patients than in the original article will be included (from the same Database) [21] .

For the purpose of this study, a GWA - QT (quantitative trait) analysis is performed. The method uses the hippocampus volume from neuroimaging as the quantitative trait (in the original study, hippocampal grey matter density was used), and examines which SNPs (as proxies for genes) influence the quantitative trait differently for AD and healthy controls. The trait used here is the volume of hippocampus in AD subjects vs normal since MRI studies have suggested that reductions in hippocampus over time can be particularly useful in predicting AD before the beginning of clinical symptoms, and in assessing the efficacy of pharmacological treatment in clinical trials [31] [32]. Therefore, in this GWA study hippocampal volume is used as an imaging phenotype to reveal genes that potentially influence hippocampal atrophy and dementia in the context of AD. The genes which influence hippocampal grey matter concentration differentially in AD and healthy subjects may provide important information regarding the mechanisms of disease-related atrophy.

To sum up, a GWAS for SNP association with volume of hippocampus in Alzheimer's disease will be made. It will be performed using different data from ADNI cohorts and randomizations. Previously the database (DDBB) will be analyzed in order to extract the required information.

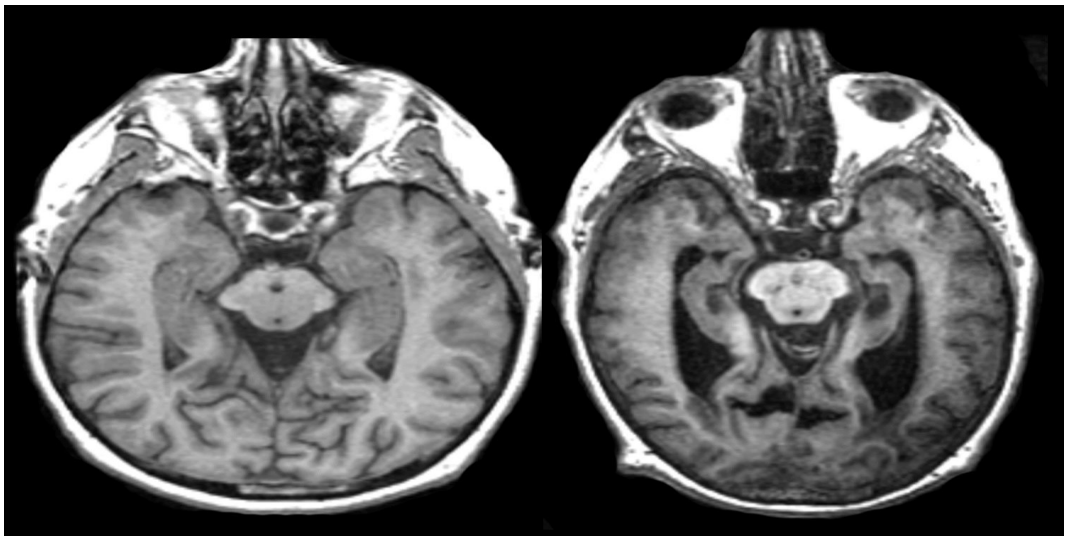


Figure 2.2: Healthy patient brain (left) and brain affected by AD (right). Image from Sunnybrook Health Science Centre

Chapter 3

Materials and Methods

3.1 ADNI Project

Data used in this project has been obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (`http://adni.loni.usc.edu/`). This database is constantly changing and growing. In 2003, the ADNI was launched by the National Institute of Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), non-profit organizations and private pharmaceutical companies. The main goal of the initiative was to test whether some biological markers (such as magnetic resonance imaging (MRI) or positron emission tomography (PET)), clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early AD. Determination of sensitive and specific markers of very early AD progression would help to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials.

ADNI is the result of efforts of many co-investigators from a broad range of academic institutions and private corporations, and the Principal Investigator of this initiative is Michael W. Weiner, M.D., VA Medical Center and University of California (San Francisco). Subjects have been recruited from all over North America, and their number has been increasing from 800 to more than 2000 patients, coming from different phases of the initiative (ADNI1, ADNIGO, ADNI2, ADNI3; See Figure 3.1):

- **ADNI1:** 400 subjects diagnosed with MCI, 200 with the early AD, and 200 elderly control subjects.

- **ADNIGO:** Existing ADNI1 cohort along with 200 new participants with early mild cognitive impairment (EMCI).

- **ADNI2:** Participants from the ADNI1/ADNI GO phases in addition to the following new participant groups: 150 elderly controls, 100 EMCI subjects, 150 late mild cognitive impairment (LMCI) subjects, and 150 mild AD patients. A new cohort, Significant Memory Concern (SMC), was also added in ADNI2 to address the gap between healthy controls and MCI; a key inclusion

criteria is a self-reported significant memory concern from the participant. The study included 107 SMC subjects.

- **ADNI3:** Subjects from previous phases and 135-500 Normal Controls, 150-515 MCI, 85-185 AD.

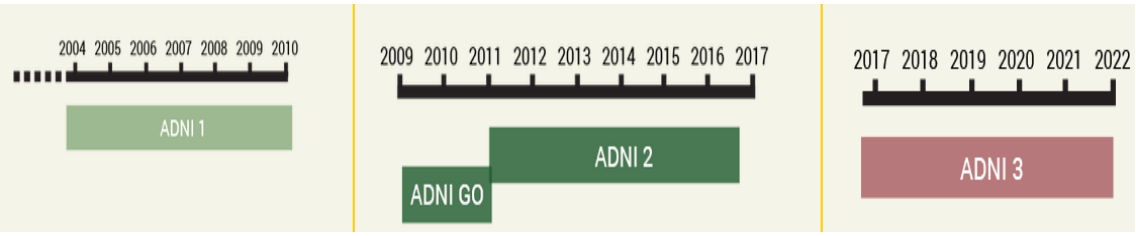


Figure 3.1: Temporal line of ADNI phases

The diagnosis of the patients is made according to the results they obtain in some tests (Mini- Mental State Exam (MMSE), Clinical Dementia Rating (CDR), Alzheimer's Disease Assessment Scale-Cognitive (ADAS)) and medical observations. For normal subjects: MMSE scores between 24-30 (inclusive), a CDR of 0, non-depressed, non MCI, and nondemented; MCI subjects: MMSE scores between 24-30 (inclusive), a memory complaint, have objective memory loss, a CDR of 0.5, absence of significant levels of impairment in other cognitive domains, essentially preserved activities of daily living, and an absence of dementia; Mild AD: MMSE scores between 20-26 (inclusive), CDR of 0.5 or 1.0, and meets NINCDS/ADRDA (Nacional Institute of Neurological and Communicative Disordes and Stroke-Alzheimer's Disease and Related Disordes Association) criteria for probable AD.

More up-to-date information can be found at `http://adni.loni.usc.edu/about/`.

3.2 Participants

All subjects studied here were part of different cohorts of ADNI. The initiative enrolls participants between the ages of 55 and 90 who are recruited at 57 sites in the United States and Canada. After obtaining informed consent, participants undergo a series of initial tests that are repeated at intervals over subsequent years, including a clinical evaluation, neuropsychological tests, genetic testing, lumbar puncture, and MRI and PET scans. In the original article, 172 AD subjects and 209 healthy controls were included.

Some GWAS studies begin with small numbers of participants in the initial scan but carry forward large numbers of SNPs to minimize false-negative results. Other studies begin with more participants but carry forward a smaller proportion of associated SNPs. Optimal proportions of study participants and SNPs have

yet to be determined, but carrying forward a small proportion ($\leq 5\%$) of SNPs will often mean limiting the associations ultimately identified to those having a relatively large effect [22] In this project, a total of 1492 subjects were included in the analysis (757 from ADNI1 and 735 from ADNI2, 1111 patients in total more than in the original article) with nearly 621.000 SNPs (516.645 in the original article).

In order to compare proportions of a categorical outcome according to AD and CN groups, several statistical tests such as chi-squared test and Fisher's exact test are considered. The chi-squared test and Fisher's exact test can assess for independence between two variables when the compared groups are independent and not correlated. The chi-squared test applies an approximation assuming the sample is large, while the Fisher's exact test runs an exact procedure especially for small-sized samples [51]. Tables 3.1 ,3.2 , 3.3 represent the demographic characteristics of the cases and controls of the non randomized samples analyzed for independence (cohorts ADNI1, ADNI2 and ADNI1 merged with ADNI2). Also, a median comparison (when possible) is performed. In all comparisons, healthy control group had significantly more education than people with Alzheimer's, and also they scored high in MMSE (a cognitive test), as expected.

Furthermore, no statistical difference in the distribution of APOE $\epsilon 4$ alleles between ADNI1 and ADNI2 cohorts was found (Fisher's Exact Test for Count Data, p-value = 0.7808).

Category	*Control*	AD	p-value
Number of Subjects	214	175	
Gender (Male/Female)	115/99	93/82	0.410 Chi Square (Yates Correction)
Age	75.67 ± 4.9	75.4 ± 7.4	$\geq$ 0.6743 Welch Two Sample t-test
MMSE	26.1 ± 0.99	23.27 ± 2.05	<2.2e-16 Welch Two Sample t-test
Years of Education	16.07 ± 2.78	14.6 ± 3.16	<2.586e-06 Welch Two Sample t-test
Ethnicity (Hisp-Latino/Not Hisp-Latino/Unknown)	2/211/1	4/169/2	0.6302 Fisher's exact
Race(Am Indian-Alaskan/Asian/Black/More than one/White)	0/2/15/0/197	0/2/8/2/163	0.4881 Fisher's exact
Copies of e4(0/1/2)	156/53/5	58/85/32	<2.2e-16 Fisher's exact

Table 3.1: Baseline characteristics for both CN and AD groups in ADNI 1

Category	*Control*	AD	p-value
Number of Subjects	149	110	
Gender (Male/Female)	77/72	65/45	0.2898 Chi Square (Yates correction)
Age	73.84 ± 5.96	74.59 ± 8.57	$\geq$ 0.432 Welch Two Sample t-test
MMSE	29 ± 1.3	23 ± 2.1	<2.2e-16 Welch Two Sample t-test
Years of Education	16.48 ± 2.50	15.75 ± 2.61	<0.02488 Welch Two Sample t-test
Ethnicity (Hisp-Latino/Not Hisp-Latino/Unknown)	7/141/1	2/107/1	0.5264 Fisher's exact
Race(Am Indian-Alaskan/Asian/Black/More than one/White)	1/5/9/0/134	0/3/3/0/104	0.4972 Fisher's exact
Copies of e4(0/1/2)	112/33/4	37/52/21	<8.504e-12 Fisher's exact

Table 3.2: Baseline characteristics for both CN and AD groups in ADNI 2

Category	*Control*	AD	p-value
Number of Subjects	363	285	
Gender (Male/Female)	192/171	158/127	0.5713 Chi Square (Yates correction)
Age	74.92 ± 5.43	75.09 ± 8.87	$\geq$ 0.7623 Welch Two Sample t-test
MMSE	29 ± 1.12	23 ± 2.07	<2.2e-16 Welch Two Sample t-test
Years of Education	16.24 ± 2.68	15.05 ± 3.01	<2.31e-07 Welch Two Sample t-test
Ethnicity (Hisp-Latino/Not Hisp-Latino/Unknown)	9/352/3	6/276/3	0.353 Fisher's exact
Race(Am Indian-Alaskan/Asian/Black/More than one/White)	1/7/24/0/331	0/5/11/2/267	0.7563 Fisher's exact
Copies of e4(0/1/2)	268/86/9	95/137/53	<2.2e-16 Fisher's exact

Table 3.3: Baseline characteristics for both CN and AD groups in ADNI 1 and 2

3.3 Genotyping

In order to obtain the genetic information, the DNA of patients was sequenced. Nevertheless, different methods of genotyping were used: ADNI1 samples were genotyped using the Illumina Human610-Quad BeadChip and intensity data processed with GenomeStudio v2009.1 and ADNIGO/2 samples were genotyped using the Illumina HumanOmniExpress BeadChip and intensity data processed with GenomeStudio v2009.1. This leads to a difference in the presence of SNPs genotyped between phases (for ADNI1 620.901 SNPs and for ADNIGO/2 730.525 SNPs), that must be erased for the analysis of the data. Therefore, the intersection of all SNPs will be made and taken into account in analysing the data.

Moreover, none of the protocols had SNPs for APOE between the SNPs genotyped. Due to its major importance, ADNI decided to make an additional analysis and genotyped it. Therefore, the information about APOE is not included in the first analysis, but in one unique file for all subjects. Also, the codification of the APOE SNPs (See figure 3.2) is not the same as the analysis with the rest of genetic information: APOE SNPs are represented by a combination of 2, 3 and 4 (not the bases A, C, T, G). Those numbers represent $\epsilon 2$, $\epsilon 3$ and $\epsilon 4$ and need to be coded with A, C, T, G. That is made with a JAVA program following table 2.2.

