# DETERMINATION OF PERFORMANCE PARAMETERS FOR AHP BASED SINGLE NUCLEOTIDE POLYMORPHISM (SNP) PRIORITIZATION APPROACH ON ALZHEIMERS'S DISEASE DATA

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

ONAT KADIOĞLU

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
IN
BIOINFORMATICS

SEPTEMBER 2011

Approval of the Graduate School of Informatics

_____

Prof. Dr. Nazife Baykal
Director

I certify that this thesis satisfies all the requirements as a thesis for the degree of Master of Science.

_____

Assist.Prof.Dr. Didem Gökçay
Head of Department

This is to certify that we have read this thesis and that in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of Master of Science.

_____

Assist. Prof. Dr. Yeşim Aydın Son
Supervisor

Examining Committee Members

Assoc. Prof. Dr. Tolga Can (METU, CENG) _____

Assist. Prof. Dr. Yeşim Aydın Son (METU, II) _____

Assist. Prof. Dr. Zeynep Kalaylıoğlu (METU, STAT) _____

Dr. Gürkan Üstünkar (İzmir Economy University, ISE) _____

Assist. Prof. Dr.Tülin Yanık (METU, BIO)_____

**I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.**

**Name, Last Name :** **Onat Kadıoğlu**

**Signature :**

# ABSTRACT

DETERMINATION OF PERFORMANCE PARAMETERS FOR AHP BASED SINGLE NUCLEOTIDE POLYMORPHISM (SNP) PRIORITIZATION APPROACH ON ALZHEIMER'S DISEASE DATA

Kadıoğlu, Onat

MSc., Bioinformatics Program

Supervisor: Assist. Prof. Dr. Yeşim Aydın Son

September 2011, 91 pages

GWAS mainly aim to identify variations associated with certain phenotypes or diseases. Recently the combined p-value approach is described as the next step after GWAS to map the significant SNPs to genes and pathways to evaluate SNP-gene-disease associations. Major bottleneck of standard GWAS approaches is the prioritization of statistically significant results. The connection between statistical analysis and biological relevance should be established to understand the underlying molecular mechanisms of diseases. There are few tools offered for SNP prioritization but these are mainly based on user-defined subjective parameters, which are hard to standardize. Our group has recently developed a novel AHP based SNP prioritization algorithm. Beside statistical association AHP based SNP prioritization algorithm scores SNPs according to their biological relevance in terms of genomic location, functional consequence, evolutionary conservation, and gene-disease association. This allows researchers to evaluate the significantly associated SNPs quickly and objectively. Here, we have investigated the performance of the AHP based prioritization as the next step in the utilization of the algorithm in comparison to the other available tools for SNP prioritization. The user-defined parameters for AHP based prioritization have been investigated and our suggestion on how to use these parameters are presented. Additionally, the GWAS results from the analysis of two different sets of Alzheimer Disease Genotyping data with the newly proposed AHP based prioritization and the integrated software,

METU-SNP, it was implemented, is reported and our new findings on the association of SNPs and genes with AD based on this analysis is discussed.

Keywords: GWAS, SNP Prioritization, Biomarker Discovery, Analytic Hierarchy Process, Alzheimer's Disease

# ÖZ

ANALİTİK HİYERARŞİ SÜRECİNE DAYALI TEK NÜKLEOTİD POLİMORFİZMİ ÖNCELİKLENDİRME YAKLAŞIMI PERFORMANS PARAMETRELERİNİN ALZHEIMER HASTALIĞI VERİSİ İÇİN BELİRLENMESİ

Kadıoğlu, Onat

Yüksek Lisans, Biyoenformatik Bölümü

Tez Yöneticisi: Assist. Prof. Dr. Yeşim Aydın Son

Eylül 2011, 91 sayfa

Genom boyutunda ilişkilendirme çalışmaları (GWAS) genel olarak biyolojik çeşitliliğin araştırılması ve çeşitli hastalıklarla ilişkilendirilmesiyle ilgilidir. GWAS den sonraki aşama olarak tanımlanan birleşik p değeri anlamlı olarak bulunan SNPlerin genlerdeki yerlerini belirlemek ve daha sonra SNP-gen-hastalık ilişkisini saptamak için kullanılabilir. İstatistiksel olarak anlamlı sonuçları önceliklendirmek mevcut GWAS analizlerinin başlıca eksikliklerdendir. Hastalıkların moleküler mekanizmalarını daha iyi anlayabilmemiz için istatistiksel analiz ve SNP lerin biyolojik anlamlılıkları arasındaki bağlantı daha sağlam bir şekilde kurulmalıdır. SNP önceliklendirmesi için geliştirilmiş az sayıdaki yazılımlar standardizasyonu güç olan kullanıcı tanımlı öznel paramatrelere dayalı uygulamalardan öteye geçememektedirler. AHP (Analitik Hiyerarşi Süreci) tabanlı yapılandırılmış SNP önceliklendirmesi için grubumuz tarafından geliştirilen algoritma; SNPlerin biyolojik anlamlılıklarının genomik lokasyona, fonksiyonel sonuçlara, evrimsel korunmaya ve gen-hastalık ilişkilendirilmesine göre skorlandırılmalarına dayanmaktadır. Böylece istatistiksel olarak anlamlı SNP ler araştırmacılar tarafından kolayca nesnel olarak değerlendirilmiş olur ve yüksek skora sahip olanlar, onaylama ve daha sonraki muhtemel uygulamalar için kullanılabilir. Bu çalışmada AHP tabanlı önceliklendirme yaklaşımının performansını, algoritmanın uygulanmasında sonraki adım olarak diğer SNP önceliklendirme metotlarıyla karşılaştırarak değerlendirdik. AHP tabanlı önceliklendirme için

kullanıcı tanımlı parametreler araştırıldı ve bu parametrelerin nasıl kullanılması gerektiği sunuldu. Ek olarak, iki adet Alzheimer Hastalığı (AD) Genotipleme datasının yeni oluşturulan AHP tabanlı önceliklendirme yaklaşımıyla ve bu yaklaşımın uygulandığı METU-SNP uygulaması ile yapılan analizlerinden elde edilen GWAS sonuçları sunuldu. Bu analizler ışığında SNPlerin ve genlerin AD ilişkisi ile ilgili yeni bulgular da ele alındı.

Anahtar Kelimeler: GWAS, SNP Önceliklendirmesi, Biyolojik Gösterge Bulma, Analitik Hiyerarşi Süreci, Alzheimer Hastalığı

*To My Family*

# ACKNOWLEDGEMENTS

First of all, I should thank to my supervisor Assist. Prof. Dr. Yeşim Aydın Son. She is a very supporting, well intentioned and encouraging instructor that every student wants to work with. She is a role model for me while pursuing my academical career.She has supported me from the very beginning of this study. GTalk sessions also helped a lot throughout thesis preparation process. I should also thank to Dr. Gürkan Üstünkar, who has developed the METU-SNP software during his Ph.D. project at METU IS department. He answered my technical questions and directed me while using the software.

I am grateful to Assist. Prof. Dr. Alptekin Temizel and NVIDIA Cuda Teaching Center at METU for giving me chance to use NVIDIA C2070 Tesla Computer while performing AHP analysis. It helped me a lot and saved my time throughout the study.

I am also grateful to the administrative staff of METU Informatics Institute; Sibel Gülnar, Necla Işıklar, Ali Kantar and Hakan Güler for their help.

I am also grateful to my supporting family. I thank my parents for devoting their life on me.

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# PREFACE

In this research, we have tested our recently developed Single Nucleotide Polymorphism (SNP) prioritization system based on Analytic Hierarchy Process (AHP) on two different independent Alzheimer Disease genotyping data with respect to biological relevance in comparison to SPOT, which is one of the most widely used web-based SNP prioritization tool. Performance measures have been determined for the AHP based SNP prioritization approach (METU-SNP) in various aspects such as sensitivity, specificity and biological relevance measures.

**Thesis Organization**

This thesis is composed of four main chapters. Brief contents are given below:

**Chapter 1** presents the biological background and introduces genome wide association studies with an emphasis on the SNP prioritization process.

It is composed of;

- Biological background
- Genome wide association studies (GWAS), GWAS bottlenecks
- Genetic variations in the human genome
- SNP prioritization,available tools and analytic hierarchy process (AHP) based SNP prioritization approach
- Complex diseases, Alzheimer disease (AD)

**Chapter 2** provides a literature review on Alzheimer's Disease and GWAS findings. It is composed of;

- Literature review on AD
- Previously reported GWAS results for AD
- AD databases
- AD associated gene list

**Chapter 3** reports our results from the analysis of two different sets of Alzheimer Disease Genotyping data with the newly proposed AHP based prioritization and the integrated software, METU-SNP. It is composed of;

- Analysis of two independent AD genotyping data with METU-SNP and AHP based prioritization results
- Comparison of AHP based prioritization with SPOT in terms of biological relevance

**Chapter 4** reports classification performance of METU-SNP in terms of specificity, sensitivity measures in various p-value thresholds to determine the pre-prioritization cutoff value.

It is composed of;

- METU-SNP classification measures in different p-value thresholds

**Chapter 5** provides estimation of AHP score cutoff for two independent AD genotyping data depending on biological relevance measures. Additionally, the user-defined parameters for AHP based prioritization and our suggestion on how to use these parameters, and the provided AHP score is discussed.

It is composed of;

- AHP score distributions of the AD genotyping data after SNPs are AHP prioritized
- Biological relevance measures of SNP list based on AHP score ranking
- User defined parameters of AHP based prioritization
- Classification performance in different AHP score ranges
- Utilization of AHP prioritization score

**Chapter 6** Conclusion and Future Work

- Dıscussion of findings from AD data
- Discussion of AHP parameter and cut-off score estimation
- Future Work
- Conclusion

# CHAPTER 1

# INTRODUCTION

## 1.1 Biological Background

The genome defines the hereditary nature of each organism. Deoxyribo nucleic acid (DNA) which is the building block of genome carries the genetic information. Complete set of this information in an organism is called its genotype. There are 4 different bases in DNA; Adenine (A), Tymine (T), Guanine (G) and Cytosine (C). Purine bases are A and G whereas pyrimidine bases are C and T. Hydrogen bonds are established between A-T and G-C forming the double helix structure of DNA. Those bonds can be broken and reformed which is essential in DNA replication. Double helix structure involves backbones at opposite edges and paired bases meet in the middle as visualized in Figure 1.1 [1].

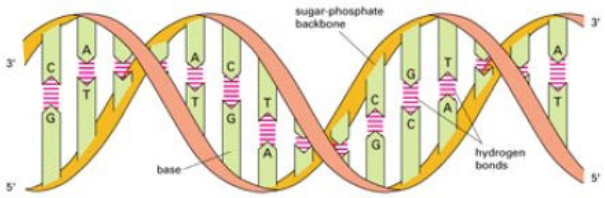



**Figure 1.1 DNA structure [1]**

Central dogma of molecular biology involves transcription (RNA produced from DNA), RNA processing (mature RNA formation) and translation (protein coding from mature RNA) respectively. Gene expression process begins with transcription where RNA complementary to DNA sequence of the coding region is coded. Template DNA strand is utilized during transcription. Processing of RNA and splicing is followed by RNA production. In this step, coding regions are joined together for the protein synthesis. Non-coding parts (introns) are excised out and coding parts (exons) are combined together during processing and splicing of RNA. Translation is the synthesis of protein from the mature RNA. Aminoacids (building blocks of proteins) are coded depending on three base sequences in RNA (codons). There are twenty aminoacids with the exception of some modified aminoacids utilized by organisms living on extreme conditions. Some aminoacids are coded via multiple codons whereas some are coded by only one codon (methionine is coded by only AUG codon and every protein begins with

methionine, AUG is referred as the start codon). Moreover there are three stop codons (UAG, UGA and UAA) indicating the termination of translation [1].

The genome of eukaryotic organisms are densely packed into chromosome structures prior to the cell division. Each chromosome consists of a linear array of genes located at a particular location, referred as the genetic locus. A gene is a unique DNA sequence directing the production of a specific protein specialized in a specific function influencing a particular characteristic in an organism. Alleles of a gene can be defined as the different forms that are found at the corresponding locus. Diploid organisms have one paternal (inherited from father) and one maternal (inherited from mother) sets of chromosomes, likewise each gene has two copies of paternal and maternal alleles [1].

Human genome consists of 23 pairs of chromosomes, 46 in total, where each half inherited maternally or paternally. The crossing over and recombination of chromosomal pairs that occurs during the meiotic division are the sources of genetic variance in the production of egg and sperm. Egg and sperm cells are haploid and carry 23 chromosomes and zygote formation leads to a diploid cell production. Mitotic divisions at the zygote contribute to the development of the progeny.

Human genome is around 3.3 billion base pairs in length and coding regions are only about 3 percent of the whole genome, and the rest is known as noncoding regions, which have no annotated functions yet.

Differences in the genome sequence of species are referred as variations and there are various varations having impact on biological function. Mutations are much rarely observed in a population compared to polymorphisms, which are observed in at least 1% of a population and they are the major causes of variations observed among individuals.

Mutations are stable changes in DNA sequence and their lethality is determined mostly by their genomic location. Mutations at coding regions are potentially more severe than mutations on the non-coding regions of genome. Lethal mutations are not observed frequently within a population since they are rarely inherited. Mostly mild mutations are observed within a population and contribute to disease formation under certain environmental conditions, also giving rise to different traits observed within a population [2].

Similar to mutations, genetic polymorphisms are observed when multiple genotypes at a locus coexist. Polymorphisms are changes in DNA sequence which are observed in more than 1 percent of a population and usually involve milder effects compared to mutations. By definition, changes at single nucleotide that are observed in more than 1 percent in a population are referred as single nucleotide polymorphisms (SNP). The location in the genome where a SNP, copy number variation, insertion or deletion occur is referred to as an allele [2]. The dominant allele with the highest frequency within a population is called the major allele, while less frequent

observed ones are called the minor alleles. Therefore, the genotype of an individual is referred as the combined allele information for a particular locus.

The alleles on the same chromosome tend to be inherited together rather than independent assortion, predicted by Mendelian laws, due the genetic linkage event caused by the organization and packing of genes into chromosomes. Genetic linkage is measured by the unit of recombination, which is used for construction of genetic maps. Distance between mutations and variations in terms of recombination frequencies is determined by genetic mapping. Recombination frequencies do not represent the actual physical distances since frequencies can be distorted relative to the physical distance between sites [2]. There are some markers such as restriction fragment length polymorphisms (RFLPs) and SNPs which can be referred as the basis for linkage maps. SNPs are widely preferred for linkage studies and genetic mapping since they are widespread in the human genome (there is a SNP ~every 100-300 bp). Newly discovered disease genes and associations can be identified via locating them between the nearest SNPs. The frequency of polymorphism and the unique combination of SNPs or RFLPs in a specific region is referred as the **haplotype**. Figure 1.2 represents an example of a haplotype involving 3 SNPs [2].

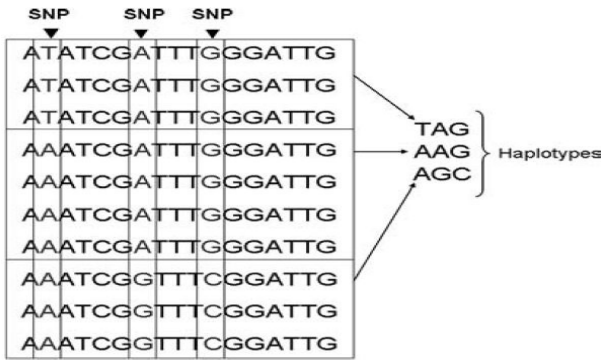



**Figure 1.2 Example of determination haplotype of various alleles of a genomic locus [2]**

The concept of a haplotype was originally introduced to describe the genetic constitution of the major histocompatibility locus, a region specifying proteins of importance in the immune system. The concept now has been extended to describe the particular combination of alleles or restriction sites (or any other genetic marker) present in a defined area of the genome and they are inherited as single haplotype blocks [3]. SNPs are conserved in a genome usually within haplotype blocks. If we assume that there is a gene with two SNPs and two alleles of each SNP (A and B) there are four possible combinations (A-A, A-B, B-A, B-B). In most cases, SNPs occur with a different frequency than would be expected to occur only by chance from a random haplotype distribution. Non-random associations of alleles at two or more loci constitute the phenomenon of Linkage Disequilibrium (LD). Linkage disequlibrium is defined by the allelic association between SNPs. High LD SNPs have higher probability to be inherited together than

lower LD SNP pairs. Linkage disequlibrium in a human chromosome 20 region is visualized in Figure 1.3, where darker regions indicate strong correlations and thus high LD among SNPs [2].

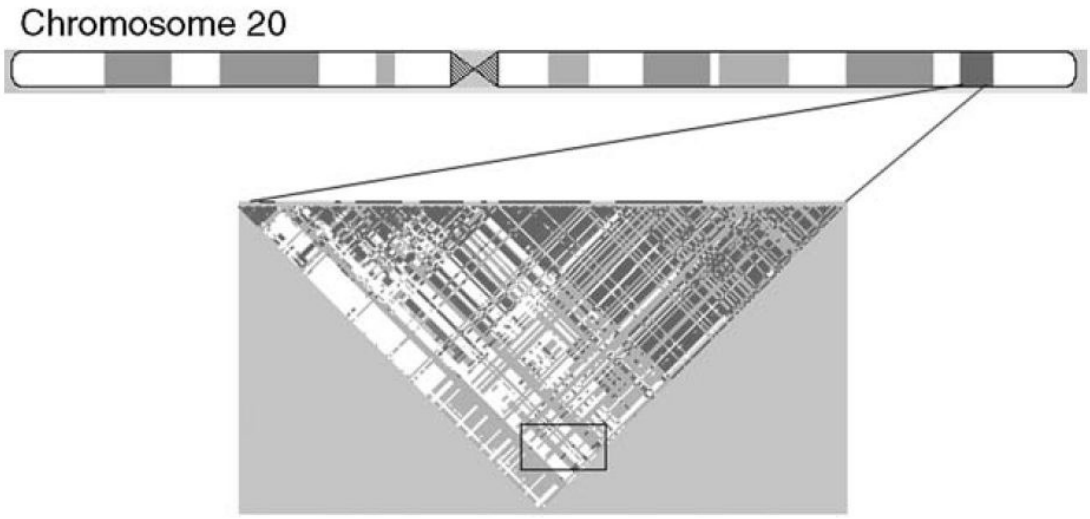



**Figure 1.3 Linkage disequlibrium in a region on human chromosome 20 [2]**

## 1.2 Genome Wide Association Studies (GWAS)

Genome wide association studies (GWAS) can be defined as genetic association studies in which the genetic marker density and the extent of linkage disequilibrium is sufficient to cover a large proportion of the common variation in the population under study [4]. The sample size of people provides sufficient power to detect variants of modest effect [4]. GWAS mainly focuses on identification of significant variations that can be associated with certain disease and phenotypes among individuals in population in a holistic and agnostic manner. They are hypothesis-free studies and thus unbiased since they focus on entire genome to find associative variations. SNPs are the most widely referred variations in GWAS due to their low genotyping cost and abundance in the genome. Resources such as the HapMap Project and the 1000 Genomes Project provides a catalog of the SNPs in the genome [5].

After DNA genotyping is performed for control and case subjects, allele frequencies are compared and significantly different variations between cases and controls are aimed to be identified. Large sample size on the order of thousands is usually preferred in GWAS to achieve a high statistical power (sensitivity). Statistical tests are then applied in order to identify SNPs that demonstrate allele frequency differences between cases and controls. Identification of candidate associative variations can be followed by new strategies to detect, treat and prevent the disease.

GWAS are widely referred since 2005 to discover genetic variations that contribute to various complex diseases, such as asthma, cancer, diabetes, heart disease and mental illnesses. After completion of the Human Genome Project in 2003 and the International HapMap Project in

2005, the number of studies aiming to identify associative genetic variations with various diseases have dramatically increased. Since GWAS serve as a promising strategy for the identification of genetic variations associated with various phenotypes, they have become a widely preferred approach in identifying genetic determinants and biomarkers of complex phenotypes and common diseases.

Many tools including computerized databases containing the reference human genome sequence helped researchers for the analysis of whole-genome samples for genetic variations. Since a considerable cost reduction in sequencing a human genome is being achieved via various next generation seuquencing platforms, in the near future patients will probably be provided with individualized information about their certain disease risks depending on the genetic variations they carry. The genetic makeup of an individual will determine the treatment strategies and even the doses of the drugs to be prescribed. This individualized information provides the approach of personalized medicine.

Since the first successful GWAS in 2005, over 1200 GWAS have been reported according to the NIH GWAS catalog as can be seen from the Figure 1.4 [6]. Chromosomes are visualized in the figure and circles with different colors mapped on chromosomes represent associative variations with different diseases and traits.

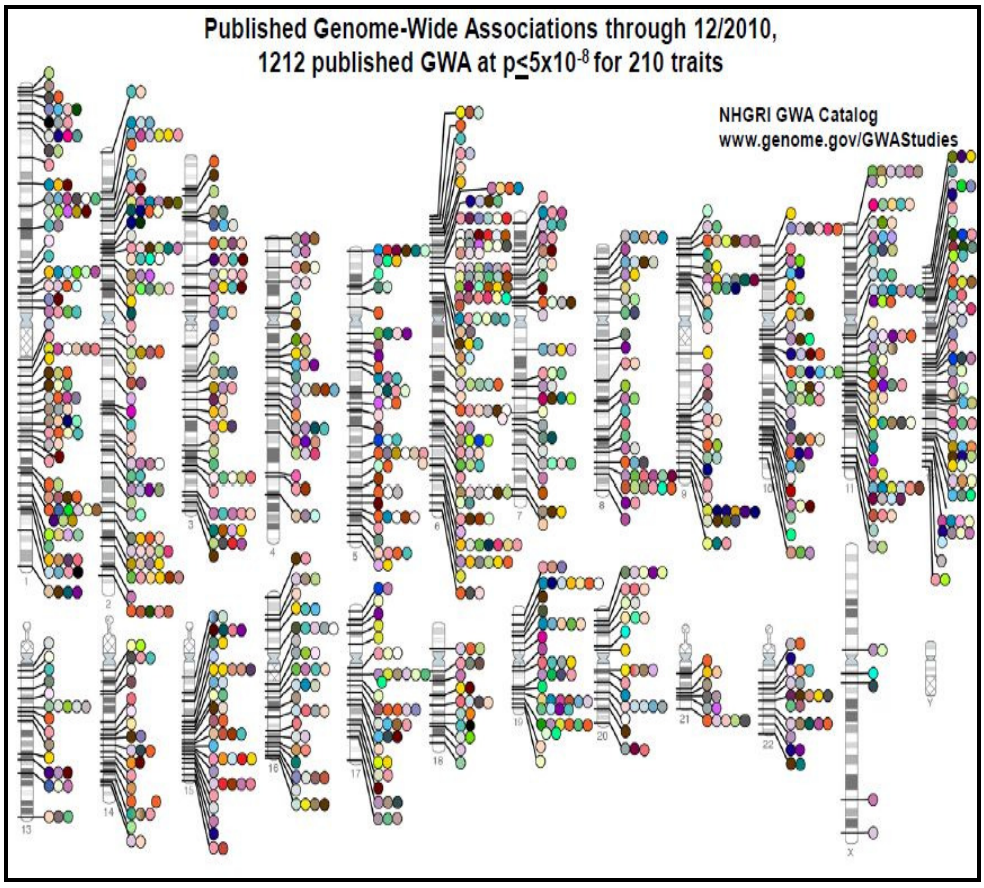



**Figure 1.4 Published GWA studies and the mapping of associative variations [6]**

Analysis of GWAS results depend on two major steps; **statistical** and **biological** approaches. The former approach involves p-value computation to find out statistically significant variations between control and case subjects. Allele frequencies of variations that are significantly different among cases than controls are referred as candidate associative variations for a certain disease or trait. p-value can be defined as a statistical measure that indicates the probability of a certain event to occur just by chance, smaller the p-value higher the significance of the event. In GWAS, variations that have p-values smaller than 0.05 are usually referred as significant and potential candidates to be associated with certain diseases or phenotypes. A widely used software developed by Shaun Purcell at the Center for Human Genetic Research, **PLINK** provides a toolkit involving various statistical tests for single-locus analysis, haplotype analysis and allelic-based interaction analysis [7]. Statistical analysis might also involve 'second wave GWAS' strategy where the combined p-values for genes can be used for identification of enriched genes and pathways significantly associated with disease [8]. After the detection of significant variations such as SNPs and their mapping to genes, "combined p-value" can be computed for genes via Fisher's tests. *Fisher's combination test* can be used to compute combined p-value for a gene, whereas *Fisher's exact test* can be used to compute combined p-value for a pathway[8].

Fisher's combination test to combine p-values of all K independent SNPs within the gene is performed as follows;

$$Z_F = -2\sum_{i=1}^{K} \log P_i$$

which follows a $\chi^2_{(2K)}$ distribution.

Fisher's exact test is performed to combine p-values of genes within pathways. p-value of observing k-significant genes in the pathway is calculated as follows;

$$P = 1 - \sum_{i=0}^{K} \frac{\binom{S}{i}\binom{N-S}{m-i}}{\binom{N}{m}}$$

where total number of genes of interest is N, number of genes that are significantly associated with disease according to Fisher's combination test is S, number of genes in the pathway is m, and the number of significantly associated genes in the pathway is K [8].

Gene and pathway based analysis in a second wave GWAS involves independent p-values of single SNPs in the gene combined into an overall p-value for the gene and independent p-values of a single gene in the pathway into an overall p-value for the pathway. Fisher's method for combination might yield unreliable results when there are large correlations among SNPs in a gene. Gene and pathway-based GWAS that consider correlations among the SNP and genes might be an effective strategy for the second wave GWAS analysis and will probably be carried out in the near future [8]. The combined p-value approach for GWAS involving several disorders

such as bipolar disorder (BD), coronary artery disease (CAD), Crohn's disease (CD), hypertension (HT), rheumatoid arthritis (RA), type I diabetes (T1D), type II diabetes (T2D), Parkinson's disease (PD), age-related eye disease (AREDS) and Amyotrophic lateral sclerosis (ALS) identified genes that include individually significant SNPs and also new genes containing SNPs with small disease risks but have jointly significant effect and strong association with diseases [8].

Statistical association on its own is not adequate to identify predisposing loci or genes effectively due to the limited power and information it provides together with high dimensional data and multiple testing problems. Achievement of a reliable GWAS heavily depends on biological knowledge to correctly prioritize results for downstream genetic studies. Next step after statistical analysis is search for biological relevance where statistically significant variations are evaluated depending on different biological properties such as their genomic location, evolutionary conservation and gene association. Being the most widely referred variations in GWAS, SNPs have various biological features relevant to disease risks. For instance, SNPs located in non-synonymous coding region involve a higher probability to be associated with a disease than SNPs located within the intron. Moreover, some non-gene features of SNPs might also yield clues related to disease risk. The connection between statistical analysis and biological relevance for SNP biomarkers should be established more firmly to be able to understand the underlying molecular mechanisms of a disease. Recent studies point out that conservation, natural selection and microRNA binding are contributing factors to human disease susceptibility [9].

GWAS involve various approaches and terms, brief definitions of those are provided in Appendix A [10 , 11, 12] .

### 1.2.1 Current Challenges of GWAS

GWAS should meet three essential elements to be counted as a reliable study [13]:

- sufficiently large sample size from a population that effectively provide genetic information regarding the research question
- polymorphic alleles covering the whole genome and those alleles should be efficiently genotyped
- statistically powerful analytic methods that can be utilized for the identification of the genetic associations in an unbiased fashion

GWAS failing to achieve the criteria stated above are unlikely to yield informative results. Moreover, traditional GWAS heavily depend on statistical analysis are inadequate to identify associative variations in the human genome. Since most widely investigated variations in GWAS are single nucleotide polymorphisms (SNPs) rather than duplications, deletions, insertions or copy number variations, major bottleneck of current GWAS is the prioritization of

statistically significant and determination of biologically relevant SNP associations after the statistical analysis. Small p-value choice for SNPs might not be optimal in cases when the joint action of multiple SNPs within a gene involves more variance than the most significant SNP. Small p-value choice approach might also lead to biases of favouring large extensive pathways and genes with greater numbers of SNPs. Another drawback of GWA studies is varying levels of genome coverage across samples [14].

Genetic variations associated by GWAS can only explain a small proportion of the genetic risks associated with the complex diseases. New strategies and approaches are required to compansate this lack of explanation. The approach to investigate variations individually might lead to inadequate results. One reason for the inadequacy is that genetic variants with small individual effect sizes but jointly significant genetic effects would be missed by single-SNP analysis. Consequently, identified genetic variants involve a small fraction of heritability for most studied traits. GWAS that focus on discovery of biological pathways rather than prediction of individual risk loci associated with polygenic traits and diseases are more preferable. Detection of SNPs that have marginally weak but jointly strong effects is a difficult task for GWAS focusing on invidiual SNP analysis. Jointly analyzing SNPs within the same biological pathway compansates the individual SNP analysis, providing new insights to the understanding of complex human traits.

A preprocess step to filter out SNPs that are unable to achieve the different threshold values such as Hardy-Weinberg Equilibrium, Minor Allele Frequency is essential since using all available SNPs per gene cause computational challenges in addition to significant amounts of noise into the analysis [15].

Detection of associations with common SNPs becomes harder for a GWA study when the phenotypes are poorly or inconsistently defined, controls are poorly screened for exclusion of disease. Limited statistical power and inadequate environmental data might lead to disruption in the detection of gene–environment interactions [16]. Another strategy to increase statistical power is using nonrisk cases and risk controls in the study. Power increases due to enrichment of disease favouring alleles in nonrisk cases while risk controls are enriched for protective alleles [17].

Limitations and bottlenecks of GWAS to detect candidate variants associated with complex diseases can be summarized as follows. Genetic variants that involve a insignificant risk of disease individually but have a considerable contribution when considered jointly with other variations, would be probably missed in the ‘most significant SNPs/genes’ approach. Small sample size is another important problem in GWAS since those variants that confer a larger effect may not always be tested. Complex diseases are thought to be caused by multiple risk genes mutually together with various environmental factors rather than most significant genes only. Their pathogenesis involve dysfunction of several metabolic pathways [18].

### 1.2.2 Genetic Variations In The Human Genome

While GWAS usually target identification of associative single nucleotide polymorphisms (SNPs), there are various different types of variations widespread throughout the human genome. Figure 1.5 and 1.6 [19,20] provide a visualization of the genetic variations.

*Copy Number Variations (CNV)*

Duplications, deletions and inversions are classified as copy number variations (CNV), most of them are known to be tagged by SNPs and therefore focusing on those CNVs directly is unlikely to identify many new variants, targeting rare CNVs and structural variants might yield valuable results. Structural variation has not been as deeply studied as SNPs because their detection is less accurate, biological confirmation is costly, and smaller copy number variants (<100 kb) are not very reliably detected. Genetic risks for common disorders and complex diseases cannot be fully explained by common SNPs. Rare variants should be investigated more firmly to identify further associations. As mentioned in the GWAS terms, rare variant is usually defined by a frequency lower than 1%. One approach to find out rare variants might be sequencing a chromosomal region from many people and focusing on that region only. Indeed some rare structural variants have already been identified as a risk factor for diseases. For instance, Hras-1 VNTR mini satellite rare alleles are found to be associated with various cancers and uncommon translocations such as t(11:22)(q23:q11) are found to be associated with breast cancer. Another example is DiGeorge syndrome caused by large deletions on chromosome 22q11, and 20% of patients with this syndrome also develop schizophrenia. Therefore, structural variants that are individually rare but involve clustering within specific regions are valuable resources to identify disease associations in the human genome. Genotyping of rare variants is not an easy task with the existing technology however new arrays are being designed for this purpose. 1000 Genomes project contributes to identification of such variants. Some of the associations which are currently attributed to common variants might be actually caused by some rare variants. Next-generation sequencing technologies such as Ion Torrent, Oxford Nanopore, IBM Transistor, PacBio SMRT (single molecule, real time) providing affordable whole-genome sequencing, will lead a shift in the search for rare variants from genotyping arrays to whole-genome sequencing [21].

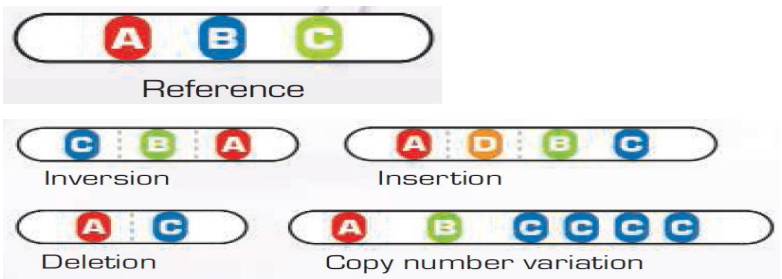



**Figure 1.5 Genetic variations in the human genome (inversion,insertion,deletion,copy number variation) [19]**

*Single Nucleotide Polymorphism (SNP)*

If a variation is observed in more than 1% of a population, they are referred as **polymorphisms** and the substitution of one single nucleotide for another at a homologous site in a population is called **single nucleotide** polymorphism, which are the most widely observed polymorphisms in the human genome (~90%). National Center for Biotechnology Information's (NCBI) current SNP variation database (dbSNP build 131) holds about 30 million validated SNPs within the human genome. SNPs exist both within coding regions and non-coding regions of the entire human genome. Different types of SNPs are summarized at Table 1.1 [22].

**Table 1.1 Single Nucleotide Polymorphism Types In The Human Genome [22]**

| **CODING REGION** | **Definition** |
|---|---|
| *Non-synonymous* | results in an aminoacid change,cause of most monogenic disorders such as cystic fibrosis and hemophilia |
| *Synonymous* | do not lead to aminoacid change |
| *Frameshift* | causes a frameshift |
| *Stop loss* | results in loss of a stop codon |
| *Stop gained* | results in gain of a stop codon |
| **NON-CODING REGION** | |
| *Essential splice site* | in the first or last 2 bp of an intron |
| *Splice site* | 1-3 bps into an exon or 3-8 bps into an intron |
| *Upstream* | within 5 kb upstream of the 5'end of a transcript |
| *Regulatory region* | in regulatory region annotated by NCBI or Ensembl |
| *5 ' UTR* | located within 5 ' UTR |
| *Intronic* | located within intron |
| *3'UTR* | located within 3'UTR |
| *Downstream* | within 5 kb downstream of the 3'end of a transcript |
| *Intergenic* | more than 5 kb away from a transcript |

In GWAS, an informative SNP subset which are referred as tag SNPs, are genotyped in case and control individuals. After the computation of tag SNP statistics, the genomic regions

that involve a linkage disequilibrium (LD) with the most significantly associated tag SNPs are expected to contain the causal polymorphisms [23].

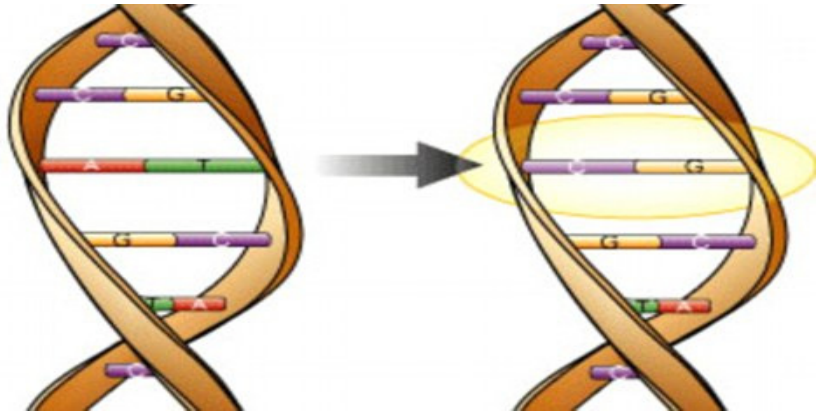


**Figure 1.6 Genetic variations in the human genome (single nucleotide polymorphism) [20]**

### 1.3 SNP Prioritization Approaches

GWAS yield highly promising results in terms of identifying regions associated with a variety of complex traits and diseases. Most of the studies point out SNPs as the likely causal variants that are in complete linkage disequilibrium and their prioritization is challenging since individually weak SNPs might involve a strong association jointly if it is in high LD with associative and causal SNPs. Although SNPs located at coding regions and cause a aminoacid change or at a regulatory region are highly potential candidates to be associated with a disease, in many complex diseases, the causal SNPs are found to be located within noncoding regions, intergenic regions which makes the prioritization of them difficult based on likely function [24].

Major aim of a biomarker discovery study is the identification of potentially significant differential variants across control and case groups. GWAS solely based on statistical analysis are prone to many false positive results, as simultaneous testing of hundreds of thousands of SNPs yields a high number of hits occurring by chance with a traditional p-value threshold of 0.05. Linkage disequlibrium is another factor for the inadequacy of traditional statistical analysis in GWAS. Considering p-values alone cannot provide a reliable identification of associative variants due to strong physical association between certain SNPs. Biological relevance should be strongly involved in GWA studies to acquire reliable results and informative associations during SNP prioritization. There are various approaches to prioritize SNPs after GWAS such as **meta analysis** and **pathway based analysis**.

#### 1.3.1 Meta Analysis

Pooling information from multiple GWAS to increase the statistical power and thus chances of finding true positives among the false positives is referred as meta analysis. Combining p-values via Fisher's method and converting the test statistics into z scores via

referring to odds ratios together with the regression coefficients are two major meta analysis methods utilized in GWAS. Meta analysis approach involves combination of different GWAS results rather than using the original data which can be computationally demanding [13].

### 1.3.2 Pathway Based Analysis

GWAS employing pathway based analyses are promising since they can identify many causal variants that cannot be identified with traditional statistical methods. Pathway based analysis is preferable since testing a few hundred pathways to identify the subsets of genes associated with diseases, eliminates the need for difficult task of huge multiple testing. For this reason, pathway based analysis can improve the power of GWAS, and identification of subset of genes in biologically relevant pathways associated with diseases or traits becomes easier. In addition, a gene which involves a number of SNPs with medium size effects might involve a increased effect after the information of SNPs are combined by the gene based score as a result of pathway based analysis [25].

Pathways are tested in terms of associations with diseases or traits and variations within the associated pathways are said to be potential causative variants even if they cannot reach genome wide significance level. For instance, some of the SNPs located at genes involved in the axon guidance pathway are found to be causal SNPs associated with Parkinson's Disease, even though none of them are found to reach the significance threshold. In pathway analysis approach, pathway heterogeneity is an important factor since disruptions in different pathways might lead to the same disorder. Moreover, the variants within those pathways are likely to differ despite affected individuals may share the same disrupted pathways.

Post-GWAS studies are likely to provide a clearer picture of the true role of common variants in common complex disorders and there are various tools for pathway enrichment analysis such as the Database for Annotation, Visualization and Integrated Discovery (DAVID) and Protein ANalysis THrough Evolutionary Relationships (PANTHER) [13].

## 1.4 SNP Prioritization Tools

There are various tools focusing on single nucleotide polymorphism prioritization following statistical analysis of GWAS. Biological information and functional properties of SNPs are integrated to the algorithms of those tools. However an adequate biological relevance level have not been achieved yet with those tools. In this part some of the most widely used ones will be introduced such as FastSNP, SNPLogic,SPOT, SNPinfo, SNPit and the recently developed AHP based SNP prioritization approach.

### 1.4.1 FastSNP

It is a webserver providing identification and prioritization of high risk SNPs depending on their phenotypic risks and deleterious functional effects. Risk determination requires access to

various biological databases and analytical tools, FastSNP retrieves biological information via 11 external web servers making it possible to perform updated querying. SNP prioritization based on phenotypic risks is essential due to the fact that only a small portion of them are causal and associative polymorphisms contributing to various disease phenotypes. SNPs can be cateogorized depending on their genomic location and corresponding functional effects like below;

- Nonsynonymous SNPs effecting protein structures via changing single amino acids
- SNPs located in transcription factor binding sites in promoter or enhancer regions can modulate transcriptional regulation
- SNPs in splice sites may disrupt alternative splicing regulation

FastSNP depends on a decision tree to determine risk factors of SNPs depending on solely their genomic location ignoring essential biological relevance points such as evolutionary conservation, gene and pathway association. Table 1.2 summarizes the SNP categorization of FastSNP and risk factors assigned to each class of SNPs [26].