Phase	ID	RID	SITEID	VISCODE	USERDATE	USERDATE2	APTESTDT	APGEN1	APGEN2	APVOLUME
ADNI1	4	2	107	sc	2005-08-23		08/22/05	3	3	5
ADNI1	6	3	107	sc	2005-08-23		08/22/05	3	4	10
ADNI1	8	4	10	sc	2005-08-23		08/22/05	3	3	9.2
ADNI1	10	5	107	sc	2005-08-29		08/29/05	3	3	10
ADNI1	12	7	10	sc	2005-09-06		09/05/05	3	4	9
ADNI1	14	9	10	f	2005-09-06		09/05/05	3	3	5.5
ADNI1	16	8	107	sc	2005-09-06		09/05/05	2	3	10
ADNI1	18	11	107	f	2005-09-21		09/19/05	3	4	7.8
ADNI1	20	14	10	sc	2005-09-21		09/19/05	3	3	7.2
ADNI1	22	16	107	sc	2005-09-21		09/19/05	3	4	9.5
ADNI1	24	18	10	f	2005-09-21		09/19/05	3	3	1
ADNI1	26	15	4	sc	2005-09-28		09/26/05	3	4	10
ADNI1	28	20	32	f	2005-09-28		09/26/05	4	4	10
ADNI1	30	24	32	f	2005-09-28		09/26/05	3	3	9

Figure 3.2: APOE information from genotyped samples

3.4 ADNI Database and Data Processing

In order to carry out the analysis of the data, PLINK (the software used for that purpose) needs several files with a especial format. The first step is to find the information of the patients in the database: genotype (list of SNPs), phenotype (hippocampus volume) and other information such as gender, age, Family ID, Paternal ID and Maternal ID. This is a tough step, because of the dimension of the database and the varied origin of data. Information is only found for subjects from ADNI1, ADNIGO and ADNI2, so the last phase (ADNI3) will not be used here.

3.4.1 DDBB Organization

Information about ADNI initiative is stored in a repository from California University `http://adni.loni.usc.edu/`. Here, all the mentioned data can be reached and the information is divided in three different sections: Image Collections, Genetic Data and Study Data.

On Image Collections there is a IDA searcher where one can find all information about the images collected from patients (MRI, PET...). Regarding the genetic data , ADNI1/GO/2 samples were whole-genome sequenced (WGS): There are two releases of the ADNI array data, and the second one provides more information, so it will be the one used here. This data comes on a .zip file containing all patients, a .csv file per individual, so it is not possible to look for a patient in particular, the whole data set must be download. Also, each project has its own .zip file containing its patients.

In genetic data section more information can be found: gene expression, DNA methylation profile and data prepared for challenges.

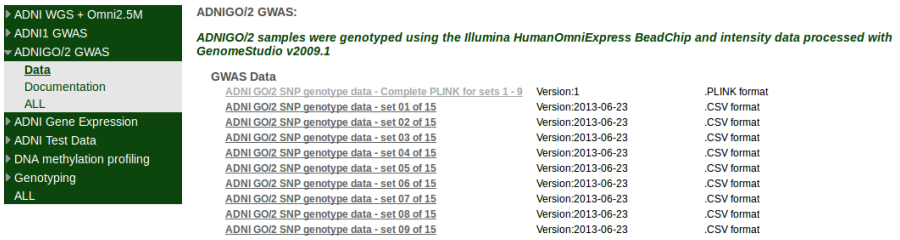


Figure 3.3: Organization of Genetic Data in ADNI

In what concerns to Study Data, it is divided into (See Figure 3.4) :

- Assessments: It contains files about diagnosis and Neuropsychological data, such as results of tests before mentioned for diagnosing (MMSE, ADAS).

- Biospecimen: One of the goals of ADNI is the collection of biospecimens, including blood, urine, and cerebrospinal fluid (CSF), blood for APOE and Cell Immortalization from participants.

- Enrollement: Information about visits, non-clinical tracing of patients.

- Genetic: Genotype results for TOMM40 gene and information about ADNI genetic data.

- Images: For MRI and PET images, information related to analysis and quality of them.

- Medical History: List of adverse events, drugs consumed, medical history and physical/neurological exams of the patients.

- Neuropathology: Neuropathology exams results.

- Study Info: Information about different studies and standardization of data presentation. It also contains packages for programs used in the studies.

- Subject Characteristics: Family History and demographic information about the subjects.

- Test Data: Other data for challenges, but it contains no images.

3.4.2 Information Collection

In order find the desired data, these files are taken from ADNI DDBB:

- *DATADICT.csv*: Description of the columns in other files. It is placed in *"Study Info"*.

- *APOERES.csv*: This file includes the APOE gen sequenced. It is located in *"Biospecimen"*.

- WGS: Genetic data for ADNI1 and ADNI2 patients, it is found in *"Genetic Data"* and consists of a .csv file per patient in which all the SNPs are listed and their sequence given.

- *ROSTER.csv*: This file contains information about patients identification (Individual ID and related). It can be found on *"Study Data"* - *"Enrollment"*.

- *PTDEMOG.csv*: This includes demographic data and other information such as gender, date of birth, profession... It can be located in *"Subject Chracteristics"*.

- *ADNIMERGE.csv*: It contains some key variables merged into one data table and also information about the phenotype studied here (hippocampus volume). It is placed in *"Study Info"*.

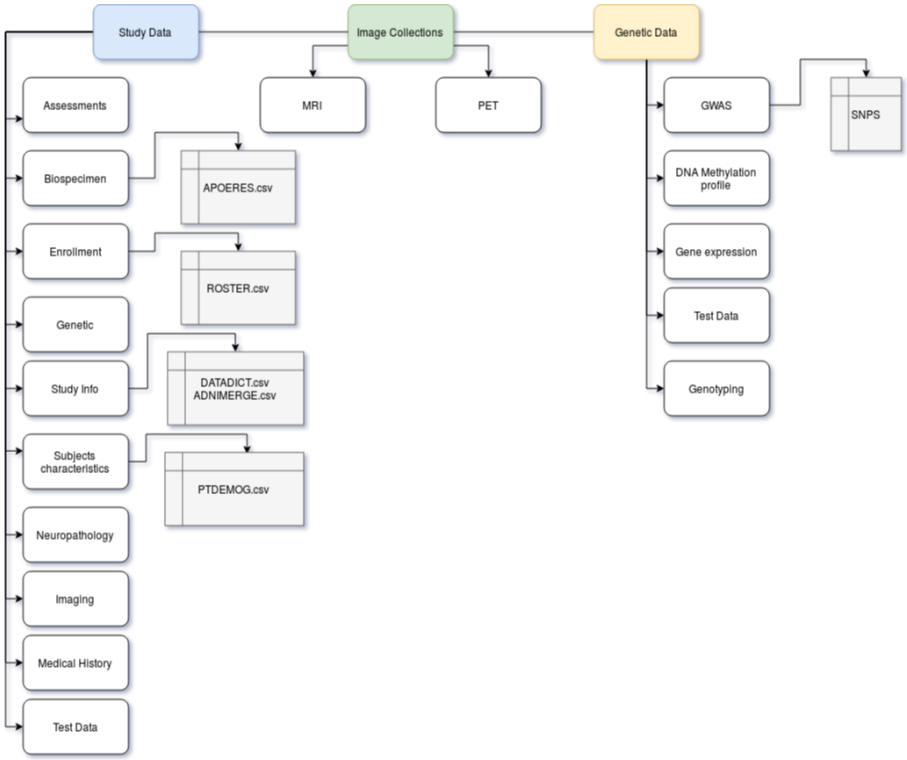


Figure 3.4: ADNI DDBB

Once all the information is located, it is time to convert the files into the format required by the software, detailed below.

3.5 PLINK

The software used for the analysis is PLINK v1.07 (`http://pngu.mgh.harvard.edu/purcell/plink/`) [34]. This software is being developed by Shaun Purcell at the Center for Human Genetic Research (CHGR), Massachusetts General Hospital (MGH), and the Broad Institute of Harvard and MIT, with the support of others. Is a free, open-source whole genome association analysis toolset, designed to perform a range of basic, large-scale analyses in a computationally efficient manner and its focus is purely on analysis of genotype/phenotype data, so previous steps (study design and planning, generating genotype or CNV calls from raw data) have no support on PLINK.

These files are required by PLINK in order to do the analysis, and have been generated in collaboration with the group using JAVA, unix shell and R:

MAP This file is a list with all the SNPs wanted for the analysis. All patients must have information about the same SNPs, so the intersection of all patients' lists is made and is the one used from now on. The colums are:

Chromosome [1] : Chromosome number

SNP ID : SNP consensus name

SNP position : In centimorgans [2]

SNP position : In base pairs (bp).

```
1       rs3094315       0       742429
1       rs12562034      0       758311
1       rs12124819      0       766409
1       rs4475691       0       836671
1       rs3748597       0       878522
1       rs13303118      0       908247
1       rs9777703       0       918699
1       rs3934834       0       995669
1       rs9442372       0       1008567
1       rs3737728       0       1011278
1       rs6687776       0       1020428
```

Figure 3.5: Example of a MAP file

FAM FAM files have the relevant information of the patients. The columns are:

Family ID : There is no Family ID found in data, so it is a counter.

[1] Chromosomes are the most prominent structures, and most genetic processes occur in chromosomes including transcription, DNA replication and repair, and repression of gene expression, which are modulated by interactions with gene-regulating proteins. [36]

[2] A centimorgan is defined as the distance between chromosome positions for which the expected average number of intervening chromosomal crossovers in a single generation is 0.01. [37]

Individual ID

Paternal ID : Set to 0 if patient has no father on the database.

Maternal ID : Set to 0 if patient has no mother on the database.

Sex : 1 = "male", 2 ="female", 0 = "unknown".

Phenotype : 1 ="case", 2 ="control", -9 = "unknown". Here it is set as "unknown" because the information will be upload from the *pheno* file.

```
1 014_S_0520 0 0 2 -9
2 005_S_1341 0 0 2 -9
3 012_S_1175 0 0 1 -9
4 012_S_0803 0 0 2 -9
5 018_S_0055 0 0 1 -9
6 027_S_0118 0 0 1 -9
7 027_S_0403 0 0 1 -9
8 053_S_0389 0 0 2 -9
9 041_S_0262 0 0 1 -9
10 005_S_1224 0 0 1 -9
```

Figure 3.6: Example of a FAM file

LGEN LGEN files contain the genotype (alleles) of each SNP for all patients (one SNP per row, for patient). The columns are:

Family ID :The same Family ID used in other files.

Individual ID: The same Individual ID used in other files.

SNP ID: The same SNP ID used in other files.

Allele 1: A = Adenine, T = Timine, C = Citosine, G = Guanine.

Allele 2 :A = Adenine, T = Timine, C = Citosine, G = Guanine.

```
495 062_S_1182 200003 A G
495 062_S_1182 200006 T C
495 062_S_1182 200047 A G
495 062_S_1182 200050 G C
495 062_S_1182 200052 T T
495 062_S_1182 200053 T T
495 062_S_1182 200070 G C
495 062_S_1182 200078 C C
495 062_S_1182 200087 T T
495 062_S_1182 200091 A C
495 062_S_1182 200096 C C
```

Figure 3.7: Example of a LGEN file

PED PED file contains the genotype (alleles) of all SNP in a row, one row per patient. It also contains information about the phenotype, sex and ID's. The columns are:

Family ID :The same Family ID used in other files.

Individual ID :The same Individual ID used in other files.

Paternal ID :The same Paternal ID used in other files.

Maternal ID: The same Maternal ID used in other files.

Sex :The same sex used in other files.

Phenotype: The same phenotype used in other files.

SNPs All the SNPs of the patient in the same row.

```
1 814_S_0520 0 0 2 -9 A C T G A A C ...
1 121_S_8120 0 0 2 -9 A T C G A G C ...
1 130_S_0370 0 0 2 -9 G C C A A G T ...
1 002_S_1966 0 0 2 -9 A C C A A G T ...
```

Figure 3.8: Example of a PED file

BIM BIM file contains the same information and structure as MAP file, but two last columns are added. All columns are:

Chromosome: Chromosome number

SNP Name: SNPs consensus name

SNP Position: In centimorgans

SNP Position: In base pairs (bp).

Minor allele: Second most common allele.

Major allele: Most common allele.

```
1       rs3094315       0       742429  C       T
1       rs12562034      0       758311  A       G
1       rs12124819      0       766409  G       A
1       rs4475691       0       836671  T       C
1       rs3748597       0       878522  T       C
1       rs13303118      0       908247  G       T
1       rs9777703       0       918699  C       T
1       rs3934834       0       995669  T       C
1       rs9442372       0       1008567 A       G
1       rs3737728       0       1011278 T       C
1       rs6687776       0       1020428 T       C
```

Figure 3.9: Example of a BIM file

BED BED is a binary file containing the genotype.

PHENOTYPE This is a plain text file containing the phenotype in case it has not been specified in FAM file or there are more than only one phenotype. It needs a header with FID (Family ID), IID (Individual ID), and one column per phenotype (See Figure 3.10). In order to be consider, the command *–pheno* is required.

FID	IID	Ventricles_bl	Hippocampus_bl	WholeBrain_bl
1	014_S_0520	18132	-9	1096540
2	005_S_1341	32530	-9	832358
3	012_S_1175	23506	7214	1104960
4	012_S_0803	32425	4648	885844
5	018_S_0055	75838	-9	1130600
6	027_S_0118	36930	-9	944188
7	027_S_0403	27337	7509	1043920
8	053_S_0389	36932	5342	1064650
9	041_S_0262	25868	7975	965385
10	005_S_1224	87077	-9	1052900

Figure 3.10: Example of a phenotype file

COVARIATE This is a file similar to Phenotype file, needs a header with FID (Family ID), IID (Individual ID), and one column per covariate (See Figure 3.11).In order to be consider, the command *–covar* is required. In case *sex* is a covariate, there is no need to add it as a column on the covariate file: the command *–sex* automatically adds it.

FID	IID	PTGENDER	PTDOBYY
1	014_S_0520	2	1928
2	005_S_1341	2	1935
3	012_S_1175	1	1934
4	012_S_0803	2	1921
5	018_S_0055	1	1930
6	027_S_0118	1	1925
7	027_S_0403	1	1929
8	053_S_0389	2	1934
9	041_S_0262	1	1920
10	005_S_1224	1	1925
11	010_S_0472	1	1934
12	035_S_0048	1	1927

Figure 3.11: Example of a covariate file

3.6 Randomizations and sample selection

In this project, three different GWAS were performed: one including the subjects from ADNI1 (757 subjects), another with ADNI2 cohort (735 subjects) and the last one was a single GWAS including all the subjects in both phases (1492 subjects). That was made in order to study which SNPs are consistently preserved through ADNI cohorts in GWAS.

In addition, an exhaustive randomized analysis of the patients was performed in order to assess the consistency of the results and the sensibility of PLINK GWAS with different selections of patients (different sample size and different patients for each sample size). Therefore, from the whole population (entire ADNI cohort), different percentages of the total sample were analyzed: The randomized analysis selected distinct $n\%$ patients of the whole ADNI cohort, where n ranged from 30 to 95% in steps of 5%. For each n, 25 different sample randomizations were made in order to make the corresponding GWAS.

In total, in this study $14 \times 25 = 350$ GWAS randomized, plus 3 GWAS with ADNI cohorts were done.

3.7 Quality control of genetic data

Once the sample selection and PLINK files are ready, a vital step that should be part of any GWAS is the use of appropriate quality control (QC). This QC is necessary because the raw data is inherently imperfect, and without a quality data check, the GWAS will not be reliable. These errors coming from data can arise for numerous reasons related with genotyping, such as poor quality of DNA samples, poor DNA hybridization to the array, poorly probes, and sample mix-ups or contamination [35].

For instance, failing to thoroughly control for these data issues has led to the retraction of an article published by Sebastiani et al. (2010) in Science (Sebastiani et al., 2010, 2011; Sebastiani et al., 2012; Sebastiani et al., 2013). The results of the retracted article were affected by technical errors in the Illumina 610 array and an inadequate QC to account for those. Even though the main scientific findings remained supported after appropriate QC, the results of the new analysis deviated strongly enough for the authors to decide to retract the article.

In order to make the quality-control analysis, SNPs that were missing in more than 2% of the population were excluded, and also were subjects who had more than 10% missing genotypes. SNPs with minor allele frequency greater or equal to 5% were included, and markers that fail the Hardy-Weinberg test at a $10^{-}6$ significance threshold were excluded. PLINK provides some parameters and filters in order to have this QC done. These parameters are presented at Table 3.4 and specified below:

Features	As inclusion criteria	Meaning
Missingness per individual	–mind N	Exclude individuals with too much missing genotype data. This option is set as follows: plink –file mydata –mind 0.1 which means exclude with more than 10% missing genotypes (this is the defalt value).
Missingness per marker	–geno N	Exclude SNPs on the basis of MAF (minor allele frequency): plink –file mydata –maf 0.05 means only include SNPs with MAF >= 0.05. The default value is 0.01. This quantity is based only on founders
Allele frequency	–maf N	Exclude SNPs on the basis of missing genotype rate, with the –geno option: the default is to include all SNPS (i.e. –geno 1). To include only SNPs with a 90% genotyping rate (10% missing) use plink –file mydata –geno 0.1 As with the –maf option, these counts are calculated after removing individuals with high missing genotype rates.
Hardy-Weinberg equilibrium	–hwe N	Exclude markers that fail the Hardy-Weinberg test at a specified significance threshold, use the option: plink –file mydata –hwe 0.001 By default this filter uses an exact test. The standard asymptotic (1 df genotypic chi-squared test) can be requested with the –hwe2 option instead of –hwe.

Table 3.4: Parameters for PLINK analysis

Missingness per individual: GWA studies rely on the "common disease, common variant" hypothesis, which suggests that genetic influences on many common diseases will be at least partially attributable to a limited number of allelic variants present in more than 1% to 5% of the population [22]. According to this and to what it is found in the literature [34], [47], [22] [48], and in order to be able to compare the results with similar studies made about the same DDBB [46], mind parameter is set to 0.02.

Missingness per marker: The list of SNPs presented is already the intersection of all patients SNPs, so the parameter is set to 0.10.

Allele frequency: Due to the same reasons for missingness per individual, the parameter was set to 0.05.

Hardy-Weinberg equilibrium: Genotype mistakes can lead to increased random error and bias in gene-disease associations, so methods have been developed to detect and, where possible, deal with genotyping error. One of these methods is Hardy-Weinberg equilibrium test. [45] Hardy-Weinberg equilibrium

stands that, in the absence of migration, mutation, natural selection, and assortative mating, genotype frequencies at any locus are a simple function of allele frequencies. This phenomenon is called "Hardy-Weinberg equilibrium", and deviation from it at particular markers may suggest problems with genotyping or population structure or, in samples of affected individuals, an association between the marker and disease susceptibility. [44]

Consider a sample of SNP genotypes for N unrelated diploid individuals measured at an autosomal locus. The sample includes 2N alleles, including n_A copies of the rarer allele and n_B copies of the common allele. Let the number of heterozygous AB genotypes be n_{AB} ,and note that the numbers of AA and BB homozygous genotypes are n_{AA} = (n_A - n_{AB}) / 2 and n_{BB} = (n_B - n_{AB}) / 2 . Note that there are (2N)! / $n_A!n_B!$ possible arrangements for the alleles in the sample and that $2^{n_{AB}}$N!/($n_{AA}!$ $n_{AB}!$ $n_{BB}!$) of these arrangements correspond to exactly n_{AB} heterozygotes. Thus, under the assumption of HWE, the probability of observing exactly n_{AB} heterozygotes in a sample of N individuals with n_A minor alleles is [44]:

$$P(N_{AB} = n_{AB}|N, n_A) = \frac{2^{n_{AB}} N!}{n_{AA}!n_{AB}!n_{BB}!} \times \frac{n_A!n_B!}{(2N)!} \tag{3.1}$$

The expression for $P(n_{AB}|N, n_a)$ given in equation 3.1 leads to natural tests for HWE. For example, one could define one-sided tests that focus on detection of a deficit of heterozygotes, by calculating the statistic $P_{low} = P(N_{AB} \leq n_{AB}|N, n_A)$, or detection of an excess of heterozygotes, by calculating the statistic $P_{high} = P(N_{AB} \geq n_{AB}|N, n_a)$. In each case, the statistic can be calculated by simply summing over equation 3.1, to include all possible values of N_{AB} that are lower (for P_{low}) or higher (for P_{high}) than those observed in the actual data [44].

In brief, let the expected proportion of heterozygotes be $p_{AB} = 2pq$ and the two homozygote proportions be $p_{AA} = p^2$ and $p_{BB} = q^2$. The distribution is stable from generation to generation and genotypes occur at frequencies of p^2, q^2 and $2pq$. The Hardy- Weinberg equilibrium is represented in 3.12.

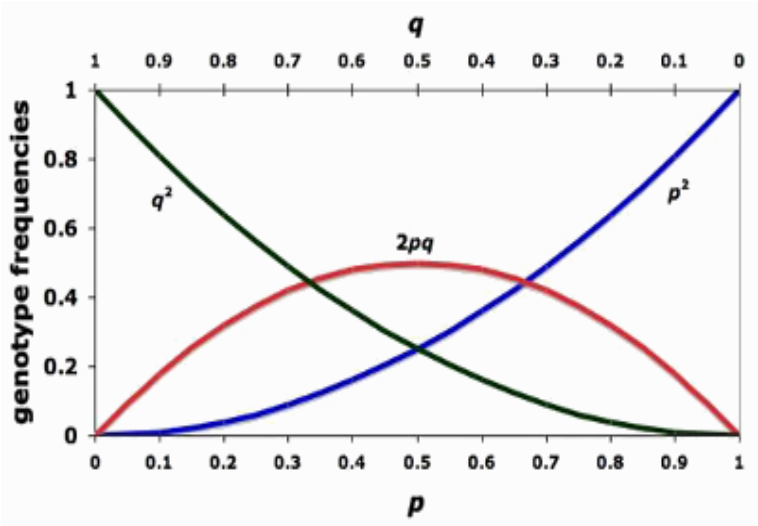


Figure 3.12: Hardy-Weinberg equilibrium

According with what is found in the literature [44], [46], here the significance threshold for HWE is 1e-06.

3.8 Association Analysis

In PLINK, a case-control analysis can be performed by an association analysis. However, quantitative traits (hippocampus volume from the last visit registered, in this case) can also be tested for association, using either asymptotic (likelihood ratio test and Wald test) or empirical significance values. Therefore, PLINK provides an implementation of the between/within model, which uses a permutation procedure (permuting genotype rather than phenotype) to control for the non-independence of individuals within the same family. The analysis of phenotype-genotype association is a standard regression of phenotype on genotype (how the mean value of phenotype for a given genotype changes with the value of genotype) that ignores family structure. Regression analysis is a reliable method of identifying which variables have impact on a topic of interest. The process of performing a regression allows to confidently determine which factors matter most, which ones can be ignored, and how these factors influence each other. The univariate regression analyzes one SNP at a time to assess its influence on the phenotype.

The regression model may be described by saying, for a given continuous phenotype (y), the SNP genotype at a given locus (x) follows a normal distribution with mean (See Equation 3.2):

$$E(y|x) = \alpha + \beta x \tag{3.2}$$

and variance σ^2 (a constant). The β is the regression coefficient or the parameter that represents the strength of association between the SNP x and the phenotype y. In order to estimate the parameters α, β and σ^2 that characterize the model, theoretical arguments lead to the following rule: α and β are estimated by *"the least squares*[3]*"* estimators, and a and b, namely the quantities which minimize the *residual* sum of squares, $\sum(y_i - Y_i)^2$ where Y_i is given by the estimated regression equation 3.3:

$$Y_i = a + bx \tag{3.3}$$

By solving this, it is found the parameter values that minimize the squared distance between observed phenotype value and the predicted phenotype value.

It can be shown by elementary calculus that a and b are given by the formulae 3.4 3.5:

$$a = \bar{y} - b\bar{x} \tag{3.4}$$

[3]The *least squares* estimators of α and β are also maximum likelihood estimators; furthermore, among all unbiased estimators they have the smallest standards errors

$$b = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \tag{3.5}$$

Also, an unbiased estimator of σ^2 is the residual sum of squares, $\sum(y - Y)^2$. See Equation 3.6 :

$$s_o^2 = \frac{\sum(y - Y)^2}{n - 2} \tag{3.6}$$

The divisor $n - 2$ on 3.6 is often referred to as the residual degrees of freedom, s_o^2 as the *residual mean square*, and s_o as the *standard deviation about regression*. The quantities a and b are called the *regression coefficients*, the term is often used particularly for b, the slope of the regression line.

When both x and y are random variables it might be useful to have a measure of the extent to which the relationship between the two variables approaches the extreme scenario in which every point on a scatter diagram fall exactly on a straight line. Such an index is provided by the *product-moment correlation coefficient* (or simply *correlation coefficient*), defined by:

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{(\sum(x - \bar{x})^2 \sum(y - \bar{y})^2)}} \tag{3.7}$$

This (3.7) provides a useful interpretation of the numerical value of r. The squared correlation coefficient is the fraction by which the sum of the squares of y is reduced to give the sum of squares of deviations from its regression on x. It is easy to note that $0 \leq r^2 \leq 1$.

In PLINK, significance is based on the following permutation procedure: genotypes are decomposed into between- and within-family components, results are applicable to study design in complex disease, especially for late-onset conditions for which parents are usually not available [38]. These two components are then permuted independently at the level of the family and are summed to form new pseudogenotype scores for each individual. That is, between components are swapped between families; within components have their sign swapped, with a 50% chance (similar for all members of the same family). This approach provides tests that give correct type I error rates accounting for the relatedness between individuals. Despite the need for permutation, one advantage is that nonnormal and dichotomous phenotypes can be appropriately analyzed. Whereas the basic test is of total association, the between and within components can also be tested separately [34]. For more details about the performance, see [38].