It utilizes a decision tree approach to assign risk rankings for SNP prioritization. The tree structure used for the prioritization only depends on genomic location of SNPs and lacks the essential information such as evolutionary conservation, gene association and biological pathways. Depending on the functional effects, each SNP is assigned a risk factor and ranking is done accordingly.

**Table 1.2 FastSNP Single Nucleotide Polymorphism Functional Properties And Risk Factors [26]**

| Coding type | Function type | Possible effects | Risk (ranking) |
|---|---|---|---|
| Coding | Non-sense | Causes premature termination of an amino-acid sequence | Very high (5) |
| | Splicing regulation (abolishing protein domain) | Breaks the exonic splicing enhancer/silencer binding site in a coding sequence, leading to abolished protein domain | Moderate to high (3~4) |
| | Splicing regulation | Breaks the exonic splicing enhancer/silencer binding site in a coding sequence containing the same protein domains | Low to moderate (2~3) |
| | Mis-sense (non-conservative change) | Alters an amino acid in a protein to one with different structure characteristics | Moderate to high (3~4) |
| | Mis-sense (conservative change) | Alters an amino acid in a protein to one with similar structure characteristics | Low to moderate (2~3) |
| | Sense/synonymous | Does not alter an amino acid in a sequence | Very low (1) |
| Non-coding | Downstream with no known effect | No known effect | No known effect (0) |
| | Upstream with no known effect | No known effect | No known effect (0) |
| | Splicing site | Breaks a consensus splicing site sequence | Moderate to high (3~4) |
| | Promoter/regulatory region | Does not alter an amino acid, but can affect the level, location or timing of a gene expression | Very low to moderate (1~3) |
| | Intronic enhancer | Alters a binding site of a transcription factor in an intronic region | Very low to low (1~2) |
| | Untranslated | Changes an UTR in a sequence | No known effect to very low (0~1) |
| | 3′utr post-transcriptional regulation | Breaks motifs likely to be involved in post-transcriptional regulation | Very low to moderate (1~3) |

### 1.4.2 SNPLogic

It integrates SNP information from numerous sources to provide a comprehensive SNP selection, annotation and prioritization system. This integration provide information about;

- the genetic context of SNPs such as chromosomal and functional locations
- genotypic data such as allele frequencies in a population
- coverage of commercial arrays like Affimetrix and Illumina
- functional predictions modeled on sequence and structure
- identified associations via biological pathways, gene ontology terms or OMIM disease terms

The interface visualized at Figure 1.7 [27] involves SNP list formation and user defined thus subjective scoring rules to rank those SNP lists. Ranking system is established by the users depending on associations, connections, annotations and functional predictions of SNPs. SNP lists can be generated based on genes of interest, chromosomal regions, biological pathways, ontology terms and disease associations, which allow grouping SNPs in biologically meaningful ways. The approach of user defined scoring pattern of SNPLogic might be inadequate in terms of biological relavance since SNP prioritization requires an objective and reliable scoring function which can be applied to every genome wide association study and traits of interest [27].

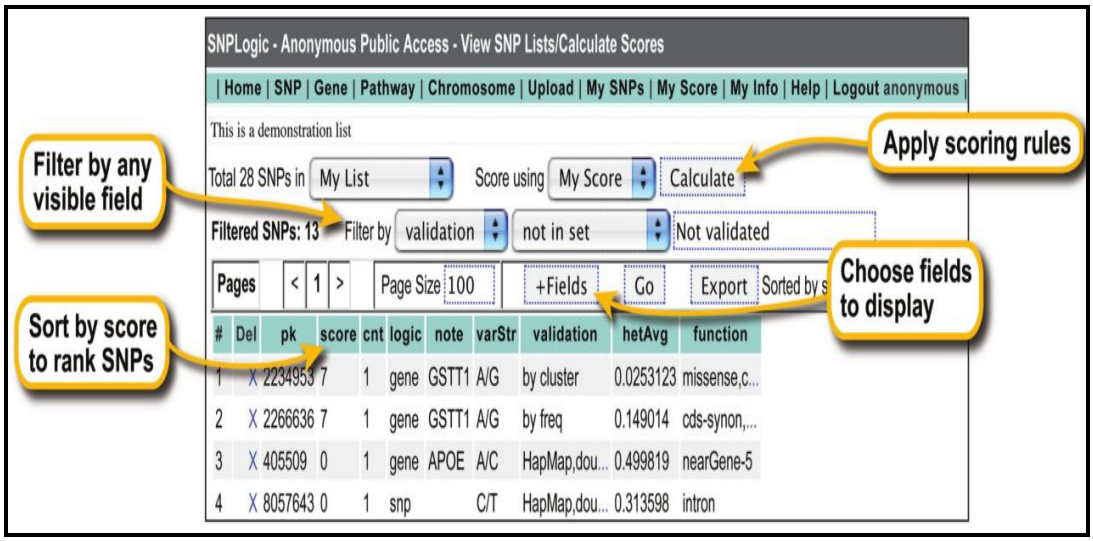



**Figure 1.7 SNPLogic web interface [27]**

### 1.4.3 SPOT

This SNP prioritization tool involves a genomic information network (GIN) approach allowing users to upload a list of SNPs and GWAS p-values providing output of a prioritized list of SNPs. GIN is a directed graph with nodes as features from a biological database. The GIN process begins with a SNP and ends in the terminal node which provides the calculation of its overall prioritization score S. S is determined by biological relevance obtained by combining

information from multiple databases. Figure 1.8 [28] represents the GIN idea for SNP prioritization and assignment of score [28].

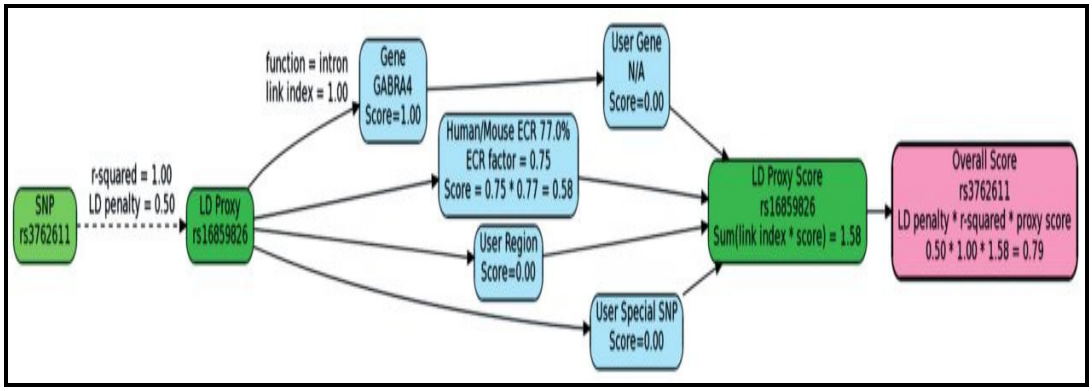



**Figure 1.8 SPOT genomic information network [28]**

### 1.4.4 SNPinfo

The web application of SNPinfo retrieves information regarding SNPs in HapMap and dbSNP and constructs LD relationships specific to ethnic groups from both sources. SNPs that were not genotyped in a GWAS, but are in LD with a SNP that was genotyped, can be screened by this way. Moreover, GWAS data generated in one ethnic group can be utilized to analyze SNPs in another ethnic groups [29].

SNPinfo web server contains various different modules suitable for different kinds of analysis during a genome wide association study [29];

- Candidate Gene SNP Selection (GenePipe)
  SNP selection for candidate genes is performed dependent on GWAS results, functional SNP prediction and LD information.
- GWAS Functional SNP Selection (GenomePipe)
  SNPs that are in high LD with GWAS SNPs provide a pool for the selection of functional SNPs
- GWAS SNP Selection in Linkage Loci (LinkagePipe)
  User defined list of linkage regions help to select small p-value GWAS SNPs in candidate genomic regions like linkage loci.
- LD Tag SNP Selection (TagSNP)
  Selection and visualization of LD tag SNP and is performed followed by formation of SNP list from various queries.
- SNP Function Prediction (FuncPred)
  Functional predictions and ethnic specific allele frequencies of SNPs serve as a query for this module.

- SNP Information in DNA Sequence (SNPseq)

  Sequence information of SNPs is provided with this module.

### 1.4.5 SNPit

SNP Integration Tool (SNPit) tool is built upon a federated data integration system, and provides current information on various SNP data sources [30]. For instance, the Human Gene Mutation Database (HGMD) provides information about a particular SNP location and its disease association depending on the gene it maps. Genomic context and location information is provided by the UCSC Browser's Genscan Gene Prediction track. Evolutionary relationships between the genomes are retrived from the ECR Browser. Recent positive selection within the human genome of certain SNPs are provided by Haplotter. SNPs affecting the protein function are predicted via SIFT. Linkage disequlibrium information is retrived from Genome Variation Server (GVS).

SNPit helps to integrate and analyze functional significance of SNPs and thus contributing in understanding the GWAS results. Inference engine plug makes it possible to analyze additional logical inference. SNPit provides information about functional annotation of SNPs via going to one source for up-to-date information. Figure 1.9 depicts the rules and heuristic weights assigned to different SNP characteristics [30]. Weights were assigned to each node in the decision tree, with the score of the final node calculated by multiplying the previous nodes in its' path. For example, SNPs that are in a coding region, that are non-synonymous and damaging, have a risk of moderate to very high; a heuristic weight of 3.375 was assigned to this branch of the tree.

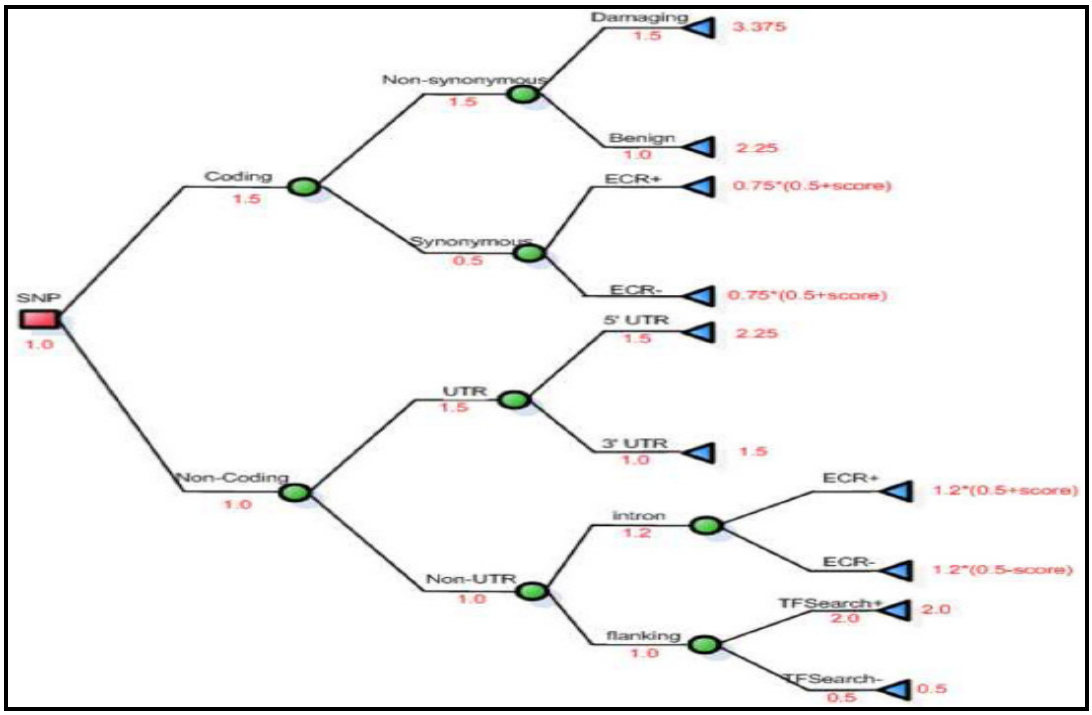



**Figure 1.9 SNPit heuristic weights [30]**

### 1.4.6 Analytic Hiearchy Process Based SNP Prioritization Approach

Analytic Hierarchy Process (AHP) is a multi criteria decision making method developed by Prof. Thomas L. Saaty to derive ratio scales from pairwise comparisons. Input can be acquired from actual measurements such as weight, height, price or from subjective opinions such as preference, satisfactory level etc. Some inconsistencies are allowed during the judgement. Principal eigenvectors determine the ratio scales whereas principal eigenvalues contribute to the computation of consistency index [31].

AHP provides an effective way to deal with complex decision making process via helping to identify and weight selection criteria. AHP involves analyzing the data collected for the criteria and determines the decision making process. It captures both subjective and objective evaluation measures, consistency of them and the alternatives are checked which lead to bias reduction in decision making [32].

There are two main steps of this process [32];

- The goal is decomposed into its constituent parts, progressing from the general to the specific involving a structure of a goal, criteria and alternatives as visualized in Figure 1.10 [33]. Further division of alternatives into an appropriate level of detail is performed, the possibility that the more criteria included, the less important each individual criterion may become should be taken into consideration.

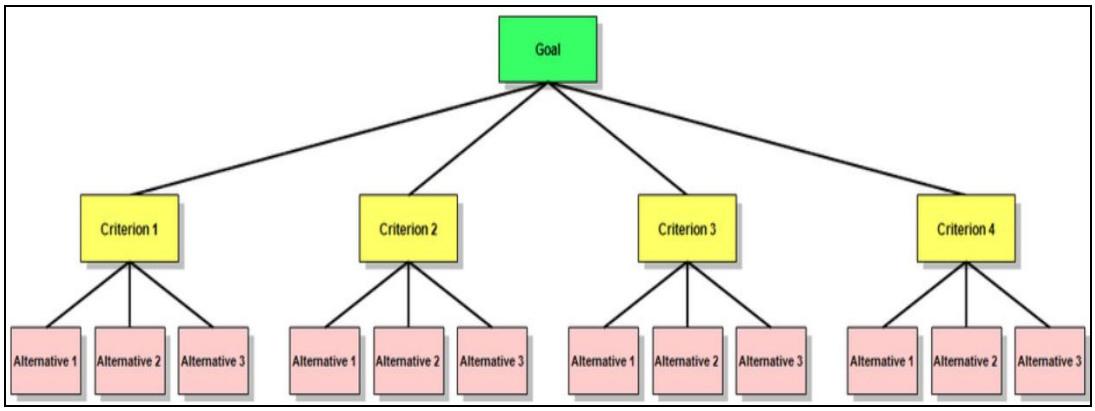



**Figure 1.10 AHP level of hierarchy example [33]**

- Next step is the assignment of relative weight to each criteria. Each criterion has a local and global priority. Global priority of a criteria represents relative importance within the overall model. Sum of the weights in a given parent criterion level must equal one. Pairwise scoring is performed for each criteria pair and this scoring is usually depending on the 9-point scale representes at Table 1.3.

**Table 1.3 AHP Pairwise Comparison Scale**

| Comparative Importance | Definition | Explanation |
|---|---|---|
| 1 | Equally important | Two decision elements equally influence the parent decision element. |
| 3 | Moderately more important | One decision element is moderately more influential than the other. |
| 5 | Strongly more important | One decision element has stronger influence than the other. |
| 7 | Very strongly more important | One decision element has significantly more influence over the other. |
| 9 | Extremely more important | The difference between influences of the two decision elements is extremely significant. |
| 2, 4, 6, 8 | Intermediate judgement values | Judgment values between equally, moderately, strongly, very strongly, and extremely |

After the pairwise comparisons are performed, comparison matrix and priority vector is constructed depending on the weights of each criteria. Table 1.4 depicts the AHP tree nodes and their weights for SNP prioritization. SNPs are evaluated based on the features embedded in the AHP tree as follows [34,35];

An indicator function is defined to assign the features that a particular SNP have;

$$I_k(SNP_i) = \begin{cases} 1, \text{ if } SNP_i \text{ has the feature at leaf node k} \\ 0, \text{ otherwise} \end{cases}$$

For instance if a SNP causes a frameshift (having the feature depicted at leaf node 1.3.2.1), then $I_{1.3.2.1}(SNP_i) = 1$ for that SNP. Then, final score of a SNP is computed via S function such that;

$$S(SNP_i) = \sum_{k=1}^{n} I_k(SNP_i) W_k \text{ for } i = 1, \ldots, m$$

where n is the number of leaf nodes, m is the total number of SNPs and $W_k$ is the weights given at Table 1.4. For instance a SNP that is known to map to a disease gene via LD and located at the 5' splice site acquires an AHP score of (0.036593+ 0.006597) = 0.04319. Table 1.5 represents the scoring order of leaf nodes depending on different biological features and genomic locations [34,35].

**Table 1.4 AHP Tree Nodes For SNP Prioritization And Combined Weights Of Nodes After Pairwise Scoring Performed By 5 Specialists [34,35]**

| GWAS Related Criteria | | | Genetic Criteria | | |
|---|---|---|---|---|---|
| **Leaf** | **Description** | **Score** | **Leaf** | **Description** | **Score** |
| 0.1 | Individual SNP | 0.033616 | 1.1.1 | Vertebrate | 0.037841 |
| 0.2.1 | Significant Gene - Via LD | 0.01598 | 1.1.2.1 | Mammalian - Significant Mouse ECR | 0.04532 |
| 0.2.2 | Significant Gene - Via Direct | 0.12099 | 1.1.2.2 | Mammalian - Other Mammalian | 0.023347 |
| 0.2.3 | Significant Gene - Via Pathway | 0.053266 | 1.3.1.1.1 | Non-Coding- UTR-3 - MiRNA Prediction | 0.001142 |
| 0.3.1 | Significant Pathway Gene - Via LD | 0.01465 | 1.3.1.1.2 | Non-Coding- UTR-3 - No MiRNA Prediction | 0.000604 |
| 0.3.2 | Significant Pathway Gene - Via Direct | 0.093825 | 1.3.1.2.1 | Non-Coding- UTR-5 - CpG Island | 0.002017 |
| 0.3.3 | Significant Pathway Gene - Via Pathway | 0.04738 | 1.3.1.2.2 | Non-Coding- UTR-5 - No CpG Island | 0.00063 |
| 1.2.1.1 | Disease Gene - Via LD | 0.036593 | 1.3.1.3 | Non-Coding - Intronic | 0.000825 |
| 1.2.1.2 | Disease Gene - Via Direct | 0.186016 | 1.3.1.4 | Non-Coding - Near Gene 3 | 0.001476 |
| 1.2.1.3 | Disease Gene - Via Pathway | 0.081725 | 1.3.1.5.1 | Non-Coding - Near Gene 5 - CpG Island | 0.002467 |
| 1.2.2.1.1 | Other Gene - Other Disease - Via LD | 0.005756 | 1.3.1.5.2 | Non-Coding - Near Gene 5 - No CpG Island | 0.000571 |
| 1.2.2.1.2 | Other Gene - Other Disease - Via Direct | 0.01818 | 1.3.1.6 | Non-Coding - Splice3 | 0.005295 |
| 1.2.2.1.3 | Other Gene - Other Disease - Via Pathway | 0.011161 | 1.3.1.7 | Non-Coding - Splice 5 | 0.006597 |
| 1.2.2.2.1 | Other Gene - Neutral - Via LD | 0.00145 | 1.3.2.1 | Coding - Frameshift | 0.103733 |
| 1.2.2.2.2 | Other Gene - Neutral - Via Direct | 0.004579 | 1.3.2.3.1 | Coding - CDS Non Syn - Polyphen Benign | 0.001997 |
| 1.2.2.2.3 | Other Gene - Neutral - Via Pathway | 0.002811 | 1.3.2.3.2 | Coding - CDS Non Syn - Possibly Damaging | 0.004713 |
| | | | 1.3.2.3.3 | Coding - CDS Non Syn - Probably Damaging | 0.009187 |
| | | | 1.3.2.3.4 | Coding - CDS Non Syn - Completely Determine | 0.024045 |

**Table 1.5 Weight Order For SNPs According To AHP Scoring [34,35]**

| | |
|---|---|
| 1.2.1.2 Disease Gene - Via Direct | 0.186016 |
| 0.2.2 Significant Gene - Via Direct | 0.12099 |
| 1.3.2.1 Coding - Frameshift | 0.103733 |
| 0.3.2 Significant Pathway Gene - Via Direct | 0.093825 |
| 1.2.1.3 Disease Gene - Via Pathway | 0.081725 |
| 0.2.3 Significant Gene - Via Pathway | 0.053266 |
| 0.3.3 Significant Pathway Gene - Via Pathway | 0.04738 |
| 1.1.2.1 Mammalian - Significant Mouse ECR | 0.04532 |
| 1.1.1 Vertebrate | 0.037841 |
| 1.2.1.1 Disease Gene - Via LD | 0.036593 |
| 0.1 Individual SNP | 0.033616 |
| 1.3.2.3.4 Coding - CDS Non Syn – Completely Determine | 0.024045 |
| 1.1.2.2 Mammalian - Other Mammalian | 0.023347 |
| 1.2.2.1.2 Other Gene - Other Disease - Via Direct | 0.01818 |
| 0.2.1 Significant Gene - Via LD | 0.01598 |
| 0.3.1 Significant Pathway Gene - Via LD | 0.01465 |
| 1.2.2.1.3 Other Gene - Other Disease - Via Pathway | 0.011161 |
| 1.3.2.3.3 Coding - CDS Non Syn – Probably Damaging | 0.009187 |
| 1.3.1.7 Non-Coding - Splice 5 | 0.006597 |
| 1.2.2.1.1 Other Gene - Other Disease - Via LD | 0.005756 |
| 1.3.1.6 Non-Coding - Splice3 | 0.005295 |
| 1.3.2.3.2 Coding - CDS Non Syn – Possibly Damaging | 0.004713 |
| 1.2.2.2.2 Other Gene - Neutral - Via Direct | 0.004579 |
| 1.2.2.2.3 Other Gene - Neutral - Via Pathway | 0.002811 |
| 1.3.1.5.1 Non-Coding - Near Gene 5 - CpG Island | 0.002467 |
| 1.3.1.2.1 Non-Coding- UTR-5 - CpG Island | 0.002017 |
| 1.3.2.3.1 Coding - CDS Non Syn - Polyphen Benign | 0.001997 |
| 1.3.1.4 Non-Coding - Near Gene 3 | 0.001476 |
| 1.2.2.2.1 Other Gene - Neutral - Via LD | 0.00145 |
| 1.3.1.1.1 Non-Coding- UTR-3 - MiRNA Prediction | 0.001142 |
| 1.3.1.3 Non-Coding - Intronic | 0.000825 |
| 1.3.1.2.2 Non-Coding- UTR-5 - No CpG Island | 0.00063 |
| 1.3.1.1.2 Non-Coding- UTR-3 - No MiRNA Prediction | 0.000604 |
| 1.3.1.5.2 Non-Coding - Near Gene 5 - No CpG Island | 0.000571 |

## 1.5 Complex Diseases

Complex diseases can be defined as disorders caused by the combined effect of various genes together with the contribution of environmental factors. They usually do not involve Mendelian characteristics. Some examples of complex diseases are obesity, diabetes, hypertension, Alzheimer disease, Parkinson disease and various types of cancers. They are also referred as polygenic diseases and they have the largest impact on the human population. For

instance in the Western world the the leading cause of death are known as cancer and heart diseases.