In the regression model 3.4, suppose that repeated sets of data are generated, each with the same n values of x but with randomly varying values of y. The statistics $\bar{y}$, a and b will vary from one set of data to another. Their sampling variances are obtained as follows:

$$var(\bar{y}) = \frac{\sigma^2}{n} \tag{3.8}$$

$$var(b) = \frac{\sigma^2}{\sum(x - \bar{x})^2} \tag{3.9}$$

$$var(a) = \sigma^2(\frac{1}{n} + \frac{\bar{x}^2}{\sum(x-\bar{x})^2}) \tag{3.10}$$

Formulae 3.10, 3.9 and 3.8 all involve the parameter σ^2. If inferences are to be made from one set of n pairs of observations on x and y, σ^2 will be unknown. It can, however, be estimated by the residual mean square, s_o (3.6). Estimated variances are, therefore,

$$var(\bar{y}) = \frac{s_o^2}{n} \tag{3.11}$$

$$var(b) = \frac{s_o^2}{\sum(x-\bar{x})^2} \tag{3.12}$$

and

$$var(a) = s_o^2(\frac{1}{n} + \frac{\bar{x}^2}{\sum(x-\bar{x})^2}) \tag{3.13}$$

and hypotheses about $\bar{y}$, a or b can be tested using the t-distribution on $n-2$ degrees of freedom [41] (Null Hypothesis H_o: $\beta = 0$ and Alternative Hypothesis $H_1 : \beta \neq 0$ for testing whether to reject the null hypothsis that there is no linear relationship between the given SNP and phenotype or not). With PLINK, a Wald test is performed. The Wald test is a parametric test for testing the truth value of the estimators based on the sample estimation. There, the maximum likelihood $\hat{\theta}$ of the parameter θ is compared to the observed value θ_0 under the hypothesis that their difference would follow a normal distribution. Generally, is the square of the difference what is compared to chi-square distribution. For univariate cases, Wald statistic is 3.14:

$$\frac{(\widehat{\theta} - \theta_0)^2}{\text{var}(\hat{\theta})} \tag{3.14}$$

which is compared against a chi-squared distribution.

When using PLINK for estimate the parameters, if the phenotype (column 6 of the PED file or the phenotype as specified with the –pheno option) is quantitative then PLINK will automatically treat the analysis as a quantitative trait analysis. The basic code to generate the analysis is:

```
plink --file mydata --assoc
```

The –assoc option will automatically perform an asymptotic version of the usual Student's-t test to compare two means.

Given a quantitative phenotype, –assoc writes regression statistics and Wald test results to *plink.qassoc* and will generate the file with fields as follows:

CHR: Chromosome number
SNP: SNP ID
BP: Physical position (base-pair)
NMISS: Number of non-missing genotypes

BETA: Regression coefficient
SE: Standard error
R2: Regression r-squared
T: Wald test (based on t-distribtion)
P: Wald test asymptotic p-value

Nevertheless, this analysis can be made on a finer way by adding some modifiers. The "qt-means" modifier causes an additional *plink.qassoc.means* file to be generated, reporting trait means and standard deviations stratified by genotype. Also, an output modifier (*–pfilter 1e-4*) for only report statistics with p-values less than 1e-4 could be set. The parameters above mentioned are set as follows:

```
./plink --bfile test --geno 0.10 --maf 0.05 --hwe
   0.000001 --mind 0.02 --pfilter 1e-4 --assoc  --
   pheno phenotypes.txt --all-pheno --out as1
```

Chapter 4

Results

4.1 Descriptive Analysis

In order to be able to make a better interpretation of results, a first descriptive step is presented here, where differences between healthy and affected by AD groups can be found for the genetic control (APOE $\epsilon 4$ copies) and the phenotype (hippocampus volume), even though it does not necessary mean that this difference is significant.

Differences on APOE $\epsilon 4$ allele distribution are represented on Figure 4.1 and it is notable that most of people from control group had no $\epsilon 4$ alleles or had just one copy. On the other hand, more people from AD than CN group present two copies of the allele.

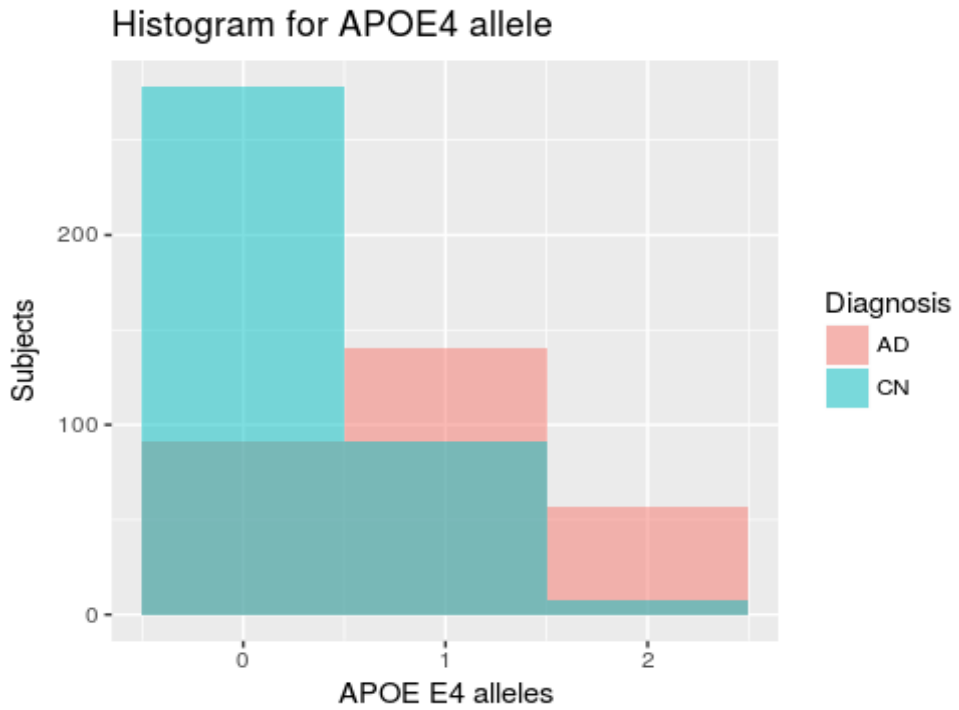


Figure 4.1: Number of APOE $\epsilon 4$ alleles by AD and CN groups on together ADNI1 and ADNI2 cohorts

Difference on Hippocampus' volume between healthy and AD groups is clear (See Figures 4.2 , 4.3, where CN controls seems to have more Hippocampus volume than AD patients. In 4.3 means of the two groups are also represented and the difference between them is remarkable.

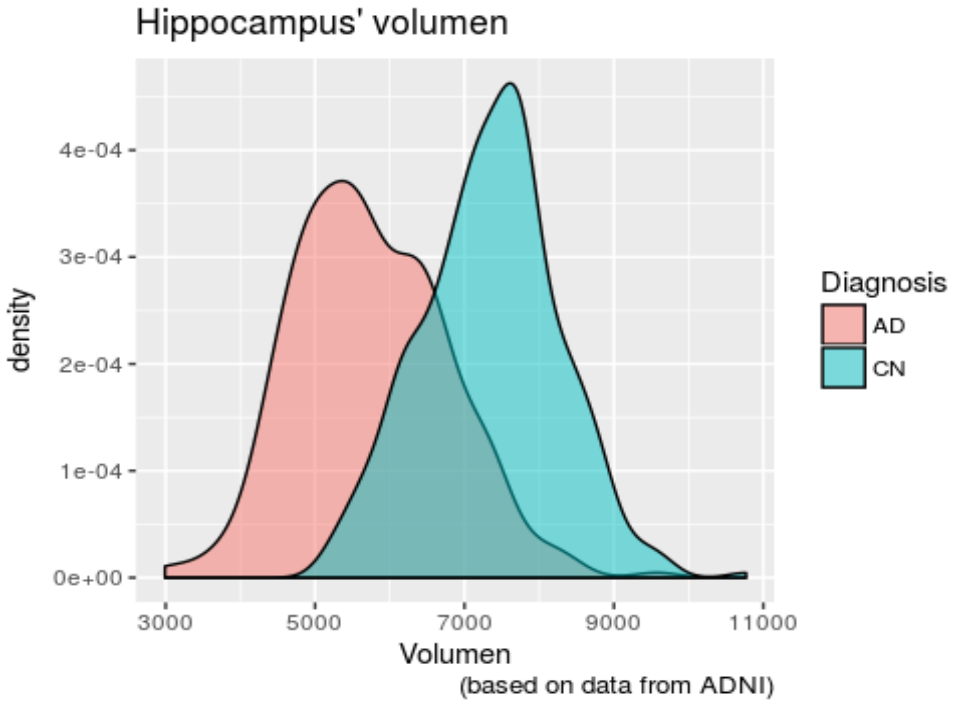


Figure 4.2: Density plot in which distributions on Hippocampus' volume between the studied groups are shown

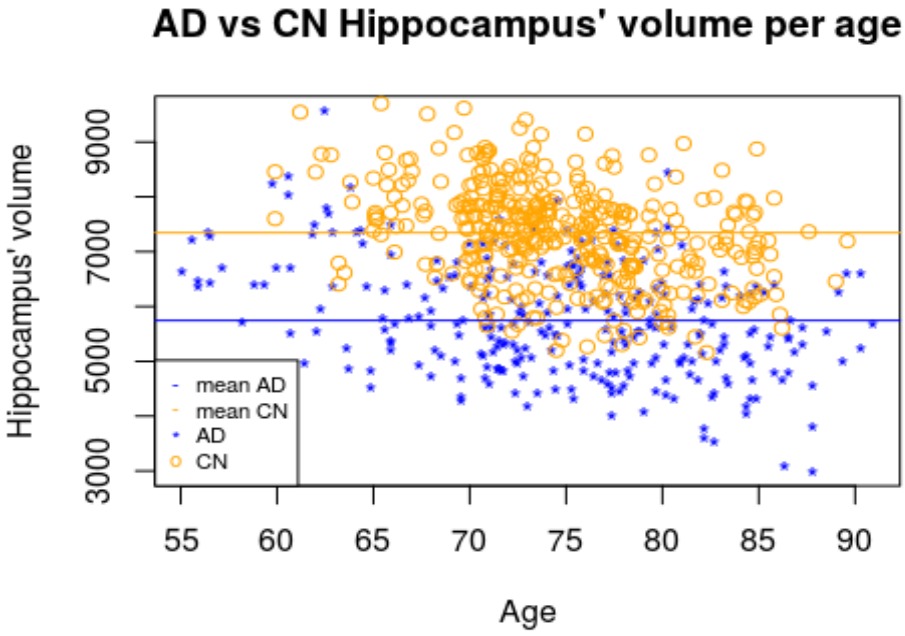


Figure 4.3: Hippocampus' volume by age and studied groups

4.2 P-value Correction

In order to be able to extract information from the GWAS, it is necessary to establish from which p-value we consider there is association between the SNP and the phenotype. At the conventional $p < 0.05$ level of significance, an association study of 1 million SNPs will result on 50.000 SNPs "associated" with the disease, almost all falsely positive and due to chance alone. The most common method of dealing with this problem is to reduce the false-positive rate by applying the *Bonferroni* correction, in which the conventional P-value is divided by the number of tests performed. The same 1 million SNP study would thus use a threshold of $p < 0.05/10^6$, or 5×10^8, to identify associations unlikely to have occurred by chance. This correction has been criticized as overly conservative because it assumes independent associations of each SNP with disease even though individual SNPs are known to be correlated to some degree due to linkage disequilibrium[1] [21]. In most of the literature, and also in Potkin et al. 2009 [46] , it is established that significant SNPs have p-values equal or less than 10^{-06}. Therefore, and in order to be able to compare the results of this study with Potkin et al. 2009 [46], 10^{-06} will be the threshold used here. Nevertheless, 10^{-08} will be considered for the representation of results.

[1]Linkage disequilibrium (LD) is the correlation between nearby variants such that the alleles at neighboring polymorphisms (observed on the same chromosome) are associated within a population more often than if they were unlinked. [49]

4.3 GWAS Results Representation

GWAS results are typically presented in Manhattan plots and Q-Q plots. Therefore, in order to visualize the results obtained in the non randomized association analysis, Manhattan and Q-Q plots are drawn with R (*package "qqman"*). This package is able to work with PLINK's outputs.

Manhattan plot

A Manhattan plot is a specific type of scatter plot widely used in genomics to study GWAS results. Each point represents a genetic variant. The X axis shows its position on a chromosome, the Y axis tells how much it is associated with a trait (hippocampus volume, in this case) [42]. Then, a Manhattan plot displays the association of the P value for each SNP in the genome (displayed as $-\log_{10}(p - value)$). Here, the horizontal lines display the cutoffs for two significance levels: blue line for high significance ($p < 10^{-6}$) and red line for genome-wide significance level ($p < 10^{-8}$). See Figures 4.5, 4.9, 4.13.