Many complex disorders including heart diseases, some types of cancer and diabetes involve a discontinous rather than normal distribution. Their phenotypic expression is dependent on various effects and interactions between genetic, social, and environmental factors. The degree of clustering within families might yield valueable information regarding the heritability of the disease. This approach requires epidemiological data from large sample size or within affected pedigrees. For instance, recurrence risk of a disease in the siblings of affected individuals give information about the multifactorial nature of the disease providing a quantitative estimate of the effect of genes on a trait.

Genetic influences on complex diseases can be more reliably evaluated via integrative studies involving bioinformatics, molecular biotechnology, and epidemiological methods to estimate disease risk. Moreover, those integrative studies provide a deeper understanding about the etiology of complex diseases and novel approaches to disease treatment and prevention [36].

There are three different approaches to screen biomarkers to find asosciations with complex diseases. Integrating data from different genetic studies and the relevant biological information might yield promising results at the systems biology level, SNPs being the most common variant among human genome serve as an important point in disease association studies. Second approach is the pathway based analysis on GWAS data to acquire enriched disease-causal information. Third approach involves integration of disease related pathways and networks for complex disease studies [2].

### 1.5.1 Alzheimer's Disease

It is the most common neurodegenerative disease accounting for about 65% of late life dementia. Neurodegeneration is referred as the progressive loss of structure or function of neurons, including death of neurons (**atrophy**). Alzheimer's disease (AD) is first described by a German Neurologist, Alois Alzheimer at 1906 [37].

Clinical characterization of AD involves **memory impairment** whereas it is pathologically characterized by **amyloid plaque** and **neurofibrillary tangle** formation in brain neurons consequently leading to brain atrophy (tissue death) [38]. Amyloid plaques are resulted from aberrancy in degradation of one type of amyloid protein (**beta amyloid**) in brain. It is produced via the synthesis of amyloid precursor protein (APP). Figure 1.11 [39] is a PET (positron emission tomography, a neuroimaging technique) scan picture representing the difference between 67 years old healthy person (left) and 79 years old AD patient brains in terms of beta amyloid production and atrophy. Beta amyloid is visualized via Carbon-11-labelled Pittsburgh B (11C-PIB) compound uptake (top) and cerebral regional glucose metabolism ($\mu$mol/min/100 mL) is evaluated via $1^8$F-fluorodeoxyglucose compound (bottom). For the top part of the figure, higher intensity values imply high beta amyloid production. For the bottom

part of the figure, higher intensity values imply higher glucose metabolism and thus lower proportion of atrophy.

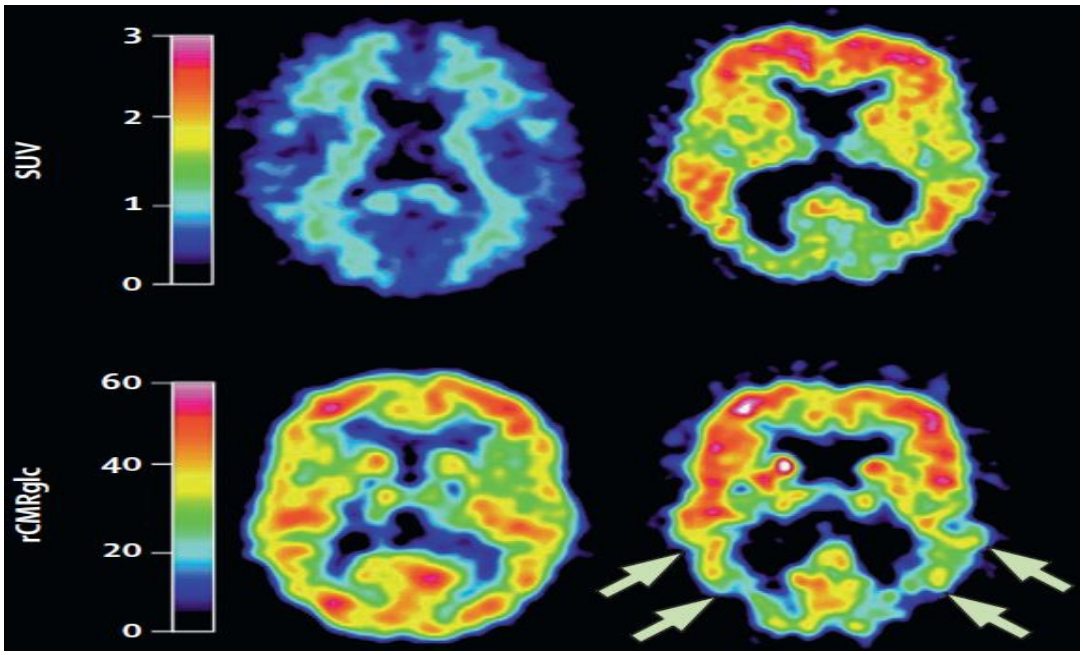



**Figure 1.11 Normal and AD brain PET scan comparison [39]**

Healthy brain breaks down the excessive beta amyloid and prevent plaque formation whereas in Alzheimer's disease case, those proteins begin to accumulate leading to the formation of insoluble plaques. Figure 1.12 [39] represents the hypothetical beta amyloid production from APP at neurons and consequent amyloid plaque formation.

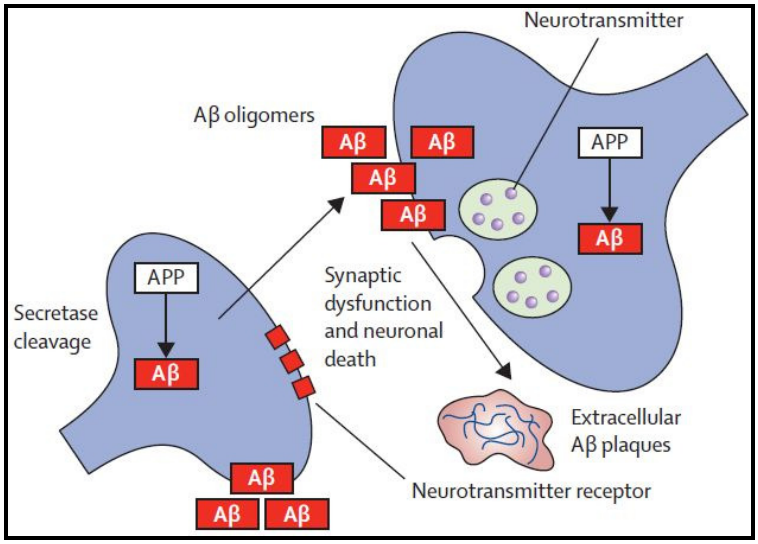



**Figure 1.12 Aβ production from APP and plaque formation [39]**

Figure 1.13 [40] represents beta amyloid plaque formation via immunofluorescence technique. Cortical amyloid plaque stained with anti-Ab antibody 4G8.

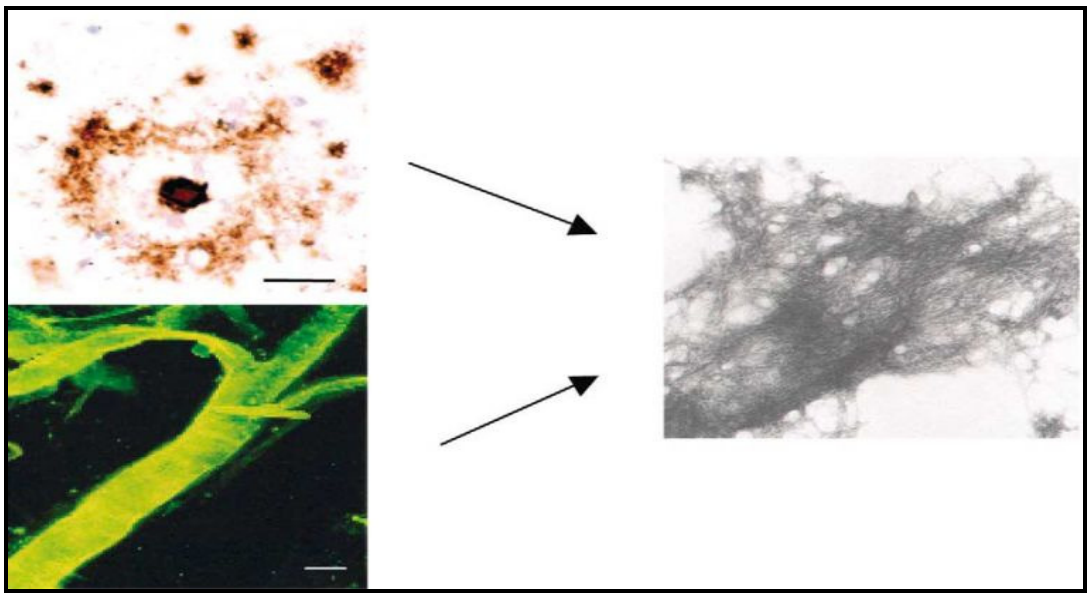


**Figure 1.13 Aβ plaque caused by Aβ deposits [40]**

Neurofibrillary tangles are insoluble twisted fibers consisting of **tau** protein, which is present in microtubule structure. Microtubule plays role in transport of nutrients and other substances inside the nerve cell. Abnormal tau protein structure in Alzheimer's disease cause neurofibrillary tangle formation. Figure 1.14 [40] represents fibrillary tangles caused by tau protein aggreagates. Hippocampal neurons stained with anti-phosphorylated tau antibody AT8.

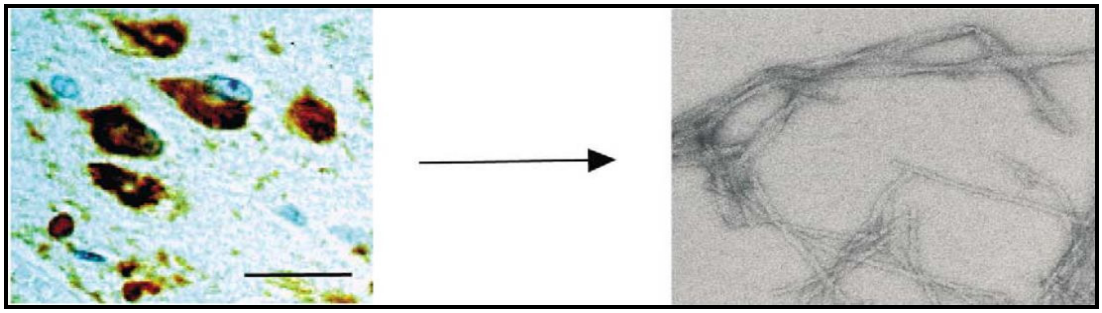


**Figure 1.14 Fibrillary tangles originated from tau aggregates [40]**

Alzheimer's disease have different categories depending on the onset. Early onset AD patients are less than 5% of the AD cases. They are diagnosed before the age of 65. More than 13% of individuals aged 65 years or older and 30-50% of people aged 80 years and older are diagnosed with AD. Those cases are referred as late onset Alzheimer's disease, most commonly observed type (LOAD) [41].

# CHAPTER 2

# LITERATURE INFORMATION AND GWAS RESULTS ABOUT ALZHEIMER'S DISEASE

This chapter mainly focuses on GWAS Analysis and literature information about Alzheimer's Disease. AD linked genes and variations are presented in this chaper.

## 2.1 Current Literature on AD and GWAS of Alzheimer's Disease

### 2.1.1 Overview of AD

Alzheimer's disease can be categorized into two major forms;

(i) strong familial clustering AD usually showing Mendelian disease transmission and early (before 65 years) or very early (before 50 years) age of onset

(ii) no familial aggregation AD occurring after 65 years of age (late onset AD, LOAD)

A 2009 review regarding GWAS in AD mentions that rare and usually highly penetrant mutations in three genes (amyloid precursor protein-**APP**, presenilin 1-**PSEN1** and presenilin 2-**PSEN2**) contribute to aberration in beta amyloid (Aβ) production leading to plaque formation. Those mutations are thought to be the major cause of the early onset AD [42]. Various other associative genes remain to be identified for early onset AD, which accounts for less than 5% of all AD cases [42].

### 2.1.2 Summary of GWAS of AD

**APOE** (apolipoprotein E) is found to consistently influence disease risk since significant association with AD is observed in various GWAS. In addition, **CLU** (clusterin, also known as apolipoprotein J) gene involves an associative SNP which resides at the intronic region with no known function [43]. According to another study [44], that intronic SNP is in strong LD with a synonymous SNP located at exon 5 of the CLU gene (rs7982; His315His), which is thought to affect alternative splicing or expressional regulation of the transcript.

Clusterin protein is predicted to bind soluble Aβ transporting it from plasma across the blood brain barrier [45]. APOE is predicted to transport Aβ in the opposite direction [46]. Clusterin together with APOE probably involve considerable roles in the regulation of cerebral Aβ levels and its transportation across the brain [42].

A 2010 review points out that the identification of APP, PSEN1 and PSEN2 associations with early onset AD provide a clear overview of the pathopysiology of AD. Moreover, APOE is mentioned to be universally accepted risk factor for late onset AD. It is a component of senile plaques and binds Aβ. APOE is thought to play role in Aβ clearance and deposition in the brain. In addition, some other functions that are not related to Aβ are suggested for APOE such as isoform specific synaptogenesis and cognition, neurotoxicity, **tau** hyperphosphorylation, neuro-inflammation, and brain metabolism [47].

Aβ is first cleaved from APP by β-secretase, followed by cleavage via γ-secretase complex. Presenilin is an essential component of the γ-secretase complex [47]. Amyloid cascade hypothesis which is mentioned briefly at Figure 1.8 involves contribution from elevated Aβ42/Aβ40 ratio or fibrillogenesis [48]. Aβ42 is referred as the toxic form of beta amyloid and its elevation influences inflammation, synaptic loss, ionic imbalance, and abnormal phosphorylation of proteins such as tau consequently leading to cell death and underlying clinical dementia [47].

There are alternative hypotheses such that tau [49] or dominant negative loss of presenilin function [50] might involve additional pathophysiologic mechanisms underlying AD.

A 2011 paper introduces a novel 'endophenotype 'approach coupled with gene expression analysis for identification of associative genes with AD. Endophenotype concept was first introduced in 1966 for Drosophila to define phenotypes that are microscopic and internal [51]. Various genetic studies involving gene expression level endophenotypes related with disease risk suggest that this combined approach contribute to increase in power for gene discovery and understanding of their mode of action [51]. For a feature to be referred as an endophenotype, it should involve an association with the disease in the general population. Moreover, endophenotypes should influence detectable changes in the clinically unaffected but at risk subjects such as family members of patients [51]. For instance Aβ can be used as an effective endophenotype for a linkage study of LOAD.

Endophenotype approach is promising since there can be several disease associative variants that cause changes in gene expression levels. Such variants might be identified by combining gene expression endophenotypes with existing disease GWAS to;

- identify novel disease genes and pathways
- validate potentially valuable results from disease GWAS
- estimate the mechanism of action of newly discovered disease genes

With their high potential in associative variant studies, gene expression endophenotypes are expected to be preferred for neurodegenerative diseases in the years to come [51].

Another 2011 paper points out three SNPs (rs744373, rs12989701 and rs7561528) at **BIN1** (bridging integrator 1) locus to be strongly associated with AD [38]. Moreover, Gleevec pathway is suggested to be candidate associative pathway linked with AD progression. Gleevec

is a cancer drug approved for the treatment of chronic myeloid leukemia. It was recently shown to reduce γ-secretase cleavage for APP [52] and to bind to a γ-secretase modulator [53]. Gleevec linked studies can yield valuable results if further validated about the potential mechanism and mode of action involved in AD [38].

A replication study of two GWAS [54, 55] confirmed the association of CLU ($p = 8.5 \times 10^{-10}$), **PICALM** (phosphatidylinositol binding clathrin assembly protein) ($p = 1.3 \times 10^{-9}$) and **CR1** (complement component (3b/4b) receptor 1 (Knops blood group)) loci ($p = 3.7 \times 10^{-9}$) [37]. **ABCA7** (ATP-binding cassette transporter), **MS4A** (membrane spanning 4 domains subfamily A), **CD2AP** (CD2 associated protein), **CD33**(sialic acid binding immunoglobulin like lectin) and **EPHA1** (ephrin receptor A1) are confirmed to be associated with LOAD via two GWA studies [56, 57]. The significant SNPs found and mapped to genes are represented below;

- ABCA7 (*rs3764650*) ($p = 5.0 \times 10^{-21}$).
- MS4A (*rs4938933*) ($p = 8.2 \times 10^{-12}$)
- CD2AP (*rs9349407*) ($p = 8.6 \times 10^{-9}$)
- EPHA1 (*rs11767557*) ($p = 6.0 \times 10^{-10}$)
- CD33 (*rs3865444*) ($p = 1.6 \times 10^{-9}$)

Those identified loci together with the new implicated pathways are summarized at Figure 2.1 [58] and below;

- *Immune system function* – CLU, CR1, ABCA7, MS4A, CD33, and EPHA1
- *Cholesterol metabolism* – APOE, CLU and ABCA7
- *Aβ metabolism* – APOE, CLU and ABCA7
- *Synaptic dysfunction and cell membrane processes* – PICALM, BIN1, CD33, CD2AP and EPHA1 [58].