Q-Q plot

The Q-Q plot, or quantile-quantile plot, is a graphical tool that helps to assess if a set of data plausibly came from some theoretical distribution such as a Normal or exponential. It is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, the points should form a line that is roughly straight [43].

Concerning GWAS, population structure should be assessed and reported, typically by examining the distribution of test statistics generated from the thousands of association tests performed and assessing their deviation from the null distribution (that is expected under the null hypothesis of no SNP associated with the trait) in a "Q-Q," plot. In these plots, observed association statistics or calculated P-values for each SNP are ranked in order from smallest to largest and plotted against the values expected had they been sampled from a distribution of known form. Deviations from the diagonal identity line suggest that either the assumed distribution is incorrect or that the sample contains values arising in some other manner, as by a true association [21]. Here, more than one point above the diagonal can be found (See Figures 4.7, 4.11, 4.15).

4.4 Whole Population Results

The top 20th associated SNPs results of the GWAS in PLINK for ADNI1 together with ADNI2 cohort (the whole population) are shown in Figure 4.4, including all parameters before mentioned (Chromosome number,SNP ID, base-pair position of the SNP,non-missing genotypes, β regression coefficient, SE, r^2 measurement, Wald test based on t-distribtion and Wald test asymptotic p-value):

CHR	SNP	BP	NMISS	BETA	SE	R2	T	P
19	rs429358	50103781	1289	-408.6	48.7	0.05187	-8.391	1.256E-16
19	rs2075650	50087459	1289	-327	53.26	0.02846	-6.14	1.095E-09
14	rs7149001	96851587	1288	-229.5	49.25	0.01661	-4.66	0.000003485
3	rs10510380	7957083	1285	228.5	50.44	0.01574	4.53	0.000006454
1	rs301798	8411152	1288	-226.9	50.16	0.01566	-4.523	0.00000665
13	rs7328292	77246192	1289	257.1	56.88	0.01562	4.519	0.000006782
10	rs1002598	15467228	1286	-234.9	52.6	0.0153	-4.466	0.000008657
11	rs1267476	34283047	1289	211.2	47.46	0.01515	4.45	0.00000932
7	rs4628172	15461675	1289	275.5	62.67	0.01479	4.395	0.00001196
22	rs9610216	20428687	1289	227.3	52.27	0.01447	4.347	0.00001486
14	rs7156868	96898501	1289	-224.9	52.04	0.0143	-4.321	0.00001674
4	rs2306174	76708606	1289	-305.1	71.98	0.01377	-4.239	0.0000241
19	rs157580	50087106	1286	209.3	49.44	0.01377	4.234	0.00002458
14	rs7161686	96898748	1286	-220.5	52.14	0.01373	-4.229	0.00002519
6	rs2690106	25365559	1284	194.2	46.62	0.01335	4.165	0.00003322
11	rs7115847	1884793	1285	199.5	48.2	0.01318	4.139	0.00003712
1	rs3765971	8367947	1289	-206.1	49.98	0.01304	-4.123	0.00003976
1	rs953043	8697038	1287	-202.8	49.89	0.01269	-4.064	0.00005119

Figure 4.4: Output of the analysis of ADNI1 and ADNI2 together (top 20 SNPs)

The analysis in the whole cohort of patients reported the expected association for APOE *rs429358* and TOMM40 *rs2075650* SNPs according to the current knowledge (p-values of 1.256e-16 and 1.095e-09, respectively). Nevertheless, they are not strongly correlated (r^2 is 0.05187 and 0.02846). In addition to *rs429358* and *rs2075650*, and close but below to the 10^{-6} threshold, six SNPs were found associated in the analysis of the 100% of the population: *rs301798* (located in chromosome 1, p-value=6.65e-06), *rs1267476* (chromosome 11, p-value=9.32e-06), rs1002598 (chromosome 10, p-value = 8.657e-06), rs7149001 (chromosome 14, p-value = 3.485e-06), *rs10510380* (chromosome 13, p-value = 6.454e-06) and *rs7328292* (chromosome 3, p-value = 6.782e-06). These SNPs will be tracked in the rest of the GWAS' and colored in green on the Manhattan plots. The Manhattan and Q-Q plots representation of the results can be found in Figures 4.5 and 4.7 .

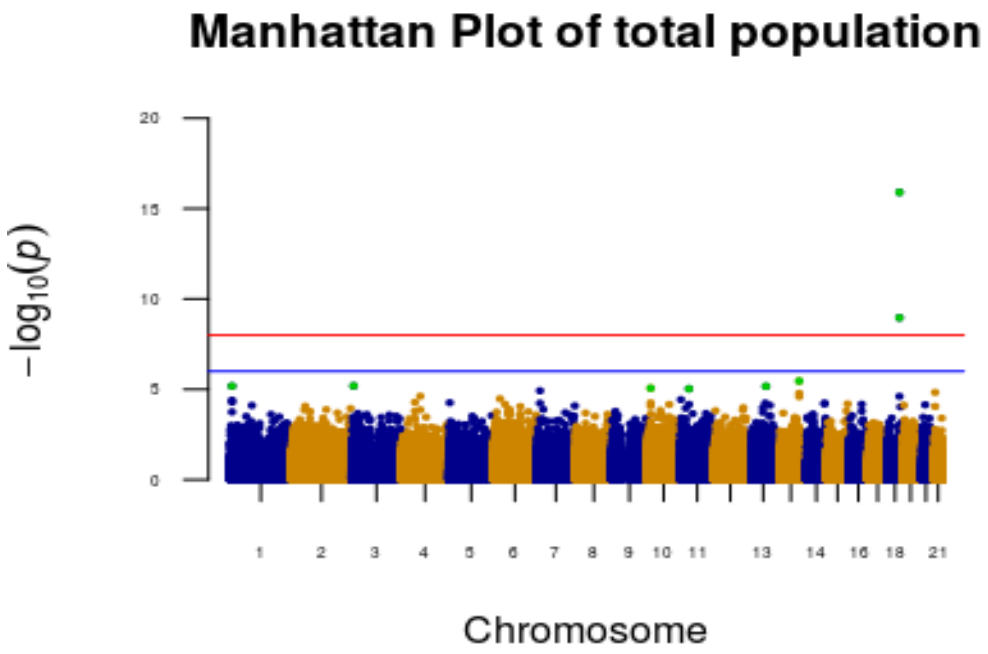


Figure 4.5: Manhattan plot of the quantitative trait (hippocampus volume) genome wide association analysis for the total population. In green, the significant SNPs and the most close to the 10^{-6} threshold ones.

In order to have a better visualization of APOE and TOMM40, another Manhattan plot, zooming the 19th chromosome, is made 4.6:

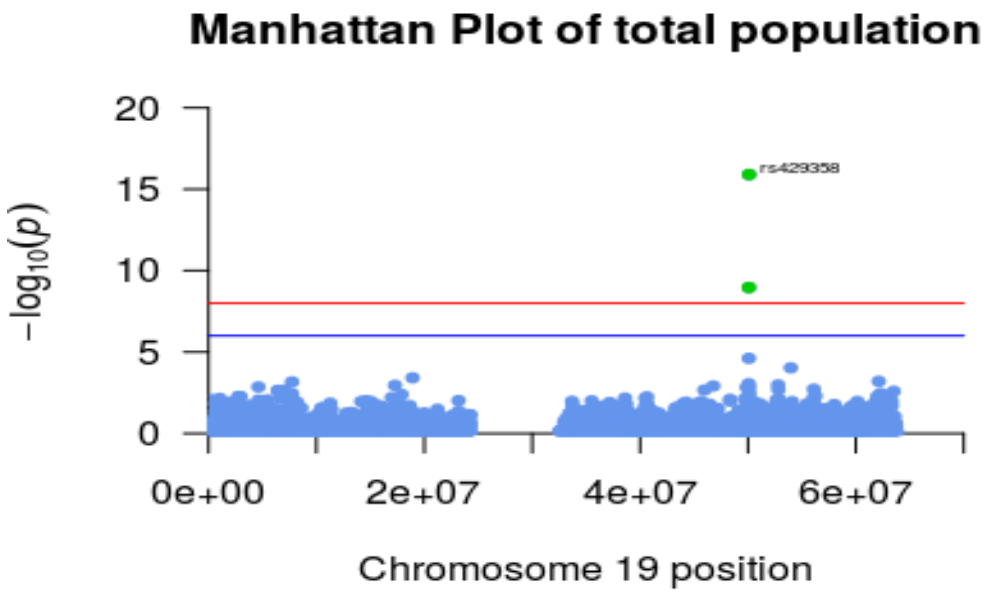


Figure 4.6: Manhattan plot for the 19th chromosome of the quantitative trait (hippocampus volume) genome wide association analysis for the total population. In green, APOE and TOMM40 SNPs. The SNP with biggest p-value is annotated.

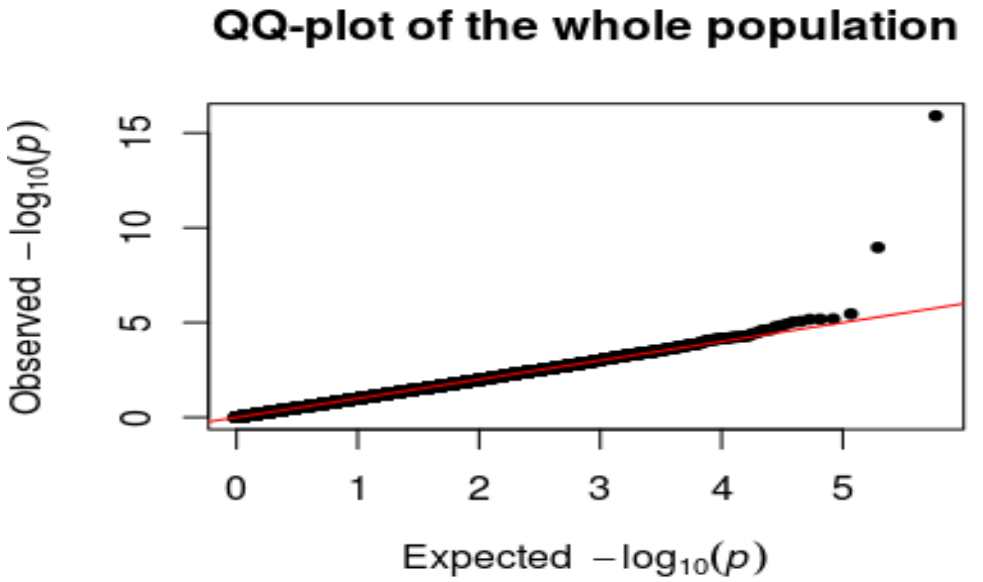


Figure 4.7: Q-Q plot of the quantitative trait (hippocampus volume) genome wide association analysis.

4.5 ADNI1 Cohort Results

The top 20th associated SNPs results of the GWAS in PLINK for ADNI1 are shown in Figure 4.8, including all parameters specified above:

CHR	SNP	BP	NMISS	BETA	SE	R2	T	P
19	rs429358	50103781	619	-501.1	67.62	0.08173	-7.41	4.162E-13
19	rs2075650	50087459	619	-393.2	73.69	0.0441	-5.336	1.339E-07
16	rs8051428	83424503	619	-310.2	69.97	0.03088	-4.434	0.00001096
2	rs7601712	65119499	619	296.1	67	0.03068	4.419	0.00001169
6	rs494991	87540413	617	452.4	105.9	0.02884	4.274	0.0000223
25	rs5941380	88991421	607	-318.4	74.76	0.02911	-4.259	0.00002379
7	rs6947191	123584877	618	362.9	85.5	0.02842	4.245	0.00002521
15	rs10520754	92760413	619	275.6	65.78	0.02767	4.19	0.00003192
19	rs2240747	5407930	619	-351.2	83.85	0.02764	-4.188	0.00003223
6	rs1066335	87469692	618	424.8	102.6	0.02705	4.139	0.00003978
14	rs10141863	99346460	619	-267	64.92	0.02669	-4.113	0.00004431
4	rs6835813	139803551	617	306.1	75.14	0.02628	4.074	0.00005224
1	rs17017668	209761131	618	455.7	112.1	0.02612	4.065	0.00005433
21	rs2834569	34952085	617	305.5	75.8	0.02573	4.03	0.00006274
6	rs13211072	84582503	619	-273.6	68.31	0.02535	-4.006	0.00006928
10	rs4075854	48105962	619	329.2	82.9	0.02492	3.971	0.00007996
11	rs7115982	81936646	618	308.7	77.77	0.02493	3.969	0.0000807
14	rs8003168	71433141	619	275.1	69.32	0.02489	3.968	0.00008088
6	rs722903	88572771	618	-265.7	67.03	0.02487	-3.964	0.00008248

Figure 4.8: Output of the analysis of ADNI1 cohort (top 20 SNPs)

The results reported only a positive association for APOE *rs429358* and TOMM40 *rs2075650* SNPs in ADNI1 (p-values of 4.162e-13 and 1.339e-07, respectively). Nevertheless, they are not strongly correlated (r^2 is 0.08173 and 0.0441, respectively). A representation of the results can be found in Figures 4.9 and 4.11:

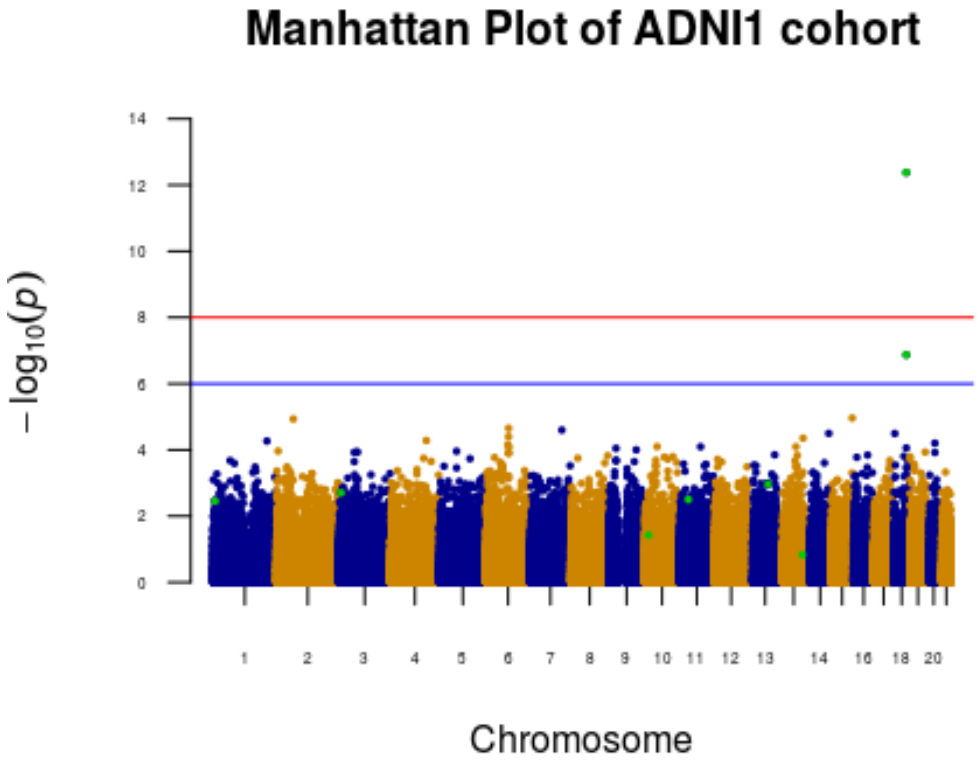


Figure 4.9: Manhattan plot of the quantitative trait (hippocampus volume) genome wide association analysis for ADNI 1 cohort. In green, the significant SNPs and the most close to the 10^{-6} threshold ones from the whole population GWAS.

In order to have a better visualization of APOE and TOMM40, another Manhattan plot, zooming the 19th chromosome, is made 4.10:

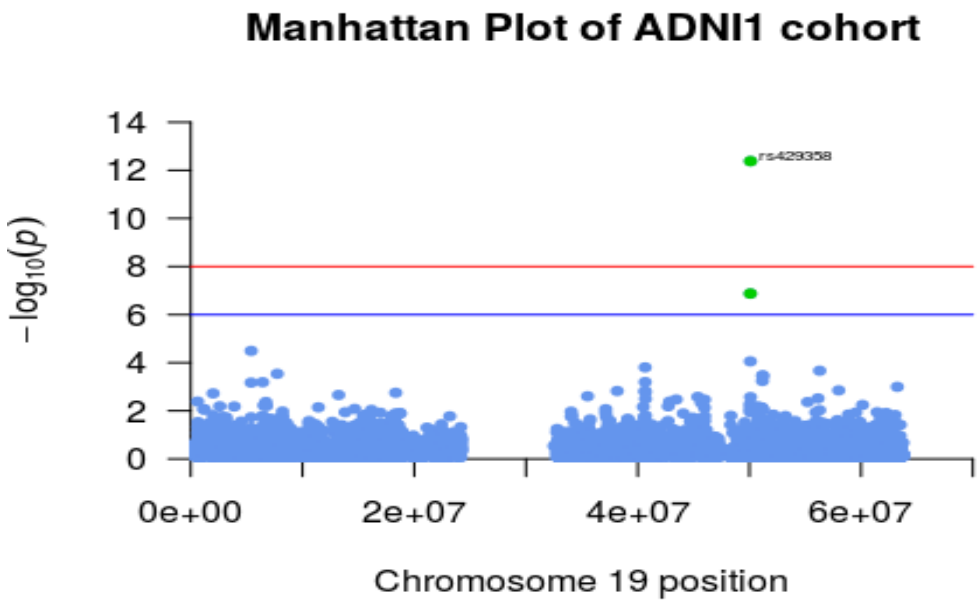


Figure 4.10: Manhattan plot for the 19th chromosome of the quantitative trait (hippocampus volume) genome wide association analysis for ADNI1 cohort. In green, APOE and TOMM40 SNPs. The SNP with biggest p-value is annotated.

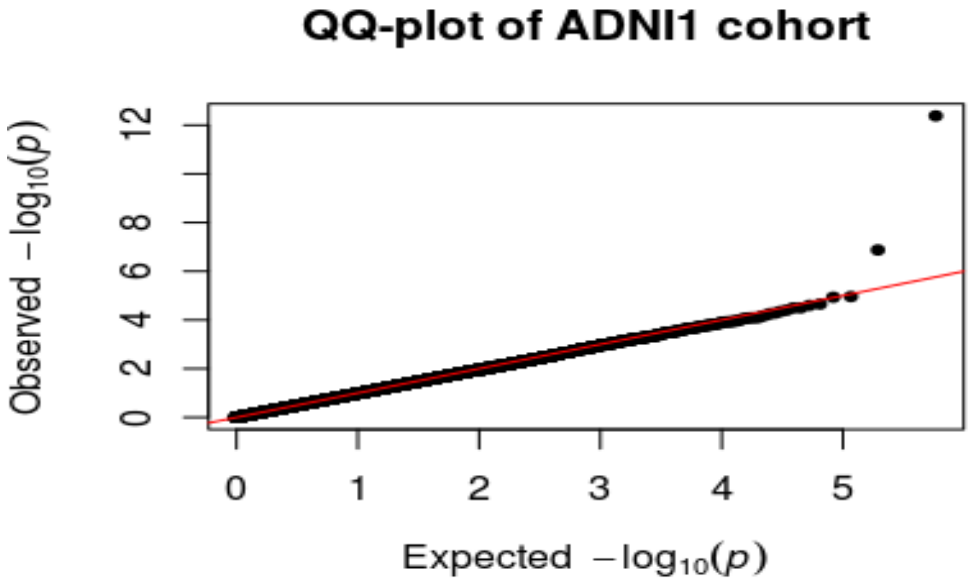


Figure 4.11: Q-Q plot of the quantitative trait (hippocampus volume) genome wide association analysis for ADNI1 cohort.

4.6 ADNI2 Cohort Results

As in previous sections, the top 20th associated SNPs results of the GWAS in PLINK for ADNI2 are shown in Figure 4.12, including all parameters specified above:

CHR	SNP	BP	NMISS	BETA	SE	R2	T	P
14	rs7149001	96851587	669	-321.9	64.08	0.03646	-5.024	6.502E-07
23	rs12400933	3505213	670	-325.4	66.52	0.03458	-4.891	0.000001255
21	rs2830241	26767184	670	285.5	61.17	0.03158	4.667	0.000003688
10	rs1002598	15467228	668	-315.9	69.37	0.03019	-4.553	0.000006281
9	rs933034	13520711	670	266.9	61.45	0.02747	4.344	0.00001619
13	rs4374042	38671999	670	-276	63.59	0.02743	-4.341	0.0000164
8	rs7833031	66808103	670	281.4	64.87	0.02741	4.339	0.00001655
20	rs17339921	4824430	669	-582.8	134.9	0.02723	-4.321	0.00001792
14	rs2001558	48544898	669	344.4	79.79	0.02718	4.317	0.00001825
21	rs979144	27447203	670	482.4	112.9	0.0266	4.273	0.00002211
2	rs6750398	235683421	669	289.2	68.28	0.0262	4.236	0.00002593
14	rs10483891	77678494	670	349.9	82.95	0.02595	4.219	0.00002795
7	rs17147407	8330741	670	-320.8	76.96	0.02536	-4.169	0.00003464
8	rs6472207	66636751	670	292	70.51	0.02503	4.141	0.00003899
8	rs10808746	66807267	670	252.7	61.1	0.02496	4.135	0.00003995
14	rs7161686	96898748	670	-283.7	68.66	0.02492	-4.132	0.00004055
14	rs7156868	96898501	670	-284	68.87	0.02483	-4.124	0.00004188
6	rs11968561	43940780	670	-304.3	74.26	0.02451	-4.097	0.00004695
13	rs9532358	38691979	669	-263	64.19	0.02455	-4.097	0.00004695
6	rs12205787	16907583	670	-318.1	78.03	0.02427	-4.076	0.00005126
18	rs953845	74068537	668	274.3	67.76	0.02402	4.049	0.00005759
17	rs7219555	57094904	669	260.8	64.57	0.02388	4.039	0.00005987
19	rs429358	50103781	670	-268.6	66.5	0.02383	-4.038	0.00006004
18	rs9676093	72211572	670	270.6	67.04	0.02381	4.036	0.0000606

Figure 4.12: Output of the analysis of ADNI2 cohort (top 20 SNPs)

In this case, the GWAS performed for the ADNI2 cohort did not show a significant association between them and the hippocampal volume (p-value 6.004e-05 for APOE SNP and 0.002591 for TOMM40). A representation of the results can be found in Figures 4.13 and 4.15:

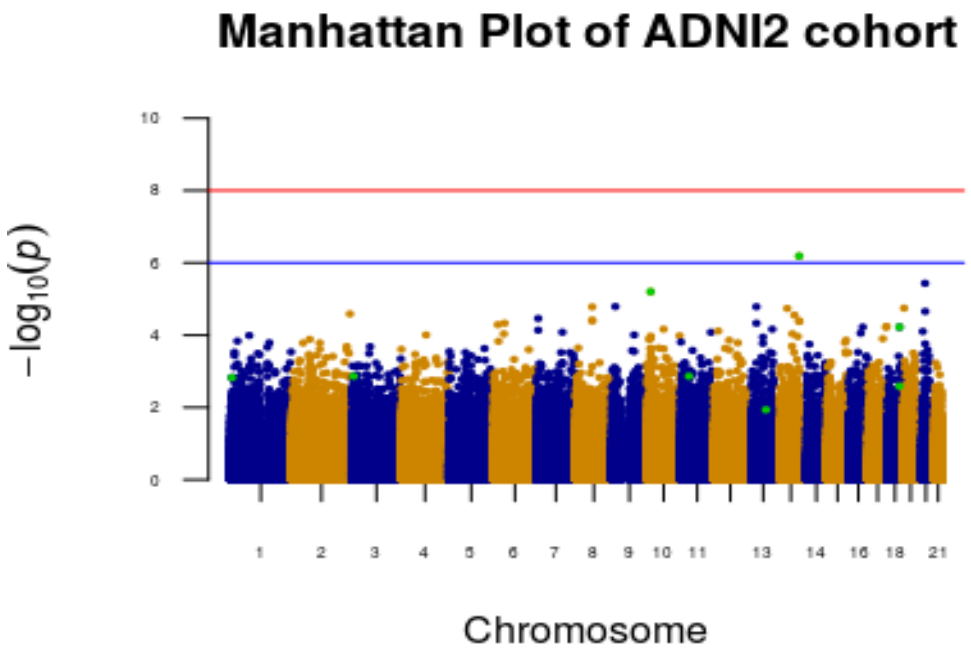


Figure 4.13: Manhattan plot of the quantitative trait (hippocampus volume) genome wide association analysis for ADNI 2 cohort. In green, the significant SNPs and the most close to the 10^{-6} threshold ones from the whole population GWAS.