Aβ might involve a modulatory effect on these new pathways as indicated by the blue arrows;

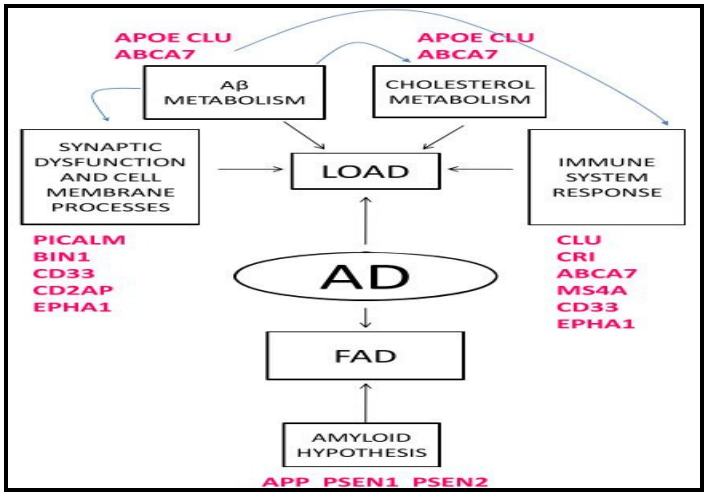



**Figure 2.1 Pathways and genes associated with AD [58]**

### 2.1.3 AD Databases Online

**AlzGene** is a widely referred database for genome wide association study findings regarding Alzheimer's disease [59]. It is a regularly updated database that provides a comprehensive and unbiased information regarding Alzheimer's Disease genetic association studies and thus it serves as a valuable source to identify AD linked genes. Meta-analyses are also available for all eligible polymorphisms with sufficient data.

Appendix B provides all the genes and locus on human chromosomes found to be potentially associated with AD [59]. AD linkage regions on chromosomes are marked with red whereas '*Top Results Gene*' with yellow. Top Results Gene list involves genes or loci having at least one variant representing a nominally significant summary odds ratio (OR) in the analysis of all studies, or those limited to samples of a specific ethnicity. There is a ranking system such that genes are ranked depending on the genetic variant with the best '*overall HuGENet/Venice grade*' [59]. HuGENet/Venice grade can be summarized with three main categories for meta-analyzed associations in AlzGene database[59]; a*mount of evidence*, *replication consistency* and *protection from bias*.

Ranking for the first category (amount of evidence) depends on the total number of minor alleles of cases and controls combined in the meta analysis.

- 'A' grade is assigned when the total number > 1,000
- 'B' grade is assigned when the total number is between 100 and 1,000
- 'C' grade is assigned when the total number < 100

Ranking for the second category (replication consistency) depends on a inconsistency measure ($I^2$) which describes the percentage of total variation across studies that is due to heterogeneity rather than chance, higher values imply higher inconsistency.

- 'A' grade is assigned if $I^2 < 25\%$
- 'B' grade is assigned if $I^2$ is between 25–50%
- 'C' grade is assigned if $I^2 > 50\%$

Ranking for the third category (protection from bias) depends on various potential sources of bias potentially affecting the association. Errors in phenotypes, genotypes, confounding (population stratification) and errors or biases at the meta-analysis level (publication and other selection biases) are examples for biases.

- 'A' grade implies that there is probably no bias
- 'B' grade that there is no demonstrable bias but important information is missing for its reliable prediction
- 'C' grade that there is evidence for potential or clear bias

For genes with identical grades, ranking is based on p-value, if their p-values are also identical then ranking is based on effect size (odds ratio) [59].

**AlzGenes** is another database that depends on catalogue information from National Human Genome Research Institute (NHGRI) [6,60]. AD linked genes are classified into 6 categories;

- Cholesterol and lipoprotein-related
- Cytokines
- Oxidative stress
- Nuclear receptor and related
- Proteases
- Miscellaneous

**OMIM** and **GeneRIF** are databases that involve valuable information about disease associations and genomic variations. They are also referred to acquire information about AD linked genes.

### 2.1.4 AD Associated Gene List

After referring to various databases, a list that involves AD linked genes should be decided to objectively evaluate the results in terms of biological relevance while performing comparisons between METU-SNP and SPOT. Firstly, **OMIM** (Online Mendelian Inheritance in Man) entries are investigated and a gene is said to be AD linked if phenotype information is observed to involve Alzheimer's Disease at the corresponding OMIM entry thus located at AD loci. Figure 2.2 represents the corresponding OMIM entry for a AD linked gene, APOE [61].

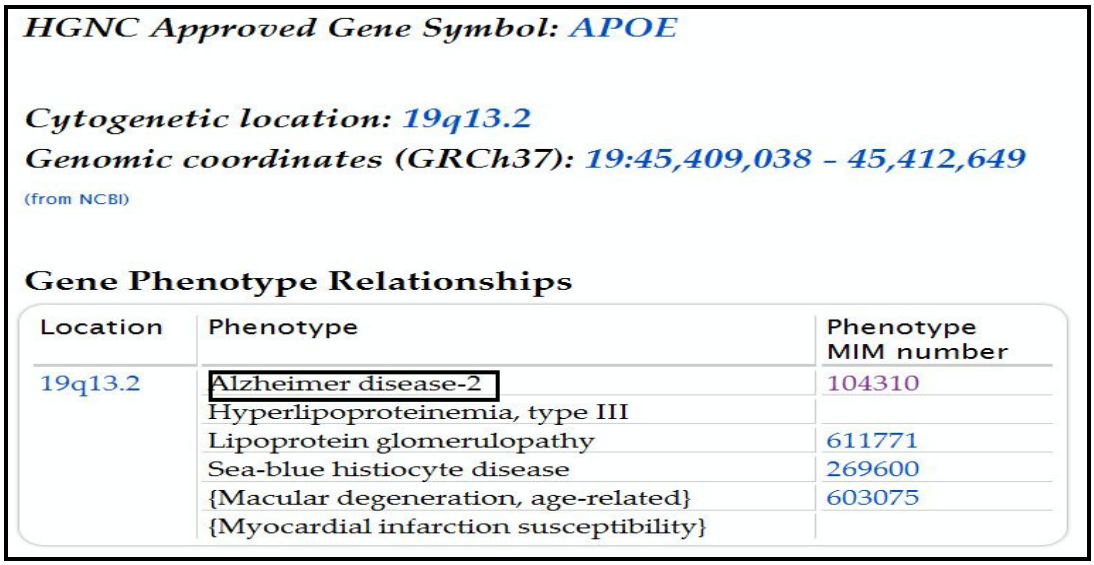



**Figure 2.2** APOE OMIM entry [61]

Final list of AD linked genes are retrieved from OMIM genes that involve AD phenotype entry (26 genes) together with the genes observed at AlzGenes and AlzGene databases (58 genes). Table 2.1 depicts the genes in AD loci and thus the AD linked genes list depending on 3 different databases.

**Table 2.1** 84 Genes Selected For AD Linked Genes List

| Gene Symbol | Gene Name |
|---|---|
| **A2M** | ALPHA-2-MACROGLOBULIN |
| **ABCA1** | ATP-BINDING CASSETTE, SUB-FAMILY A (ABC1), MEMBER 1 |
| **ACE** | ANGIOTENSIN I-CONVERTING ENZYME |
| **AD5** | ALZHEIMER DISEASE, FAMILIAL 5 |
| **AD6** | ALZHEIMER DISEASE, FAMILIAL 6 |
| **AD7** | ALZHEIMER DISEASE, FAMILIAL 7 |
| **AD8** | ALZHEIMER DISEASE, FAMILIAL 8 |
| **AD9** | ALZHEIMER DISEASE, FAMILIAL 9 |
| **AD10** | ALZHEIMER DISEASE, FAMILIAL 10 |
| **AD11** | ALZHEIMER DISEASE, FAMILIAL 11 |
| **AD12** | ALZHEIMER DISEASE, FAMILIAL 12 |
| **AD13** | ALZHEIMER DISEASE, FAMILIAL 13 |
| **AD14** | ALZHEIMER DISEASE, FAMILIAL 14 |
| **AD15** | ALZHEIMER DISEASE, FAMILIAL 15 |
| **AD16** | ALZHEIMER DISEASE, FAMILIAL 16 |
| **AGER** | ADVANCED GLYCOSYLATION END PRODUCT-SPECIFIC RECEPTOR |
| **ALDH2** | ALDEHYDE DEHYDROGENASE 2 |
| **APBB2** | AMYLOID BETA A4 PRECURSOR PROTEIN-BINDING, FAMILY B, MEMBER 2 |
| **APOA1** | APOLIPOPROTEIN A-I |
| **APOA4** | APOLIPOPROTEIN A-IV |
| **APOC1** | APOLIPOPROTEIN C-I |
| **APOC2** | APOLIPOPROTEIN C-II |
| **APOC3** | APOLIPOPROTEIN C-III |
| **APOE** | APOLIPOPROTEIN E |
| **APP** | AMYLOID BETA A4 PRECURSOR PROTEIN |
| **BIN1** | BRIDGING INTEGRATOR 1 |
| **BCHE** | BUTYRYLCHOLINESTERASE |
| **BLMH** | BLEOMYCIN HYDROLASE |
| **CBS** | CYSTATHIONINE BETA-SYNTHASE |
| **CCL2** | CHEMOKINE, CC MOTIF, LIGAND 2 |
| **CCR2** | CHEMOKINE, CC MOTIF, RECEPTOR 2 |

**Table 2.1 (cont.)** 84 Genes Selected For AD Linked Genes List

| Gene Symbol | Gene Name |
|---|---|
| **CD14** | MONOCYTE DIFFERENTIATION ANTIGEN CD14 |
| **CD36** | CD36 ANTIGEN |
| **CETP** | CHOLESTERYL ESTER TRANSFER PROTEIN, PLASMA |
| **CFH** | COMPLEMENT FACTOR H |
| **CHRNA7** | CHOLINERGIC RECEPTOR, NEURONAL NICOTINIC, ALPHA POLYPEPTIDE 7 |
| **CLU** | CLUSTERIN |
| **CR1** | COMPLEMENT COMPONENT RECEPTOR 1 |
| **CRP** | C-REACTIVE PROTEIN, PENTRAXIN-RELATED |
| **CST3** | CYSTATIN 3 |
| **CYP19A1** | CYTOCHROME P450, FAMILY 19, SUBFAMILY A, POLYPEPTIDE 1 |
| **ESR1** | ESTROGEN RECEPTOR 1 |
| **GNB3** | GUANINE NUCLEOTIDE-BINDING PROTEIN, BETA-3 |
| **GSTM1** | GLUTATHIONE S-TRANSFERASE, MU-1 |
| **GSTT1** | GLUTATHIONE S-TRANSFERASE, THETA-1 |
| **HFE** | HEMOCHROMATOSIS, HFE GENE |
| **HLA-A2** | MAJOR HISTOCOMPATIBILITY COMPLEX, CLASS I, A2 |
| **HMGCR** | 3-HYDROXY-3-METHYLGLUTARYL-CoA REDUCTASE |
| **HMOX1** | HEME OXYGENASE 1 |
| **HTR6** | 5-@HYDROXYTRYPTAMINE RECEPTOR 6 |
| **ICAM1** | INTERCELLULAR ADHESION MOLECULE 1 |
| **IL18** | INTERLEUKIN 18 |
| **IL1B** | INTERLEUKIN 1-BETA |
| **IL1RN** | INTERLEUKIN 1 RECEPTOR ANTAGONIST |
| **IL6** | INTERLEUKIN 6 |
| **LDLR** | LOW DENSITY LIPOPROTEIN RECEPTOR |
| **LIPA** | LIPASE A, LYSOSOMAL ACID |
| **LPA** | APOLIPOPROTEIN A |
| **LPL** | LIPOPROTEIN LIPASE |
| **LRP1** | LOW DENSITY LIPOPROTEIN RECEPTOR-RELATED PROTEIN 1 |
| **LRP6** | LOW DENSITY LIPOPROTEIN RECEPTOR-RELATED PROTEIN 6 |
| **MEF2A** | MADS BOX TRANSCRIPTION ENHANCER FACTOR 2, POLYPEPTIDE A |
| **MMP1** | MATRIX METALLOPROTEINASE 1 |
| **MMP3** | MATRIX METALLOPROTEINASE 3 |
| **MPO** | MYELOPEROXIDASE |
| **MTHFR** | 5,10-METHYLENETETRAHYDROFOLATE REDUCTASE |
| **NGB** | NEUROGLOBIN |

**Table 2.1 (cont.)** 84 Genes Selected For AD Linked Genes List

| Gene Symbol | Gene Name |
|---|---|
| **NOS3** | NITRIC OXIDE SYNTHASE 3 |
| **OLR1** | LOW DENSITY LIPOPROTEIN, OXIDIZED, RECEPTOR 1 |
| **PAXIP1** | PAX TRANSCRIPTION ACTIVATION DOMAIN-INTERACTING PROTEIN 1 |
| **PICALM** | PHOSPHATIDYLINOSITOL-BINDING CLATHRIN ASSEMBLY PROTEIN |
| **PLAU** | PLASMINOGEN ACTIVATOR, URINARY |
| **PON1** | PARAOXONASE 1 |
| **PON2** | PARAOXONASE 2 |
| **PPARA** | PEROXISOME PROLIFERATOR-ACTIVATED RECEPTOR-ALPHA |
| **PSEN1** | PRESENILIN 1 |
| **PSEN2** | PRESENILIN 2 |
| **PTGS2** | PROSTAGLANDIN-ENDOPEROXIDE SYNTHASE 2 |
| **SERPINE1** | SERPIN PEPTIDASE INHIBITOR, CLADE E MEMBER 1 |
| **SORL1** | SORTILIN-RELATED RECEPTOR |
| **SREBF1** | STEROL REGULATORY ELEMENT-BINDING TRANSCRIPTION FACTOR 1 |
| **TGFB1** | TRANSFORMING GROWTH FACTOR, BETA-1 |
| **TLR4** | TOLL-LIKE RECEPTOR 4 |
| **TNF** | TUMOR NECROSIS FACTOR |

# CHAPTER 3

# GWAS RESULTS AND COMPARISON OF AHP WITH OTHER PRIORITIZATION APPROACHES ON BIOLOGICAL RELEVANCE FOR ALZHEIMER'S DISEASE GENOTYPING DATA SETS

## 3.1 Alzheimer's Disease Genotyping Data Sets

### 3.1.1 ADNI AD Genotyping Data

Whole genome association data for Alzheimer's disease with the below properties was obtained from the Alzheimer's Disease Neuroimaging Initiative (*ADNI*) database [62];

- 149 AD cases and 182 controls
- 555.850 SNP-genotype fields from the Illumina 610Quad chip

The performance of our AHP based SNP prioritization system in terms of biological relevance is evaluated via comparison with SPOT.

### 3.1.2 GenADA AD Genotyping Data

*GenADA* is a multi-site collaborative study with the contribution of GlaxoSmithKline Inc and nine medical centres in Canada aiming to associate variations in candidate genes with Alzheimer's disease phenotypes. Both patients with an existing diagnosis of AD and newly diagnosed patients are enrolled in this study. For that reason, clinical data was retrospectively or prospectively obtained on Day 1 of entry. Data is retrieved from the database of Genotypes and Phenotypes (dbGAP) in National Center for Biotechnology Information (NCBI) [63]. GenADA whole genome association data for Alzheimer's disease involve the below properties;

- 852 AD cases and 866 controls
- 262264 SNP-genotype fields from the Affymetrix Mapping 250K chip

The performance of our AHP based SNP prioritization system is tested and compared with SPOT in terms of biological relevance.

### 3.2 GWAS

#### 3.2.1 GWAS Results of ADNI (p-value associations)

Initial quality control based filtering and preprocessing was performed by using the default thresholds of the system;

- Minor Allele Frequency = 0.05
- SNP Missingness Rate = 0.1
- Individual Missingness Rate = 0.1
- Hardy Weinberg Equilibrium = 0.001

After the quality control based filtering and preprocessing, GWAS is performed. As can be seen from Figure 3.1, p-value distribution of ADNI data involves a dispersed pattern with respect to chromosomes with SNPs rs4795895 and rs1233651 mapping to chromosome 17 and rs12457258 mapping to chromosome 18 having lowest p-values.

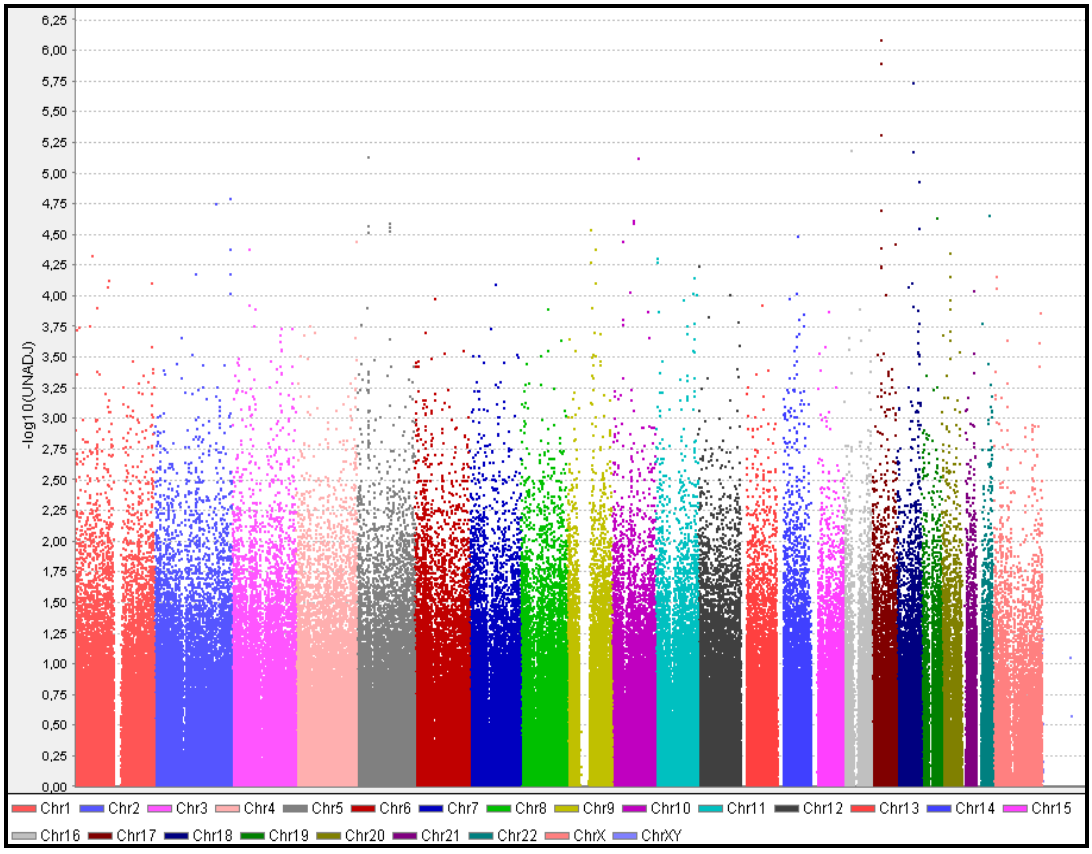



**Figure 3.1** ADNI AD genotyping data p-value distribution by chromosomes

### 3.2.2 GWAS Results of GenADA (p-value associations)

Initial quality control based filtering and preprocessing was performed by using the default thresholds of the system;

- Minor Allele Frequency = 0.05
- SNP Missingness Rate = 0.1
- Individual Missingness Rate = 0.1
- Hardy Weinberg Equilibrium = 0.001

After the quality control based filtering and preprocessing, GWAS is performed with a p-value threshold of 0.05. As can be seen from Figure 3.2, p-value distribution of GenADA data involves a dispersed pattern with respect to chromosomes with SNPs rs6980733 mapping to chromosome 8 and rs17123958 mapping to chromosome 14 having lowest p-values.