Here, also another Manhattan plot zooming the 19th chromosome is made 4.14:

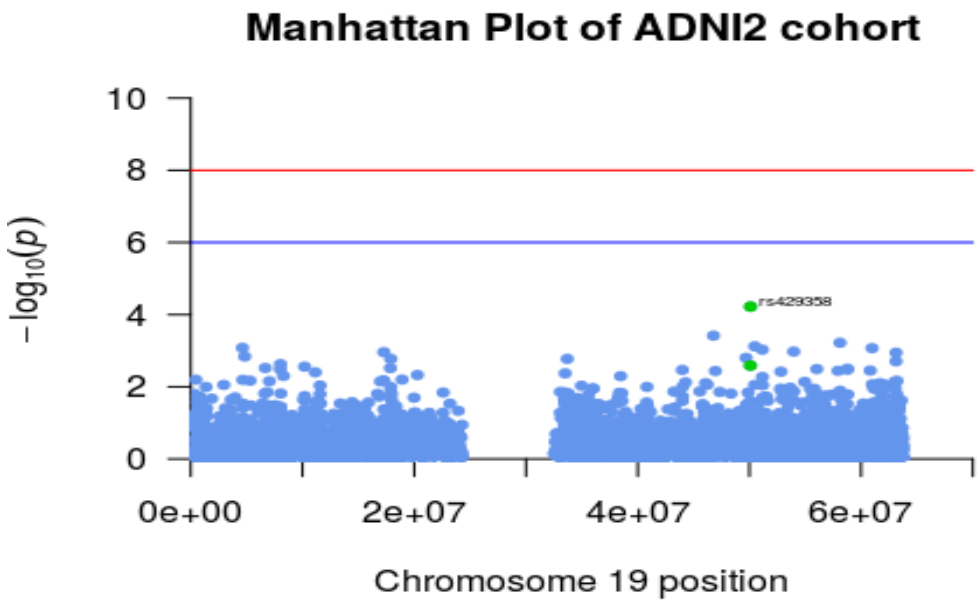


Figure 4.14: Manhattan plot for the 19th chromosome of the quantitative trait (hippocampus volume) genome wide association analysis for ADNI2 cohort. In green, APOE and TOMM40 SNPs. The SNP with biggest p-value is annotated.

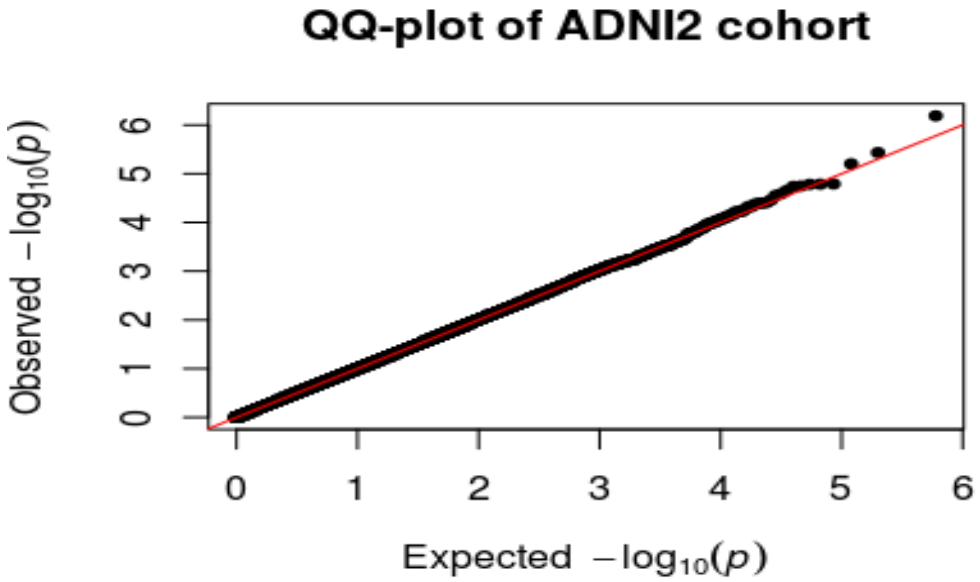
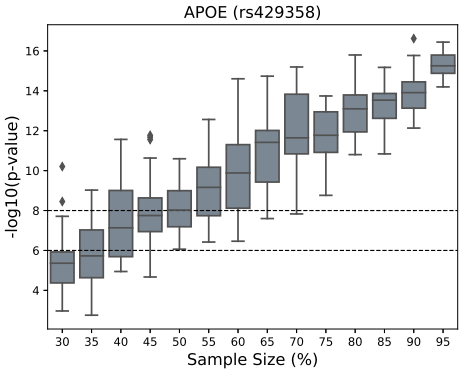


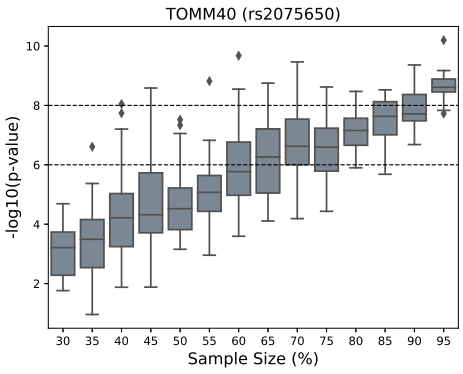
Figure 4.15: Q-Q plot of the quantitative trait (hippocampus volume) genome wide association analysis for ADNI2 cohort.

4.7 Randomized Population Results

In what concerns to the randomized population, significant NPs from the whole population were tracked through the different percentages of sample size and the consecutive 25 randomizations. Figures 4.17, 4.16 and 4.18 show the results of the p-values for the different GWAS:



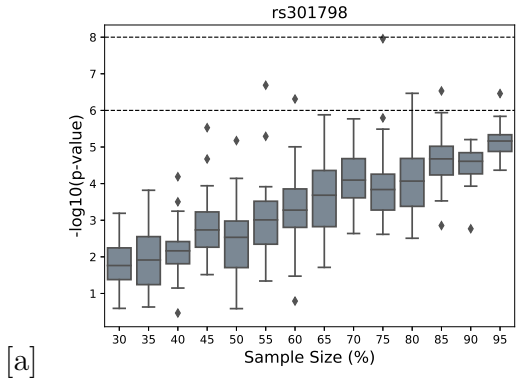
[a]



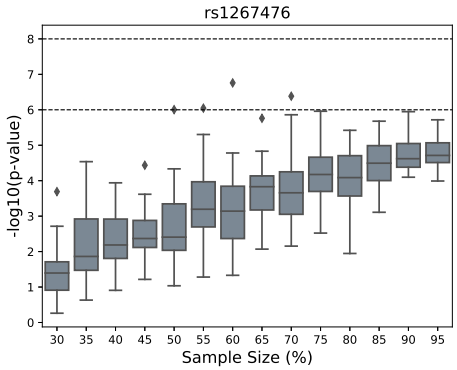
[b]

Figure 4.16: Statistical distribution of the p-values (log 10 (Observed p-value)) obtained in the randomized analysis. (a) Results for APOE SNP.(b) Results for TOMM40 SNP.

Although these are not significant results, after performing a randomized analysis similar to the one applied for APOE and TOMM40, we noticed that there is a positive trend in the association significance between these SNPs and the hippocampal volume, suggesting a possible association with a higher sample size (See Figures 4.17 and 4.18.)

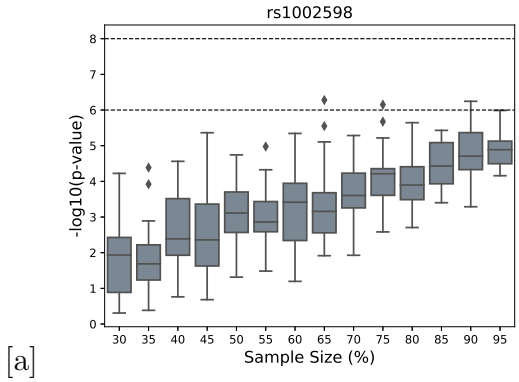


[a]

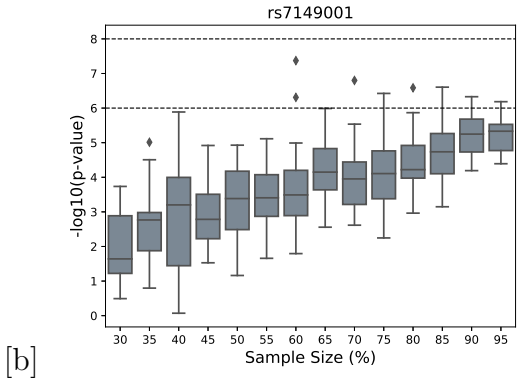


[b]

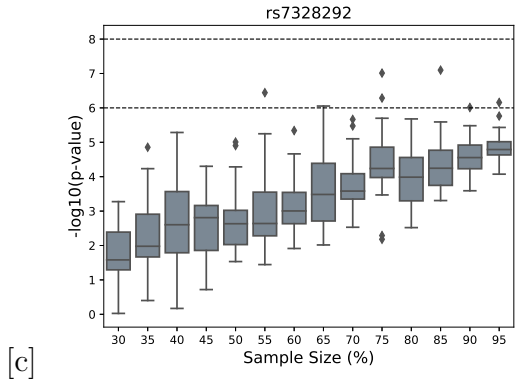
Figure 4.17: Statistical distribution of the p-values (log 10 (Observed p-value)) obtained in the randomized analysis for *rs301798* (a) and *rs1267476* (b) SNPs associated to some brain-related disorders, such as Schizophrenia and Parkinson [39], [40]



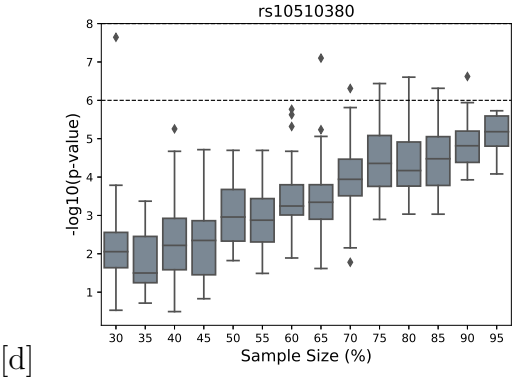
[a]



[b]



[c]



[d]

Figure 4.18: Statistical distribution of the p-values (log 10 (Observed p-value)) obtained in the randomized analysis. (a) Results for the SNP *rs1002598*.(b) Results for the SNP *rs7149001*.(c) Results for the SNP *rs7328292*.(d) Results for the SNP *rs10510380*.

To sum up, the results reported a positive association for APOE *rs429358* and TOMM40 *rs2075650* SNPs, only in ADNI1 (p-values of 4.162e-13 and 1.339e-07, respectively). The GWAS performed for the ADNI2 cohort did not show a significant association between them and the hippocampal volume (p-value 6.004e-05 for APOE SNP and 0.002591 for TOMM40). The analysis in the whole cohort of patients reported the expected association for APOE and TOMM40 according with the current knowledge (p-values of 1.256e-16 and 1.095e-09, respectively). From the exhaustive randomized analysis, the association was consistent with the results in the whole ADNI cohort from a sample size percentage of 80-90% (See Figure 4.16)

Chapter 5

Discussion and conclusions

To ensure that the strongest associations do not reflect genotyping artifacts, additional checks on genotyping quality should be made. Associations with any known "positive controls" (such as APOE here), should be reported in order to increase confidence in the consistency of findings with prior reports [21]. The results presented here are consistent with the state of the art and with those from the original article *Potkin et al.*. Hence, the power of this approach is validated, and emphasizes the value of the results implicating novel genes in candidate neurodegenerative mechanisms. Nevertheless, not all the SNPs that *Potkin et al.* pointed out were found in this study (*rs10074258*, *rs12654281*, *rs10781380*, *rs8115854*, *rs6031882*, *rs2073145*, *rs10867752*, *rs1082714* and *rs11626056* were found on the original article but not here). That could be due to the variability of the results depending on the sample size.

Additionally, a typical GWA study has 4 parts: (1) selection of a large number of individuals with the disease or trait of interest and a suitable comparison group; (2) DNA isolation, genotyping, and data review to ensure high genotyping quality; (3) statistical tests for associations between the SNPs passing quality thresholds and the disease/trait; and (4) replication of identified associations in an independent population sample or examination of functional implications experimentally. Association studies as the one presented here essentially identify a genomic location related to disease but do not provide new information on gene function. For such porpuse, there are aditional studies. The examination of known SNPs in high linkage disequilibrium with the associated SNP may identify variants with plausible biologic effects. Also, tissue samples or cell lines can be examined for expression of the gene variant. Other functional studies may include genetic manipulations in cell or animal models, such as knockouts or knock-ins. The three first steps have already been taken, and a future line of research could be based on step 4: not only for TOMM40 and APOE SNPs, but specially for*rs301798, rs1267476, rs1002598, rs7149001 , rs10510380* and *rs7328292*, because there is not information or biologic studies related to them, and they could be a very interesting source of diagnosis and knowledge.

Other applications of GWA studies is that they can also demonstrate gene-gene interactions, or modification of the association of one genetic variant by another (as with GAB2 and APOE in Alzheimer's disease), and can detect high-risk haplotypes or combinations of multiple SNPs within a single gene. Therefore, more work could

be done in this direction. For example, by performing a multivariate regression that analyze all SNPs jointly to assess the influence of each SNP on the given phenotype in the presence of all the other SNPs (this model would be more realistic for common diseases, like Alzheimer's disease, that are believed to be caused by the contribution of many different genetic loci, each having small effect on the disease susceptibility).

As above mentioned, replication is the first step in studies as similar as possible to the initial report as it is made here. Then, it may be extended to related phenotypes, different populations or different study designs to refine and extend the initial findings and increase confidence in conclusions [21].