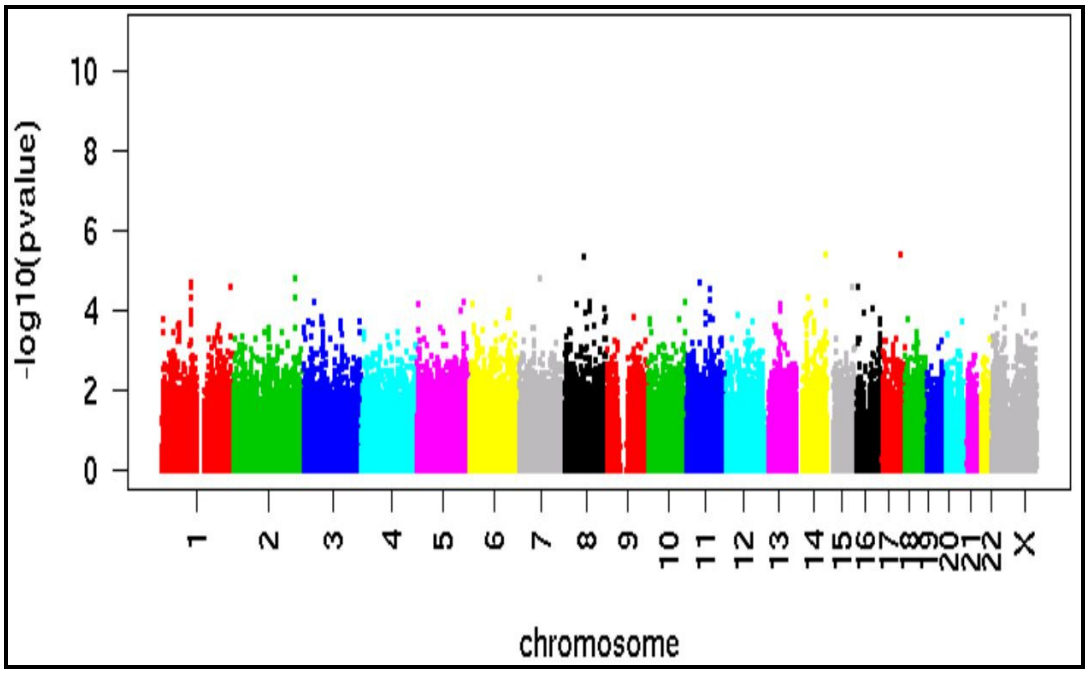



**Figure 3.2** GenADA AD genotyping data p-value distribution by chromosomes

## 3.3 AHP Prioritization

### 3.3.1 AHP Prioritization Results for ADNI (combined p-value and AHP scores)

Combined p-value for genes and pathways were computed by using 0.05 p-value threshold for both Fisher's combination test (genes) and Fisher's exact test (pathways) to find significant genes and pathways [64]. AHP based prioritization for 500K SNPs was performed within 4 hours via HCP Tesla in which NVIDIA C2070 is installed.

- 6 Gb GDDR5 RAM
- 1.5 GHz Memory Speed
- 144 Gb/sec Memory Bandwith

Table 3.1 depicts the outputs after combined p-value for genes are computed. Below is the Top 20 genes depending on their combined p-values and their OMIM associations are

represented. As can be seen from Table 3.1, 13 genes are OMIM associated. When combined p-value approach is applied after GWAS, it does not detect any AD linked gene in its Top 100 gene list as can be seen from Appendix C.

**Table 3.1 Top 20 Genes Depending On The Combined p-values And Their OMIM Associations For ADNI Data**

| Ranking | Gene | Combined p-value | OMIM association |
|---|---|---|---|
| 1 | SLC16A9 | ~0.0 | |
| 2 | RBFOX1 | ~0.0 | |
| 3 | CNTN5 | ~0.0 | Yes |
| 4 | C6orf10 | ~0.0 | |
| 5 | KDM4C | ~0.0 | Yes |
| 6 | RBMS3 | ~0.0 | Yes |
| 7 | SLC9A7 | ~0.0 | Yes |
| 8 | DSCAM | ~0.0 | Yes |
| 9 | FHIT | ~0.0 | Yes |
| 10 | LRP1B | ~0.0 | Yes |
| 11 | ANKRD44 | ~0.0 | |
| 12 | DACH1 | ~0.0 | Yes |
| 13 | SNX25 | ~0.0 | |
| 14 | FAM188B | ~0.0 | |
| 15 | CSMD1 | ~0.0 | Yes |
| 16 | HS3ST4 | ~0.0 | Yes |
| 17 | FER1L6 | ~0.0 | |
| 18 | NELL1 | ~0.0 | Yes |
| 19 | ALDH1A2 | ~0.0 | Yes |
| 20 | SUCLG2 | ~0.0 | Yes |

Table 3.2 represents Top 20 SNP list after METU-SNP AHP based prioritization, their AHP scores and the AD linked genes they map. As can be seen from Table 3.2, 5 SNPs map to AD linked genes SORL1, ABCA1 and LDLR.

**Table 3.2 Top 20 SNPs After AHP Prioritization For ADNI Data**

| Ranking | SNP ID | AHP score | AD linked gene |
|---|---|---|---|
| 1 | rs4651138 | 0.789991 | |
| 2 | rs2070045 | 0.789166 | SORL1 |
| 3 | rs2230806 | 0.757547 | ABCA1 |
| 4 | rs4652769 | 0.75215 | |
| 5 | rs3779870 | 0.737535 | |
| 6 | rs10808738 | 0.728803 | |
| 7 | rs4395923 | 0.728803 | |
| 8 | rs4936637 | 0.728803 | SORL1 |
| 9 | rs6424883 | 0.728803 | |
| 10 | rs10752893 | 0.728803 | |

**Table 3.2 (cont.) Top 20 SNPs After AHP Prioritization For ADNI Data**

| Ranking | SNP ID | AHP score | AD linked gene |
|---|---|---|---|
| 11 | rs9832203 | 0.728803 | |
| 12 | rs895286 | 0.728803 | |
| 13 | rs4358067 | 0.718534 | |
| 14 | rs10857526 | 0.718534 | |
| 15 | rs603634 | 0.718534 | |
| 16 | rs10934675 | 0.715562 | |
| 17 | rs1800464 | 0.714737 | |
| 18 | rs13390226 | 0.703919 | |
| 19 | rs1799898 | 0.703094 | LDLR |
| 20 | rs688 | 0.703094 | LDLR |

### 3.3.2 AHP Prioritization Results for GenADA (combined p-value and AHP scores)

Combined p-value for genes and pathways were computed by using 0.05 p-value threshold for both Fisher's combination test (genes) and Fisher's exact test (pathways) to find significant genes and pathways [64]. AHP based prioritization for 250K SNPs was performed within 2 hours via HCP Tesla in which NVIDIA C2070 is installed.

Table 3.3 depicts the outputs after combined p-value for genes are computed. Below is the Top 20 genes depending on their combined p-values and their OMIM associations are represented. As can be seen from Table 3.3, 15 genes are OMIM associated.When combined p-value approach is applied after GWAS, it detect only one AD linked gene (CD36) in its Top 100 gene list marked in bold and can be seen at Appendix D.

**Table 3.3 Top 20 Genes Depending On The Combined P-values And Their OMIM Associations For GenADA Data**

| Ranking | Gene | Combined p-value | OMIM association |
|---|---|---|---|
| 1 | AGAP11 | 1,089181E-08 | |
| 2 | C4orf18 | 1,187423E-08 | |
| 3 | CPT2 | 2,232619E-08 | Yes |
| 4 | OR1L1 | 1,773873E-07 | |
| 5 | NACC2 | 1,974952E-07 | |
| 6 | MCPH1 | 2,484007E-07 | Yes |
| 7 | C9orf150 | 4,419189E-07 | |
| 8 | PRSS1 | 4,765822E-07 | Yes |
| 9 | ESR2 | 7,396964E-07 | Yes |
| 10 | ARHGEF19 | 8,743196E-07 | Yes |
| 11 | STC1 | 8,987053E-07 | Yes |
| 12 | SMURF1 | 1,272100E-06 | Yes |
| 13 | CPZ | 1,276079E-06 | Yes |

**Table 3.3 (cont.) Top 20 Genes Depending On The Combined P-values And Their OMIM Associations For GenADA Data**

| Ranking | Gene | Combined p-value | OMIM association |
|---|---|---|---|
| 14 | MYPN | 1,451700E-06 | Yes |
| 15 | IFI30 | 1,649066E-06 | Yes |
| 16 | DAP | 1,937083E-06 | Yes |
| 17 | GNAZ | 1,997541E-06 | Yes |
| 18 | MRPS10 | 2,519949E-06 | Yes |
| 19 | PPT1 | 2,546921E-06 | Yes |
| 20 | HNRNPU | 2,780977E-06 | Yes |

Table 3.4 represents Top 20 SNP list after METU-SNP AHP based prioritization, their AHP scores and the AD linked genes they map. As can be seen from Table 3.2, 5 SNPs map to AD linked genes MPO, APP and A2M.

**Table 3.4 Top 20 SNPs After AHP Prioritization For GenADA Data**

| Ranking | SNP ID | AHP score | AD linked gene |
|---|---|---|---|
| 1 | rs7229 | 0.717355 | |
| 2 | rs14531 | 0.715879 | |
| 3 | rs4947 | 0.703094 | |
| 4 | rs2759 | 0.683118 | MPO |
| 5 | rs6304 | 0.683118 | |
| 6 | rs6324 | 0.681121 | |
| 7 | rs6323 | 0.681121 | |
| 8 | rs6305 | 0.681121 | |
| 9 | rs13136 | 0.679514 | |
| 10 | rs6314 | 0.669966 | |
| 11 | rs6308 | 0.66725 | |
| 12 | rs6313 | 0.665253 | |
| 13 | rs3010 | 0.655025 | |
| 14 | rs8951 | 0.654691 | |
| 15 | rs1146 | 0.654374 | APP |
| 16 | rs761388 | 0.654374 | |
| 17 | rs15966 | 0.654374 | |
| 18 | rs669 | 0.645277 | A2M |
| 19 | rs12222 | 0.644756 | |
| 20 | rs3761 | 0.644756 | |

## 3.4 SPOT

### 3.4.1 ADNI data

Table 3.5 represents Top 20 SNP list after SPOT prioritization for ADNI data, their AHP scores and the AD linked genes they map. As can be seen from Table 3.5, only one SNP map to an AD linked gene GNB3.

**Table 3.5 Top 20 SNPs After SPOT Prioritization For ADNI Data**

| Ranking | SNP ID | AD linked gene |
|---|---|---|
| 1 | rs4795895 | |
| 2 | rs17365991 | |
| 3 | rs3795263 | |
| 4 | rs4426564 | |
| 5 | rs2075650 | |
| 6 | rs12605132 | |
| 7 | rs9268368 | |
| 8 | rs10941091 | |
| 9 | rs667782 | |
| 10 | rs885691 | |
| 11 | rs1233651 | |
| 12 | rs5442 | GNB3 |
| 13 | rs12489170 | |
| 14 | rs6729218 | |
| 15 | rs13006848 | |
| 16 | rs12457258 | |
| 17 | rs6020624 | |
| 18 | rs4935801 | |
| 19 | rs3735080 | |
| 20 | rs3862683 | |

### 3.4.2 GenADA data

Table 3.6 represents Top 20 SNP list after SPOT prioritization for GenADA data, their AHP scores and the AD linked genes they map. As can be seen from Table 3.5, none of the SNPs map to an AD linked gene.

**Table 3.6 Top 20 SNPs After SPOT Prioritization For GenADA Data**

| Ranking | SNP ID | AD linked gene |
|---|---|---|
| 1 | rs1062683 | |
| 2 | rs3733472 | |
| 3 | rs10252253 | |
| 4 | rs11166412 | |
| 5 | rs3812205 | |
| 6 | rs17074644 | |
| 7 | rs3739407 | |
| 8 | rs6746030 | |
| 9 | rs4687319 | |
| 10 | rs3742261 | |
| 11 | rs17593271 | |
| 12 | rs1065035 | |
| 13 | rs560659 | |
| 14 | rs16966703 | |
| 15 | rs4545143 | |
| 16 | rs304230 | |

**Table 3.6 (cont.) Top 20 SNPs After SPOT Prioritization For GenADA Data**

| Ranking | SNP ID | AD linked gene |
|---|---|---|
| 17 | rs572846 | |
| 18 | rs8041254 | |
| 19 | rs6677080 | |
| 20 | rs17067596 | |

### 3.5 Comparison of Prioritization Approaches on Biological Relevance

#### 3.5.1 ADNI data

Top 100 SNP list after METU-SNP analysis and AHP based prioritization of ADNI data, Top 100 SNP list after SPOT based prioritization, and the AD linked genes that they are mapped to are presented at Appendix E. The genes on the previously described AD linked genes list are marked in bold. Table 3.7 represents the comparison of AHP based prioritization, combined p-value approach and SPOT prioritization tool depicting the gene associations of Top 100 SNP lists of them. For AHP based prioritization, SNPs are ranked depending on their AHP score and the resulting first 100 SNPs were analyzed based on their biological relevance and AD association. All Top 100 SNPs at the AHP list mapped to an OMIM associated gene at the Pubmed database and as can be seen from the Table 3.8, 18 out of 100 SNPs are found to be mapped to 6 AD linked genes (**SORL1**, **ABCA1**, **CHRNA7**, **LDLR**, **APP** and **IL1A**).

SPOT yields only 58 SNPs mapping to OMIM associated genes in the Top 100 SNP list after prioritization. There are 15 SNPs that do not even map to a gene. One gene is found to be AD linked in SPOT's Top 100 list; it can detect only **GNB3.**

Combined p-value for genes Top 100 gene list involves 73 genes to be OMIM associated. This approach cannot identify any AD linked gene in its Top 100 gene list.

Consequently, both combined p-value and AHP integration in METU-SNP considerably increased the biological relevance of SNP prioritization compared to SPOT. Our AHP based prioritization algorithm pinpoints the Alzheimer Disease associated genes SORL1, ABCA1, CHRNA7, LDLR, APP and IL1A successfully as can be seen from the Table 3.8 and outperforms SPOT in biological relevance in terms of SNP prioritization for AD.

**Table 3.7 Comparison Of AHP Priorization, Combined P-value Approach And SPOT In Terms Of Biological Relevance And AD Linkage For ADNI Data**

| Genes that AHP prioritized Top 100 SNPs mapped | | Top 100 genes according to combined p-value approach | | Genes that SPOT prioritized Top 100 SNPs mapped | |
|---|---|---|---|---|---|
| OMIM gene | AD loci | OMIM gene | AD loci | OMIM gene | AD loci |
| 100 | 18 | 73 | - | 58 | 1 |

**Table 3.8 AD Linked Genes That AHP Based Prioritization Points Out For ADNI Data**

| List of Genes on AD Loci within Top 100 AHP prioritized SNPs | # of Associated SNPs on the Gene |
|---|---|
| SORL1 | 6 |
| ABCA1 | 5 |
| CHRNA7 | 3 |
| LDLR | 2 |
| APP | 1 |
| IL1A | 1 |

The remaining 82 SNPs at the Top 100 SNP list of AHP priorization was analyzed in terms of biological relevance further and discovered that 66 of them are mapped to candidate AD linked genes as the literature involves various studies regarding their association and possible linkage with AD. Table 3.9 represents the candidate AD linked genes that AHP prioritized Top 100 SNPs mapped and the brief relevant literature information supporting their potential AD association.

**Table 3.9 Candidate AD Linked Genes That AHP Based Priorization Points Out For ADNI Data**

| List of Genes linked to AD within Top 100 AHP prioritized SNPs | # of Associated SNPs on the Gene | Relevant literature |
|---|---|---|
| **KALRN** | 23 | predominantly expressed in hippocampus,involved in neuronal stability and growth,underexpressed in AD hippocampus [65] |
| **ERBB4** | 23 | expressed by reactive astrocytes and microglia surrounding neuritic plaques in AD subjects,controls involve ERBB4 expression in distinct cellular compartments of hippocampal neurons [66] |
| **CTNNA3** | 8 | located at the AD6 region which is associated with LOAD [67]. 2 intronic SNPs are found to be in high LD with Aβ42 levels in 10 extended LOAD families [68] |
| **CYP7B1** | 6 | involved in 27-hydroxycholesterol metabolism and thus prevention of neurodegenerative accumulations. [69,70] Its expression is significantly lower in dentate neurons from AD [71] |
| **FGF1** | 2 | elevated concentrations in the CSF of AD patients,which can be caused by increased generation of glial cells producing FGF-1. FGF-1 expression can represent an active response to neurodegeneration [72] |

**Table 3.9 (cont.) Candidate AD Linked Genes That AHP Based Priorization Points Out For ADNI Data**

| | | |
|---|---|---|
| **MME** | 2 | significant decrease in mRNA levels in cerebral cortex of AD individuals compared to controls [73]. |
| **MAOA** | 1 | involved in the pathogenesis of mood disorders and AD. Its activity and gene expression are upregulated in different brain areas of AD patients. MAOA-VNTR polymorphism is associated with depression and AD [74] |
| **MRE11A** | 1 | a DNA repair enzyme, expression reduced in AD cortex neurons, loss of the Mre11 complex may be associated with the AD pathogenesis. MRE11A underexpression might lead to neuronal depletion [75] |

The biological relevance comparison with SPOT and METU-SNP outputs in terms of AD linked genes for the ADNI data can be summarized as follows;

- AHP based prioritization performs better than only combined p-value approach and SPOT prioritization.
- ~20% of AHP prioritized SNPs associated with AD loci and also >50% of AHP prioritized SNPs map to potential AD linked genes.
- AHP prioritization identifies many potentially AD linked genes in addition to 6 AD linked genes in the Top 100 SNP gene association list after the prioritization of 555.850 SNPs. KALRN, ERBB4, CTNNA3, CYP7B1,FGF1, MME, MAOA and MRE11A are hot candidate genes to be linked with AD as the literature involves several studies about their possible association with AD.
- AHP prioritization of GWAS SNPs will be helpful to identify disease associated genes for downstream analysis.

### 3.5.2 GenADA data

Top 100 SNP list after METU-SNP analysis and AHP based prioritization of GenADA data, Top 100 SNP list after SPOT based prioritization, and the AD linked genes that they are mapped to are presented at Appendix F. The genes on the previously described AD linked genes list are marked in bold.

Table 3.10 represents the comparison of AHP based prioritization, combined p-value approach and SPOT prioritization tool depicting the gene associations of Top 100 SNP lists of them. For AHP based prioritization, SNPs are ranked depending on their AHP score and the resulting first 100 SNPs were analyzed based on their biological relevance and AD association. All Top 100 SNPs at the AHP list mapped to an OMIM associated gene at the PubMed database

and as can be seen from the Table 3.10, 37 out of 100 SNPs are found to be mapped to 8 AD linked genes (**APP**, **A2M, ACE, PTGS2, APOA1, LDLR, LPL** and **MPO**).

SPOT yields only 75 SNPs mapping to OMIM associated genes in the Top 100 SNP list after prioritization. Moreover, 4 SNPs do not even map to a gene. The only AD linked gene that SPOT can detect in its Top 100 list is **BIN1**.

Combined p-value for genes Top 100 gene list involves 75 SNPs mapping to OMIM associated genes. This approach identify only **CD36** as an AD linked gene in its Top 100 gene list.

Consequently, both combined p-value and AHP integration in METU-SNP considerably increased the biological relevance of SNP prioritization compared to SPOT for this AD genotyping data. Our AHP based prioritization algorithm pinpoints the Alzheimer Disease associated genes APP, A2M, ACE, PTGS2, APOA1, LDLR, LPL and MPO successfully as can be seen from the Table 3.11 and outperforms SPOT in biological relevance in terms of SNP prioritization for AD.

**Table 3.10 Comparison Of AHP Priorization, Combined P-value Approach And SPOT In Terms Of Biological Relevance And AD Linkage For GenADA Data**

| Genes that AHP prioritized Top 100 SNPs mapped | | Top 100 genes according to combined p-value approach | | Genes that SPOT prioritized Top 100 SNPs mapped | |
|---|---|---|---|---|---|
| **OMIM gene** | **AD loci** | **OMIM gene** | **AD loci** | **OMIM gene** | **AD loci** |
| 100 | 37 | 75 | 1 | 75 | 1 |

**Table 3.11 AD Linked Genes That AHP Based Prioritization Points Out For GenADA Data**

| List of Genes on AD Loci within Top 100 AHP prioritized SNPs | # of Associated SNPs on the Gene |
|---|---|
| APP | 18 |
| A2M | 7 |
| ACE | 4 |
| PTGS2 | 3 |
| APOA1 | 2 |
| LDLR | 1 |
| LPL | 1 |
| MPO | 1 |

The remaining 63 SNPs at the Top 100 SNP list of AHP priorization was analyzed in terms of biological relevance further and discovered that 10 of them are mapped to candidate AD

linked genes as the literature involves various studies regarding their association and possible linkage with AD. Table 3.12 represents the candidate AD linked genes that AHP prioritized Top 100 SNPs mapped and the brief relevant literature information supporting their potential AD association.