In conclusion, the two most significant SNPs from the whole population (*rs2075650* and *rs429358*) are both located on the 19th chromosome, and form part of genes TOMM40 and APOE. In fact, these genes are close to each other (See Figure 5.1) and are part of two genes related to Alzheimer's in the literature [1, 18, 28, 30]: TOMM40 and APOE. Other SNPs that might be relevant results (notice that although they are close, they do not pass the significance threshold) could be *rs301798* and *rs1267476*, SNPs located at the intron region of RERE (arginine-glutamic acid dipeptide repeats) gene and ABTB2 (Ankyrin Repeat and BTB domain containing 2) gene, respectively. Both are known to be associated to some brain-related disorders, such as Schizophrenia and Parkinson [39], [40]. We encourage further studies for describing the specific role of the other SNPs found (*rs301798, rs1267476, rs1002598, rs7149001 , rs10510380* and *rs7328292*), which may become important in the early diagnosis of the disease.

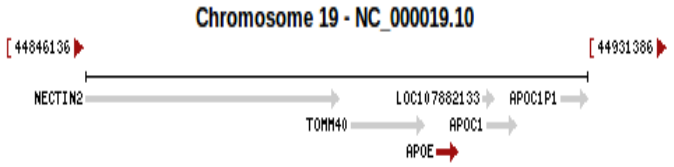


Figure 5.1: APOE gene

To sum up, in this work, the consistency of GWAS analysis through the different ADNI cohorts was studied. A positive association was found for APOE and TOMM40 SNPs with the hippocampal volume in the ADNI1 cohort and in the union of ADNI1 and ADNI2 populations. However, the association for APOE and TOMM40 was not reported in ADNI2. It can be hypothesized that the variability of association based on sample size is the reason behind that result, and this hypothesis is supported taking into account the results obtained in the randomization of the 50% of the population, but future work is needed to be able to confirm it. In addition, the study reported a weak association of SNPs that are known to be associated to brain-related disorders. For future work, the study should be extended to upcoming ADNI3 database and the associations with other brain structures.

Bibliography

[1] Alzheimer's Association. 2017 alzheimer's disease facts and figures. *Alzheimer's Dementia*, 13:325–373, 2017.

[2] Wilson RS, Segawa E, Boyle PA, Anagnos SE, Hizel LP, Bennett DA. The natural history of cognitive decline in Alzheimer's disease. *Psychol Aging*, 27: 1008–10017, 2012.

[3] Dennis J. Selkoe Alzheimer's disease results from the cerebral accumulation and cytotoxicity of amyloid -protein *Journal of Alzheimer's Disease* , 3: 75–80, 2001.

[4] Sherry L. Murphy, B.S., Jiaquan Xu, M.D., Kenneth D. Kochanek, M.A., Sally C. Curtin, M.A., Elizabeth Arias. Deaths: Final Data for 2015. *National Vital Statistics Reports*, 66:6, 2017.

[5] Kochanek KD, Murphy SL, Xu JQ, Tejada-Vera B. Deaths: Final Data for 2014, *National Vital Statistics Reports*, 65:1-122, 2016.

[6] World Health Organization. International statistical classification of diseases and related health problems. WHO Press, 2004.

[7] Hebert LE, Beckett LA,Scherr PA, Evans DA, Hebert LE, Beckett LA. Annual incidence of alzheimer disease in the united states projected to the years 2000 through 2050. Alzheimer Disease and Associated Disorders, 15:169-73, 2001.

[8] Jason Karlawish, Clifford R. Jack, Jr. Walter, A. Rocca Heather, M. Snyder, C.Carrillo. Alzheimer's disease: The next frontier—special report 2017. *Alzheimer's Dementia*, 13:374-380, 2017.

[9] Hebert LE., Weuve J., Scherr PA., Evans DA., Hebert LE., Weuve J. *Neurology*, 80:1778–1783, 2013.

[10] H. Bickel. Demenzsyndrom und alzheimer krankheit: Eine schätzung des krankenbestandes und der jährlichen neuerkrankungen in deutschland. *AG Psychiatrische Epidemiologie, Klinik und Poliklinik für Psychiatrie und Psychotherapie der Technischen Universität München.*

[11] Carter CL., Resnick EM., Mallampalli M., Kalbarczyk A., Carter CL., Resnick EM. *J Womens Health*, 21:1018–1023, 2012.

[12] Mahley RW., Rall SC. Jr. Apolipoprotein E: Far more than a lipid transport protein., *Annual Review of Genomics and Human Genetics*, 1: 507-537, 2000.

[13] Loy CT., Schofield PR., Turner AM., Kwok JBJ. Genetics of Dementia *Lancet*, 40:828–840, 2014.

[14] Holtzman DM., Herz J., Bu G. Apolipoprotein E and apolipoprotein E receptors: Normal biology and roles in Alzheimer disease. *Cold Spring Harb Perspect Med*, 2:a006312, 2012.

[15] Alzheimer's Drug Discovery Foundation. *http://www.alzdiscovery.org/cognitive-vitality/ what-apoe-means-for-your-health.*

[16] Ward A., Crean S., Mercaldi CJ., Collins JM., Boyd D., Cook MN., et al. Prevalence of apolipoprotein e4 genotype and homozygotes (APOE e4/4) among patients diagnosed with Alzheimer's disease: A systematic review and meta-analysis. *Neuroepidemiology*, 38:1–17, 2012.

[17] Chouraki V., Seshadri S. *Advance Genetics*, 87:245–94, 2014.

[18] Spinney L. Alzheimer's disease: The forgetting gene. *Nature*, 510:26-8, 2014.

[19] Lynn M. Bekris, Steven P. Millard, Nichole M. Galloway, Simona Vuletic, John J. Albers, Ge Li, Douglas R. Galasko, Charles DeCarli, Martin R. Farlow, Chris M. Clark, Joseph F. Quinn, Jeffrey A. Kaye, Gerard D. Schellenberg, Debby Tsuang, Elaine R. Peskind, and Chang-En Yua Multiple SNPs Within and Surrounding the Apolipoprotein E Gene Influence Cerebrospinal Fluid Apolipoprotein E Protein Levels *Journal of Alzheimers Disease.*, 13(3): 255–266, 2008.

[20] Sleegers K., Lambert JC., Bertram L., Cruts M., Amouyel P., Van Broeckhoven C. The pursuit of susceptibility genes for Alzheimer's disease: progress and prospects. *Trends Genet.* 26:84–93, 2010.

[21] W. Gregory Feero, M.D., Ph.D., Alan E. Guttmacher, M.D. Genomewide Association Studies and Assessment of the Risk of Disease. *Genomic Medicine*.363:166-76, 2010.

[22] Thomas A. Pearson, Teri A. Manolio. How to Interpret a Genome-wide Association Study. *JAMA*. 299(11):1335, 2008.

[23] Green RC., Cupples LA., Go R., Benke KS., Edeki T., Griffith PA. Risk of dementia among white and African American relatives of patients with Alzheimer disease. *JAMA*, 287:329–36, 2002.

[24] Robert B. Northrop, Anne N. Connor. Introduction to Molecular Biology, Genomics and Proteomics for Biomedical Engineers *CRC Press* 155, 2008.

[25] Anthony J.Brookes, The essence of SNPs, *Gene* , 234:177-186, 1999.

[26] K.C.ChengS, R.KatzA., Y.LinX, XinY.Ding. Chapter Four - Whole-Organism Cellular Pathology: A Systems Approach to Phenomics, *Advances in Genetics* , 95:89-115, 2016.

[27] K.C.ChengS, R.KatzA., Y.LinX, XinY.Ding. Genotyping, gene genealogies and genomics bring fungal population genetics above ground, *Trends in Ecology and Evolution* , 13:444-449, 1998.

[28] Saykin AJ., Shen L., Foroud TM., et al. Initiative biomarkers as quantitative phenotypes: Genetics core aims, progress, and plans, *Alzheimer's dementia: the journal of the Alzheimer's Association* 6:265-273, 2010.

[29] L.Silver. Encyclopedia of Genetics, *Academic Press*, 37, 2001.

[30] Renaud La Joie, Audrey Perrotin, Vincent de La Sayette, Stéphanie Egret, Loïc Doeuvre, Serge Belliard,Francis Eustache ,Béatrice Desgranges, Gaël Chételat Hippocampal subfield volumetry in mild cognitive impairment, Alzheimer's disease and semantic dementia, *NeuroImage: Clinical*, 3:155–162, 2013.

[31] Jack CR., Jr, Petersen RC., Xu Y., O'Brien PC., Smith GE., et al. Rates of hippocampal atrophy correlate with change in clinical status in aging and AD, *Neurology*, 55: 484–489, 2000.

[32] Du AT., Schuff N., Chao LL., Kornak J., Jagust WJ., et al. Age effects on atrophy rates of entorhinal cortex and hippocampus, *Neurobiology of Aging*, 27:733–740, 2006.

[33] Lambert, Jean-Charles et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimers disease, *Nature genetics*, 45,12:1452–1458, 2013.

[34] PLINK v1.07: Purcell S., Neale B., Todd-Brown K., Thomas L., Ferreira MAR., Bender D., Maller J., Sklar P., Bakker PIW., Daly MJ., Sham PC. PLINK: a toolset for whole-genome association and population-based linkage analysis, http://pngu.mgh.harvard.edu/purcell/plink/ *American Journal of Human Genetics* 81, 2007.

[35] Marees AT., de Kluiver H., Stringer S. et al. A tutorial on conducting genomewide association studies: Quality control and statistical analysis, *International Journal of Methods Psychiatry Results* 27, 2018.

[36] Jun Soo Kim., Igal Szleifer. Chapter Four - Crowding-Induced Formation and Structural Alteration of Nuclear Compartments: Insights from Computer Simulations, *Elsevier: International Review of Cell and Molecular Biology* ,307:73-108, 2014.

[37] Yiping Shen, Xiaohong Gong. 1 - Experimental Tools for the Identification of Specific Genes in Autism Spectrum Disorders and Intellectual Disability, *Neuronal and Synaptic Dysfunction in Autism Spectrum Disorder and Intellectual Disability*, 3-12, 2016.

[38] G. R. Abecasis, L. R. Cardon, W. O. C. Cookson. A General Test of Association for Quantitative Traits in Nuclear Families. *Human Genetics*, 66:279–292, 2000.

[39] Avik R., Kalipada P. Ankyrin repeat and BTB/POZ domain containing protein-2 inhibits the aggregation of alpha-synuclein: Implications for Parkinson's disease. *FEBS Letters*, 21:3567-3574,2013.

[40] Gaurav Kumar et al. Refinement of Schizophrenia GWAS Loci using Methylome-wide Association Data *Human genetics*, 1:77-87, 2015.

[41] P. Armitage, G.Berry. Statistical Methods in Medical Research. Science, 3rd Edition, 157:168, 1994.

[42] Yan Holtz. Manhattan plot in R: a review, *R graph gallery.*

[43] Research Data Services + Sciences. Understanding Q-Q Plots, *University of Virginia Library* .

[44] Janis E. Wigginton, David J. Cutler, Gonçalo R. Abecasis, A Note on Exact Tests of Hardy-Weinberg Equilibrium, *Human Genetics*, 76:887–883, 2005.

[45] John Attia, Ammarin Thakkinstian, Patrick McElduff, Elizabeth Milne, Somer Dawson, Rodney J. Scott, Nicholas de Klerk, Bruce Armstrong, John Thompson. Detecting Genotyping Error Using Measures of Degree of Hardy-Weinberg Disequilibrium, *Statistical Applications in Genetics and Molecular Biology*, 9(1):5, 2010.

[46] Steven G. Potkin, Guia Guffanti, Anita Lakatos, Jessica A. Turner, Frithjof Kruggel, James H. Fallon, Andrew J. Saykin, Alessandro Orro, Sara Lupoli, Erika Salvi, Michael Weiner, Fabio Macciardi. Hippocampal Atrophy as a Quantitative Trait in a Genome-Wide Association Study Identifying Novel Susceptibility Genes for Alzheimer's Disease, *PLOS One*, 4:8, 2009.

[47] Miguel E. Renterıa, Adrian Cortes, Sarah E. Medland. Using PLINK for Genome-Wide Association Studies (GWAS) and Data Analysis, *Genome-Wide Association Studies and Genomic Prediction* 8:193–213, 2013.

[48] Shaun Purcell. GWAS Tutorial with PLINK and Haploview. *Partners Healthcare* .

[49] John W.Holloway, Susan L.Prescott. Chapter 2 - The Origins of Allergic Disease, *Middleton's Allergy Essentials* 2: 29-50, 2017.

[50] Ott J., Macciardi F., Shen Y., Carta M.G., Murru A., Triunfo R., Robledo R.,Rinaldi A, Contu L. , Siniscalco M. Pilot Study on Schizophrenia in Sardinia, *Human Heredity* 70:92–96 , 2010.

[51] Hae-Young Kim. Statistical notes for clinical researchers: Chi-squared test and Fisher's exact test, *Restorative Dentistry and Endodontics* 42(2): 152–155, 2017.