**Table 3.12 Candidate AD Linked Genes That AHP Based Priorization Points Out For GenADA Data**

| List of Genes linked to AD within Top 100 AHP prioritized SNPs | # of Associated SNPs on the Gene | Relevant literature |
|---|---|---|
| **ESR2** | 3 | Two SNPs (rs1271573 and s1256043) with T/T allele are more frequent in AD women compared to control woman subjects. (p-values are 0.012 and 0.016, respectively). Moreover, ESR2 rs1271573 T/T and the s1256043 T/T genotypes involve a nearly 1.9–fold increase in the risk of AD in women [76]. |
| **UBB** | 3 | A dinucleotide deletion in UBB leads to formation of polyubiquitin causing neuritic beading, impairment of mitochondrial movements, mitochondrial stress and neuronal degeneration in primary neurons. The polyubiquitin-linked clogging of mitochondria in neurites might contribute to axon injury and neuropathology in AD [77]. |
| **EEF2** | 2 | EEF2 levels are significantly lower in AD subjects compared to controls. The decrease of total eEF2 is found to be significantly correlated with the progression of neurofibrillary degeneration [78]. |
| **MAOA** | 2 | It is involved in the pathogenesis of mood disorders and AD. Its activity and gene expression are upregulated in different brain areas of AD patients. MAOA-VNTR polymorphism is associated with depression and AD [74] |

The biological relevance comparison with SPOT and METU-SNP outputs in terms of AD linked genes for the GenADA data can be summarized as follows;

- AHP based prioritization performs better than only combined p-value approach and SPOT prioritization.
- ~40% of AHP prioritized SNPs associated with AD loci.
- AHP prioritization identifies 8 AD linked genes (APP, A2M, ACE, PTGS2, APOA1, LDLR, LPL and MPO) in the Top 100 SNP gene association list after the prioritization of 262264 SNPs. Widely accepted AD marker gene APP is the most frequently observed gene in the Top 100 list as 18 SNPs are mapped, which further supports the strength of AHP based SNP prioritization in terms of biological relevance.

- 10 SNPs in the Top 100 list are mapped to potentially AD linked genes (ESR2, UBB, EEF2 and MAOA) as the literature involves various studies regarding their linkage and association with AD.

Glycolysis and gluconeogenesis, leukocyte migration, axon guidance, actin filament polymerization, cell adhesion, DNA fragmentation during apoptosis, fatty acid metabolism and negative regulation of cell proliferation are common pathways residing at Top 100 pathways according to combined p-value for pathways that are observed in GWAS results of both data sets.

GWAS of both data with METU-SNP software and AHP based prioritization confirms the literature for Alzheimer Disease associated genes; A2M, ABCA1, ACE, APOA1, APP, CHRNA7, IL1A, LDLR, LPL, MPO, PTGS2, SORL1. rs3781835 at SORL1, rs4343 and rs4351 at ACE1 are SNPs with high AHP scores are also listed to be AD associated at PharmGKB database. Moreover, rs6313 has a high AHP ranking and maps to HTR2A gene. CT and TT genotype of rs6313 indicates resistance to the treatment with antipsychotic drugs for AD patients presenting delusional symptoms.

The candidate novel AD linked genes proposed after the analysis of both data are; **CTNNA3**, **CYP7B1**, EEF2, **ERBB4**, **ESR2**, **FGF1**, **KALRN**, MAOA, MME, **MRE11A** and UBB. They are investigated in terms of their expression localization in Genatlas database [79]. All of the proposed genes involve expression in brain except MME, MAOA, UBB and EEF2. Thus, genes marked in bold are candidate novel AD linked genes that AHP based SNP prioritization points out.

# CHAPTER 4

# EVALUATION OF USER DEFINED AHP PRIORITIZATION PARAMETERS : P-VALUE THRESHOLD OF SNPS AS A PRE-PRIORITIZATION CUTOFF

## 4.1 AHP prioritization performance of METU-SNP in different p-value thresholds for SNPs

METU-SNP is an integrative complex disease association analysis tool that allows analysis of a genotyping data in various aspects. In the METU-SNP workflow for genome-wide association study, statistical analysis and thus p-value computation via PLINK software is followed by combined p-value computation for genes and pathways. AHP prioritization is the next step contributing to biological relevance for GWAS. User can set the p-value threshold for SNPs to be selected for AHP prioritization, SNPs having p-values larger than the threshold is not AHP prioritized. After AHP based prioritization, performance of the selected SNP subset can be evaluated via k-fold cross validation provided at the 'Performance' tab of METU-SNP. WEKA (Waikato Environment for Knowledge Analysis) is implemented in METU-SNP for this performance and classification purposes. WEKA is a tool that provides machine learning algorithms implemented in Java. Various learning schemes such as decision trees, instance-based classifiers, support vector machines, Bayesian decision schemes are involved. Moreover, evaluation methods such as cross-validation, bootstrapping and attribute selection methods are also implemented in WEKA [80].

Below are some basic terms about machine learning;

- *Instance* is an object at a space of fixed dimension
- Each dimension is the *attribute* of an object, attributes are usually nominal, numerical or strings
- A *class attribute* determines the appurtenance of the instance
- Set of instances make up a *dataset*
- *Training set* is used to build a classifier.Classifier building involves learning from instances to predict the class attribute of new ones.
- *Test set* is used to evaluate a classifier

Different p-value thresholds are selected to compare the performance of the AHP based prioritization. Depending on the accuracy measures, the pre-prioritization cutoff value is the optimal p-value that yields the best performance measures.

Prediction and classification performances of AHP based SNP lists are evaluated via k-fold Cross Validation (CV) run using Naive Bayes and SMO (sequential minimal optimization) classifiers as the supervised learning scheme for the ADNI data. The following measures are used to estimate the prediction performances:

- Specificity = TN / (FP + TN)
- Sensitivity = TP / (TP + FN)
- Accuracy = (TP + TN) / (P + N)
- Negative Predictive Value (NPV) = TN / (TN + FN)
- Precision = TP / (TP + FP)

where TP denotes True Positive, TN denotes True Negative, FP denotes False Positive and FN denotes False Negative for a 2x2 confusion matrix. Table 4.1 represent the number of SNPs selected for AHP prioritization with respect to different p-value thresholds. The p-value threshold can be considered as maximum allowed false positive rate.

Different p-value thresholds are used for SNP prioritization, then Top 20000 SNPs are chosen for cross validation tests to compare AHP based prioritization in terms of performance and classification measures in various p-value thresholds.

**Table 4.1** Number Of AHP Prioritized SNPs In Different p-value Thresholds For ADNI Data

| p-value threshold | # of AHP prioritized SNPs |
|---|---|
| 0.05 | 24578 |
| 0.1 | 50453 |
| 0.2 | 101122 |
| 0.3 | 152782 |
| 0.4 | 205794 |
| 0.5 | 256730 |
| 0.6 | 308288 |
| 0.7 | 359928 |
| 0.8 | 412358 |
| 0.9 | 465213 |
| 1.0 | 516893 |

Table 4.2 depicts the 5-fold cross validation training results for different p-value threshold of SNPs with Naive Bayes as the supervised learning scheme and Table 4.3 represents the corresponding test results when 20000 SNPs are used.

**Table 4.2** 5-fold Cross Validation Training Results For ADNI Data(Learning Scheme:Naive Bayes, 20000 SNPs)

| | **Correct classification (%)** |
|---|---|
| **AHP (p-value threshold:0.05)** | 99.1416 |
| **AHP (p-value threshold:0.1)** | 99.1416 |
| **AHP (p-value threshold:0.2)** | 99.1416 |
| **AHP (p-value threshold:0.3)** | 99.1416 |
| **AHP (p-value threshold:0.4)** | 98.2833 |
| **AHP (p-value threshold:0.5)** | 96.9957 |
| **AHP (p-value threshold:0.6)** | 94.4206 |
| **AHP (p-value threshold:0.7)** | 94.4206 |
| **AHP (p-value threshold:0.8)** | 93.9914 |
| **AHP (p-value threshold:0.9)** | 93.9914 |
| **AHP (p-value threshold:1.0)** | 92.7039 |

**Table 4.3** 5-fold Cross Validation Test Results For ADNI Data(Learning Scheme:Naive Bayes, 20000 SNPs)

| | **Specificity** | **Sensitivity** | **Accuracy** | **NPV** | **Precision** |
|---|---|---|---|---|---|
| **AHP (p-value threshold:0.05)** | 0.0769231 | 0.8695652 | 0.44898 | 0.4 | 0.454545 |
| **AHP (p-value threshold:0.1)** | 0.0769231 | 0.8695652 | 0.44898 | 0.4 | 0.454545 |
| **AHP (p-value threshold:0.2)** | 0.0769231 | 0.8695652 | 0.44898 | 0.4 | 0.454545 |
| **AHP (p-value threshold:0.3)** | 0.1153846 | 0.9565217 | 0.510204 | 0.75 | 0.488889 |
| **AHP (p-value threshold:0.4)** | 0.0769231 | 0.9130435 | 0.469388 | 0.5 | 0.466667 |
| **AHP (p-value threshold:0.5)** | 0.1923077 | 0.8695652 | 0.510204 | 0.63 | 0.487805 |
| **AHP (p-value threshold:0.6)** | 0.1923077 | 0.826087 | 0.489796 | 0.56 | 0.475 |
| **AHP (p-value threshold:0.7)** | 0.153846154 | 0.826086957 | 0.469387755 | 0.5 | 0.463415 |
| **AHP (p-value threshold:0.8)** | 0.1538462 | 0.8695652 | 0.489796 | 0.57 | 0.47619 |
| **AHP (p-value threshold:0.9)** | 0.1538462 | 0.8695652 | 0.489796 | 0.57 | 0.47619 |
| **AHP (p-value threshold:1.0)** | 0.1923077 | 0.826087 | 0.489796 | 0.56 | 0.475 |

Another supervised learning scheme, SMO(sequential minimal optimization) is used for cross validation tests. Table 4.4 represents the 5-fold cross validation training results for different p-value threshold of SNPs with SMO as the supervised learning scheme when 20000 SNPs are used. Table 4.5 summarizes the corresponding test results.

**Table 4.4** 5-Fold Cross Validation Training Results For ADNI Data(Learning Scheme:SMO, 20000 SNPs)

| | **Correct classification (%)** |
|---|---|
| **AHP (p-value threshold:0.05)** | 100 |
| **AHP (p-value threshold:0.1)** | 100 |
| **AHP (p-value threshold:0.2)** | 100 |
| **AHP (p-value threshold:0.3)** | 99.1416 |
| **AHP (p-value threshold:0.4)** | 98.7124 |
| **AHP (p-value threshold:0.5)** | 97.8541 |
| **AHP (p-value threshold:0.6)** | 96.1373 |
| **AHP (p-value threshold:0.7)** | 96.1373 |
| **AHP (p-value threshold:0.8)** | 96.1373 |
| **AHP (p-value threshold:0.9)** | 95.7082 |
| **AHP (p-value threshold:1.0)** | 95.279 |

**Table 4.5** 5-Fold Cross Validation Test Results For ADNI Data(Learning Scheme:SMO, 20000 SNPs)

| | **Specificity** | **Sensitivity** | **Accuracy** | **NPV** | **Precision** |
|---|---|---|---|---|---|
| **AHP (p-value threshold:0.05)** | 0.0769231 | 1.0 | 0.510204 | 1.0 | 0.489362 |
| **AHP (p-value threshold:0.1)** | 0.0769231 | 1.0 | 0.510204 | 1.0 | 0.489362 |
| **AHP (p-value threshold:0.2)** | 0.0769231 | 1.0 | 0.510204 | 1.0 | 0.489362 |
| **AHP (p-value threshold:0.3)** | 0.1153846 | 1.0 | 0.530612 | 1.0 | 0.5 |
| **AHP (p-value threshold:0.4)** | 0.1923077 | 0.9565217 | 0.55102 | 0.83 | 0.511628 |
| **AHP (p-value threshold:0.5)** | 0.2692308 | 0.9130435 | 0.571429 | 0.78 | 0.525 |
| **AHP (p-value threshold:0.6)** | 0.2692308 | 0.9130435 | 0.571429 | 0.78 | 0.525 |
| **AHP (p-value threshold:0.7)** | 0.2692307 | 0.9130435 | 0.571428 | 0.777 | 0.525 |
| **AHP (p-value threshold:0.8)** | 0.3076923 | 0.9130435 | 0.591837 | 0.8 | 0.538462 |
| **AHP (p-value threshold:0.9)** | 0.2692308 | 0.9130435 | 0.571429 | 0.78 | 0.525 |
| **AHP (p-value threshold:1.0)** | 0.3076923 | 0.8695652 | 0.571429 | 0.73 | 0.526316 |

For both Naive Bayes and SMO cross validation tests, various p-value thresholds are applied, our analysis results are presented in Table 4.3 and 4.5. As expected lower the p-value of SNPs considered for the analysis, higher the classification measures.

A guideline for the users on how to choose p-value parameter before prioritization depending on the goal of their study is discussed in Chapter 6.

# CHAPTER 5

# EVALUATION OF USER DEFINED AHP PRIORITIZATION PARAMETERS : AHP SCORE THRESHOLD OF SNPS AS A POST-PRIORITIZATION CUTOFF

This chapter is mainly composed of AHP score cutoff estimation for two AD genotyping data depending on biological relevance measures. AHP score distribution of the SNPs after AHP prioritization is determined and as a post-prioritization cutoff value, AHP score threshold is decided depending on biological relevance measures of SNP list based on AHP score ranking.

## 5.1 AHP score distribution of the AD genotyping data after AHP based prioritization

As mentioned in Chapter 3, two genotyping data are used for this study. ADNI genotyping data is a relatively small dataset with 149 AD cases and 182 controls whereas GenADA data is a quite large dataset with 852 AD cases and 866 controls. The former study involves ~500K SNPs while the latter involves ~250K SNPs. After AHP prioritization is performed for all SNPs in the datasets, AHP score distribution of SNPs are determined and AD linkage ratios are investigated as the most pronounced biological relevance indicator. AD linkage ratio is the ratio of SNPs mapping to AD linked genes to the total number of SNPs at a specific AHP score range. For instance if there are 10 SNPs mapping to AD linked genes out of 100 SNPs, the AD linkage ratio is said to be 0.1. Depending on AD linkage ratio, AHP score cutoff is estimated for those genotyping data as post-prioritization cutoff value, implying that after this cutoff value AD linkage ratio and thus biological relevance representation of SNPs significantly decrease.

### 5.1.1 ADNI data

AHP prioritization is performed for all the SNPs in the data via choosing the p-value threshold as 1.0. Table 5.1 represents the AHP score distribution and AD linkage ratio after all 500K SNPs are AHP prioritized. AHP scores of SNPs range between 0.0 and 0.8.

**Table 5.1 AHP Score Distribution And Ratio Of SNPs Mapping To AD Linked Genes For ADNI Data**

| AHP score range | # of SNPs | # of SNPs mapped to AD linked genes | AD linkage ratio |
|---|---|---|---|
| AHP>0.7 | 21 | 5 | 0.2380952 |
| 0.6<AHP<0.7 | 1181 | 295 | 0.2497883 |
| 0.5<AHP<0.6 | 1429 | 101 | 0.0706788 |
| 0.4<AHP<0.5 | 50800 | 199 | 0.003917323 |
| 0.3<AHP<0.4 | 48856 | 34 | 0.0006959227 |
| 0.2<AHP<0.3 | 48152 | 5 | 0.0001038378 |
| 0.1<AHP<0.2 | 44007 | 0 | 0.0 |
| 0.0<AHP<0.1 | 359486 | 0 | 0.0 |

As can be seen from Figure 5.1, the AHP score distribution for the ADNI data involves an accumulation in AHP score range between 0.0 and 0.1.

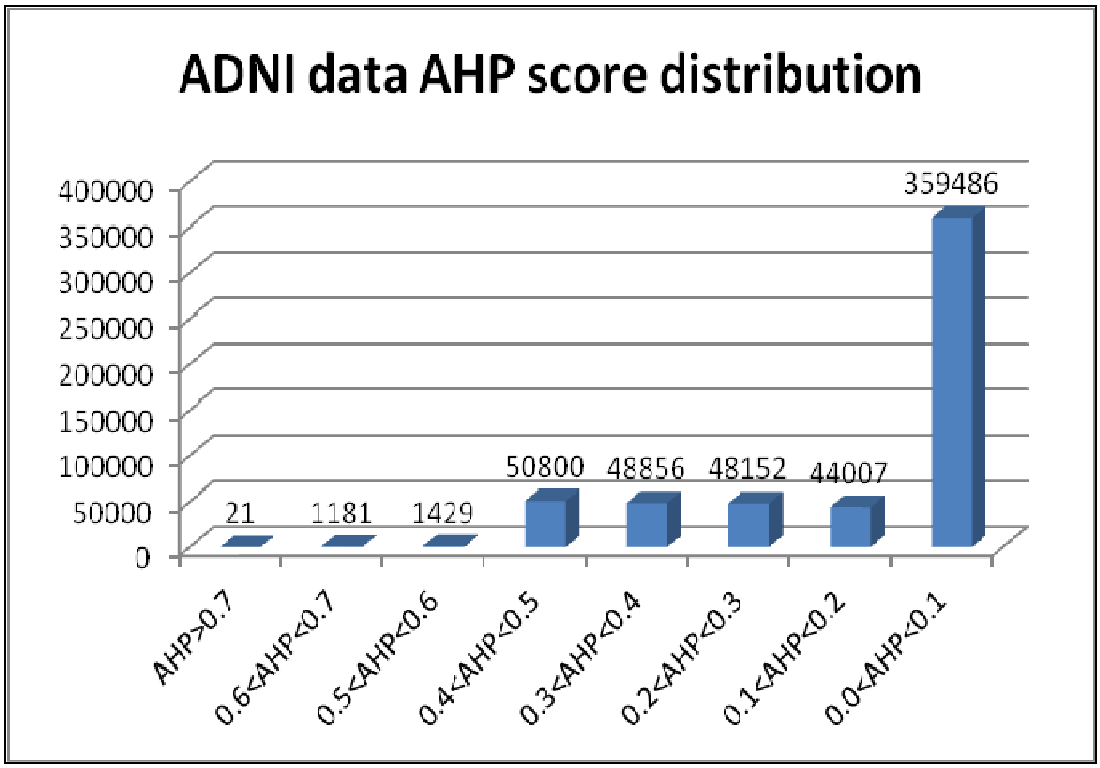



**Figure 5.1 AHP score distribution of ADNI genotyping data after all 500K SNPs are prioritized**

### 5.1.2 GenADA data

AHP prioritization is performed for all the SNPs in the data via choosing the p-value threshold as 1.0. Table 5.2 represents the AHP score distribution and AD linkage ratio after all 250K SNPs are AHP prioritized. AHP scores of SNPs range between 0.0 and 0.7.

**Table 5.2 AHP Score Distribution And Ratio Of SNPs Mapping To AD Linked Genes For GenADA Data**

| AHP score range | # of SNPs | # of SNPs mapped to AD linked genes | AD linkage ratio |
|---|---|---|---|
| AHP>0.6 | 66 | 26 | 0.393939394 |
| 0.5<AHP<0.6 | 432 | 141 | 0.326388889 |
| 0.4<AHP<0.5 | 2094 | 143 | 0.068290353 |
| 0.3<AHP<0.4 | 8471 | 64 | 0.007555188 |
| 0.2<AHP<0.3 | 10315 | 5 | 0.000484731 |
| 0.1<AHP<0.2 | 8515 | 0 | 0 |
| 0.0<AHP<0.1 | 232371 | 0 | 0 |

As can be seen from Figure 5.2, the AHP score distribution for the ADNI data involves an accumulation in AHP score range between 0.0 and 0.1.

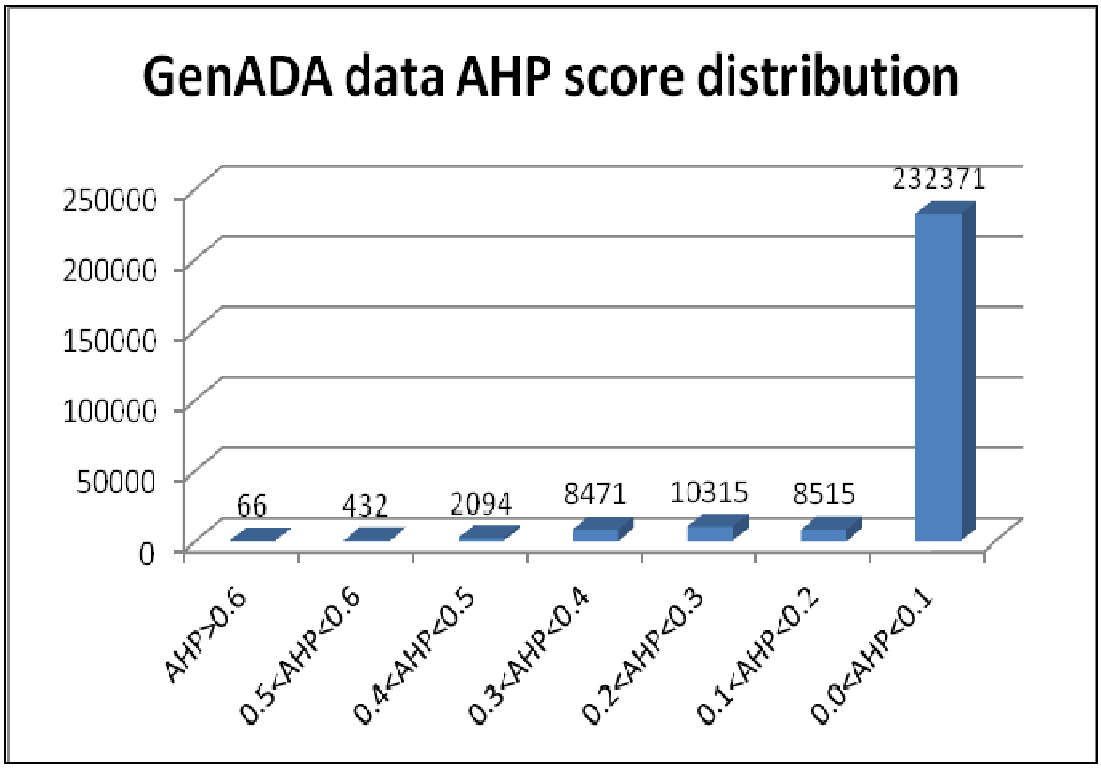



**Figure 5.2 AHP score distribution of GenADA genotyping data after all 250K SNPs are prioritized**

## 5.2 Post-priorization cutoff estimation for AD genotyping data SNPs

### 5.2.1 ADNI data

AD linkage ratio is considerably high for SNPs having AHP score higher than 0.6 as can be seen clearly from Figure 5.3. After an AHP score of 0.5, AD linked gene frequency for SNPs significantly lowers and eventually becomes 0.0. AHP score cutoff as a post-prioritization cutoff value for this data can be said to be 0.5. SNPs having AHP score higher than 0.5 are potential

associative SNPs since considerable percentage of them map to AD linked genes thus for this dataset, SNPs residing at the AHP score range higher than 0.5 can be interpreted as biologically relevant and meaningful SNPs.

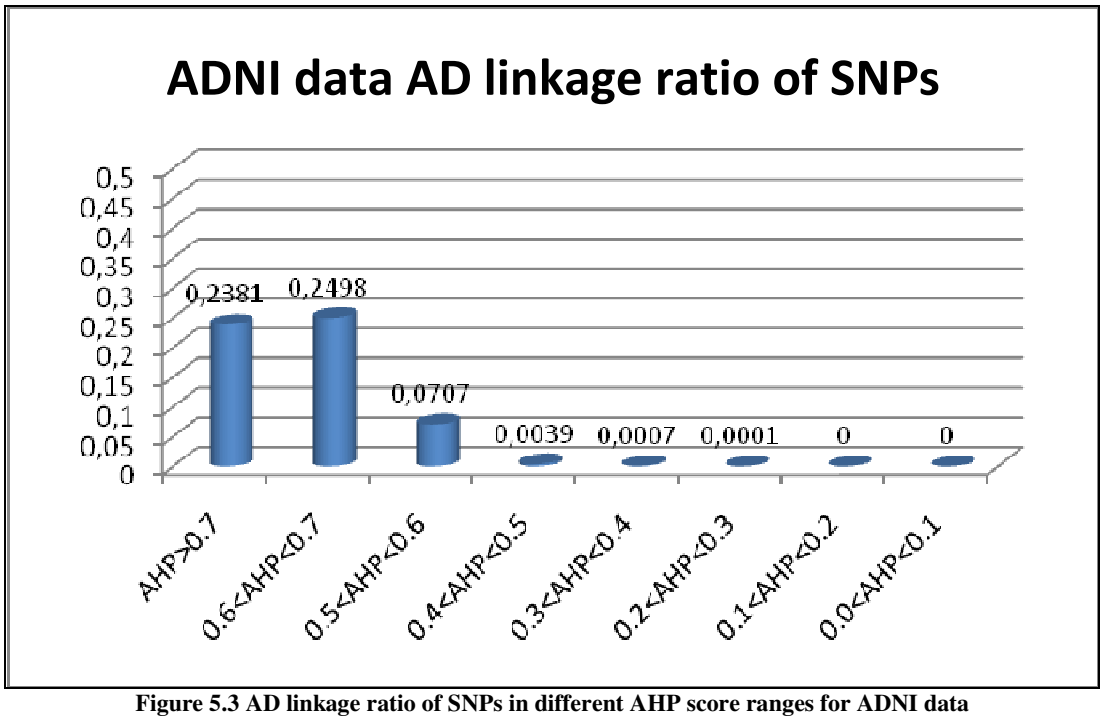



**Figure 5.3 AD linkage ratio of SNPs in different AHP score ranges for ADNI data**

### 5.2.2 GenADA data

For this dataset, AD linkage ratio is considerably high for SNPs with AHP scores higher than 0.5. Figure 5.4 depicts the AD linkage ratio for SNPs in different AHP score ranges. AHP score cutoff as a post-prioritization cutoff value for this data can be said to be 0.4 since AD linked gene frequency for SNPs significantly lowers after AHP score of 0.4. Considerable percentage of SNPs having AHP score higher than 0.4 map to AD linked genes. For GenADA data, SNPs having AHP score higher than 0.4 can be interpreted as biologically relevant and meaningful SNPs.

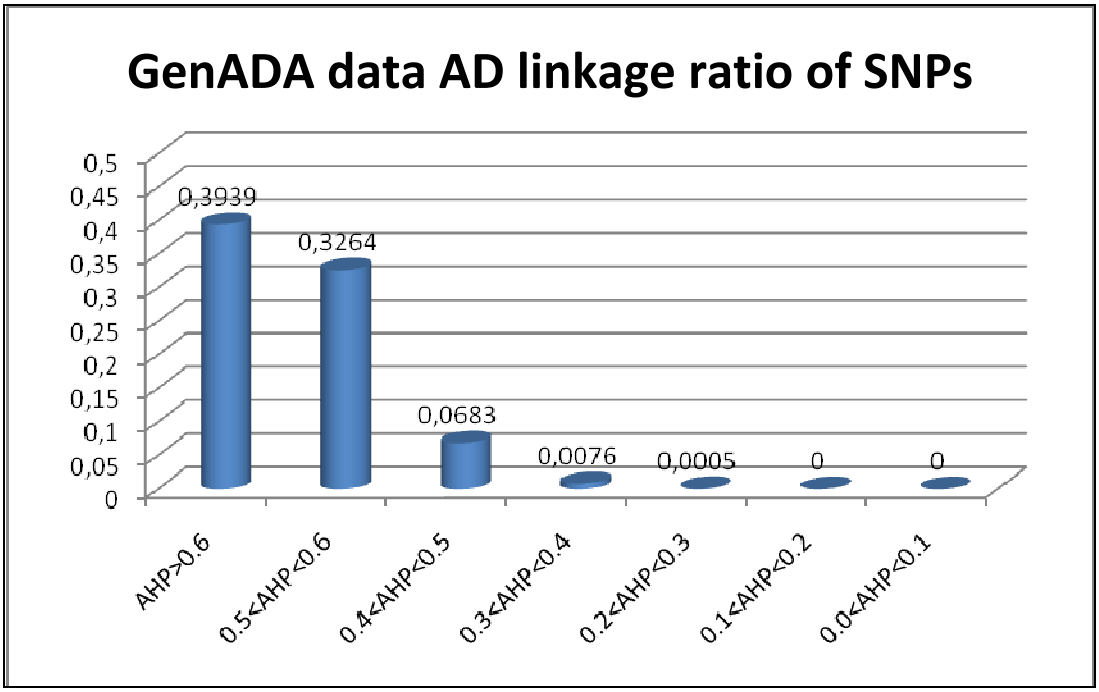



**Figure 5.4 AD linkage ratio of SNPs in different AHP score ranges for GenADA data**

### 5.2.3 AHP score cutoff classification performance for ADNI data

Classification performances in different AHP score ranges are also investigated for ADNI data after AHP prioritization to further estimate the post-prioritization cutoff value. Four AHP score ranges are determined as high, medium, low, very low and 5-fold cross validation tests are performed in those ranges by using 20000 SNPs. In order to perform classification comparison in equal conditions, AHP score increments of 0.1 are taken into account and the Top 20.000 SNP in each AHP score range is considered for the cross validation tests. For the high AHP score range, SNPs having AHP scores between 0.4 and 0.5, for the medium AHP score range SNPs having AHP scores between 0.3 and 0.4, for the low AHP score range SNPs having AHP scores between 0.2 and 0.3, for the very low AHP score range SNPs having AHP scores between 0.1 and 0.2 are considered. Table 5.3 depicts the test results in which Naive Bayes is used as the learning scheme whereas Table 5.4 depicts the test results in which SMO is used as the learning scheme. Both results imply that high AHP score range yields better performance measures. Thus, AHP score cutoff in the range of 0.4 and 0.5 is further supported for the ADNI data.

**Table 5.3** 5-Fold Cross Validation Test Results In Different AHP Score Ranges For ADNI Data(Learning Scheme:Naive Bayes, 20000 SNPs)

| | Specificity | Sensitivity | Accuracy | NPV | Precision |
|---|---|---|---|---|---|
| **High AHP score (0.4<AHP<0.5)** | **0.7692308** | **1.0** | **0.877551** | **1.0** | **0.793103** |
| **Medium AHP score (0.3<AHP<0.4)** | 0.3846154 | 0.9130435 | 0.632653 | 0.83 | 0.567568 |
| **Low AHP score (0.2<AHP<0.3)** | 0.1923077 | 0.8695652 | 0.510204 | 0.63 | 0.487805 |
| **Very Low AHP score (0.1<AHP<0.2)** | 0.3846154 | 0.8695652 | 0.612245 | 0.77 | 0.555556 |

**Table 5.4** 5-Fold Cross Validation Test Results In Different AHP Score Ranges For ADNI Data(Learning Scheme:SMO, 20000 SNPs)

| | Specificity | Sensitivity | Accuracy | NPV | Precision |
|---|---|---|---|---|---|
| **High AHP score (0.4<AHP<0.5)** | **0.6923077** | **1.0** | **0.836735** | **1.0** | **0.741935** |
| **Medium AHP score (0.3<AHP<0.4)** | 0.4230769 | 0.9565217 | 0.673469 | 0.92 | 0.594595 |
| **Low AHP score (0.2<AHP<0.3)** | 0.1153846 | 0.8695652 | 0.469388 | 0.5 | 0.465116 |
| **Very Low AHP score (0.1<AHP<0.2)** | 0.3461538 | 0.826087 | 0.571429 | 0.69 | 0.527778 |

# CHAPTER 6

# CONCLUSION AND FUTURE WORK

## 6.1 Discussion

Our main focus in this study was the evaluation of recently introduced AHP based SNP prioritization software in terms of biological relevance and performance measures by comparing its outputs with that of another SNP prioritization tool, SPOT. Two Alzheimer' s Disease (AD) genotyping data, ADNI and GenADA, are used in this study. Major aim of the AHP based SNP prioritization approach is to establish a strong connection between statistical analysis and biological relevance for a GWA study.

Chapter 1 begins with a brief biological background followed by genome wide association studies and SNP prioritization terms and tools. Major drawback of a GWA study is that it does not provide considerable biological relevance, it mainly focuses on statistical analysis and thus p-value computations. Widely used SNP prioritization tools are highly dependent on statistical analysis side rather than biological relevance. It is thus essential to eliminate this bottleneck of GWAS by boosting importance to biological relevance. Final part of this chapter focuses on the literature review of the AD.

Chapter 2 mainly involves GWA studies regarding Alzheimer's Disease and relevant literature information. Various genes and loci are found out to be associated with AD, those are clearly presented at Chapter 2 and selection of AD linked genes for the comparision list is described. Literature screening is performed and various databases are investigated to provide a reliable AD linked gene list We have build a unique AD linked gene list by integrating data from OMIM, AlzGene and AlzGenes databases. Utilizing this list allowed us to evaluate the biological relevance of the AHP prioritization results , independently from the GeneRIF data for SNP-gene-disease associations, AHP algorithm based on.

Chapter 3 focuses on AHP based SNP prioritization comparison with SPOT in terms of biological relevance with the GWAS results from both AD genotyping data. It is shown in Chapter 3 that, recently introduced AHP based SNP prioritization outperforms SPOT for both AD genotyping data in terms of biological relevance. Top 100 SNPs of the AHP list yield a much higher AD linkage ratio than that of the state of the art competitor application SPOT.

Next two chapters describes the determination of user defined performance parameters for AHP based SNP prioritization by using two approaches; pre-prioritization cutoff value (p-value threshold) estimation and post-prioritization cutoff value (AHP score threshold) estimations. Chapter 4 focuses on the former. p-value threshold determines ratio of SNPs to be selected for prioritization. Various p-value thresholds are tested and classification measures such as sensitivity, specificity, NPV and accuracy are compared at each p-value level for AHP based prioritization. Depending on our experience gained during these analysis we can suggest a guide for reserchers on how to choose appropriate p-value threshold depending on the goal of their experiment. Conservative p-value choice are suggested for focused targeted biomarker identification research, whereas moderate to less conservative p-values are suggested for larger scale association studies to investigate gene- pathway association results and building biological networks in order to reveal underlying etiology and novel association for the disease under study.

Chapter 5 involves AHP score cutoff value estimation as a post-prioritization cutoff. After AHP based prioritization is performed, SNPs are checked in terms of the genes they mapped. After a certain AHP score, AD linkage ratio of SNPs begins to decline significantly. Cutoff is defined in where a sharp decline in the AD linkage ratio occurs. Post-prioritization cutoff values are proposed for both AD genotyping data depending on our observations. AD linkage ratio of SNPs, which is the ratio of SNPs mapping to AD linked genes to the total number of SNPs in a certain AHP score range determines the AHP score cutoff value. AHP score distribution and the ratio of SNPs mapping to AD linked genes are investigated for two AD genotyping data. ADNI data yields 0.5 AHP score as the cutoff value whereas GenADA data yields 0.4 AHP score as the cutoff value since the AD linkage ratio sharply decline at AHP scores lower than those values. Moreover, classification performances are investigated in various AHP score ranges for ADNI data and the results further support the AHP score cutoff proposed since high AHP score range yields the best classification performance. For both data, the top ~2500 SNPs at the AHP ranking seem to carry the majority of the biological relevance. Researchers focusing on those data should put emphasis on the top 2500 SNPs to identify potential associative SNPs with AD.

**6.2 Future Work**

AHP score cutoff as the post-prioritization cutoff value should be investigated on other genotyping data sets to validate the findings at Chapter 5. AHP score of 0.5 seems as the common AHP score cutoff value for both datasets and a generalization for this suggestion can be defined after application of the AHP score cutoff value estimation strategy described in Chapter 5 on other data sets. Immediate candidate studies that are already under examination by other researchers in our group are GWS 16 RA data, HÜTF JIA, szhoprenia data.

The integrated METU-SNP database involves SNP information back in 2008. It is build on dbSNP 128, where today NCBI supports the newer version build db132. Our group is

developing the iSNP, which will be an Integrated, Automatically Updated SNP Database Server Over Web and it is scheduled to be finalized by December 2011. Analysis of the both AD Genotying dataset with METU-SNP and AHP prioritization approach that runs on the up-to-date iSNP database can confirm the finding of this study and might exploit new and novel SNP-gene-disease associations for AD.

In this study we have suggested association of various SNPs that map to AD linked genes or potential candidate genes supported by the literature. Most promising SNP sets should be selected for further experimental validation to prove these associations of the corresponding genes with AD, which can then be utilized as SNP biomarker for the prediction or early diagnosis of AD. Additionally new experimental studies should be designed to be able to validate the association of novel SNP-gene and pathways suggested in our study with AD and to exploit the biological networks and underlying etiology of AD.

**6.3 Conclusions**

In this study, two independent Alzheimer's Disease genotyping data are used and AHP based SNP prioritization approach is tested on them. The integrated system called METU-SNP presents a firm linkage between statistical analysis and biological relevance as proven by the corresponding comparisons to the most widely referred SNP prioritization tool, SPOT.

Performance parameters for AHP based prioritization are investigated as pre-prioritization (p-value cutoff) and post-prioritization (AHP score cutoff). Potential users of METU-SNP studying on those data are provided guidelines on p-value selection and AHP score cutoff value determination in order to define the biologically most relevant SNPs with AHP scores above the cutoff for both AD Genotyping data sets.

As presented here METU-SNP is a powerful tool for GWAS of SNP Genotyping data with a novel AHP based prioritization algorithm implemented, which can lead to discovery of new associations at SNP, gene and pathway level. In near future, we expect that these new associations described through GWAS here and in other studies will lead to development of personalized medicine approaches with application in pharmacogenomics and psychopharmacology.

# REFERENCES


[1] Alberts B, Johnson A, Lewis J, Raff M, Roberts K, and Walter P. *Molecular Biology of the Cell*, 4th ed. Garland Science; 2002.

[2] Azuaje F. *Bioinformatics and Biomarker Discovery*, 1st ed. Wiley-Blackwell; 2010.

[3] Lewin B. *Genes VIII*, 1st ed. Benjamin Cummings; 2004.

[4] http://gwas.nih.gov/01faq2.html#f1

[5] Kostem E, Lozano JA, Eskin E. Increasing Power of Genome-wide Association Studies by Collecting Additional SNPs Follow-up SNP Selection. *Science*. 2011:1-32.

[6] Hindorff LA, Junkins HA, Hall PN, Mehta JP, and Manolio TA. A Catalog of Published Genome-Wide Association Studies

[7] Yang C, Wan X, Yang Q, Xue H, Tang NL, Yu W. A hidden two-locus disease association pattern in genome wide association studies. *BMC bioinformatics*. 2011;12(1):156.

[8] Peng G, Luo L, Siu H, Zhu Y, Hu P, Hong S, Zhao J, Zhou X, Reveille JD, Jin L, Amos CI, Xiong M. Gene and pathway-based second-wave analysis of genome-wide association studies. *European journal of human genetics : EJHG*. 2010;18(1):111-7.

[9] Li M-X, Sham PC, Cherny SS, Song Y-Q. A knowledge-based weighting framework to boost the power of genome-wide association studies. *PloS one*. 2010;5(12):e14480.

[10] McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, Hirschhorn JN. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature reviews. Genetics*. 2008;9(5):356-69.

[11] http://gwas.nih.gov/09glossary.html

[12] Casto AM, Feldman MW. Genome-wide association study SNPs in the human genome diversity project populations: does selection affect unlinked SNPs with shared trait associations? *PLoS genetics*. 2011;7(1):e1001266.

[13] Cantor RM, Lange K, Sinsheimer JS. Prioritizing GWAS results: A review of statistical methods and recommendations for their application. *American journal of human genetics*. 2010;86(1):6-22.

[14] Almeida M a a, Oliveira PSL, Pereira TV, Krieger JE, Pereira AC. An empirical evaluation of imputation accuracy for association statistics reveals increased type-I error rates in genome-wide associations. *BMC genetics*. 2011;12:10.

[15] Weng L, Macciardi F, Subramanian A, Guffanti G, Potkin SG, Yu Z, Xie X. SNP-based pathway enrichment analysis for genome-wide association studies. *BMC bioinformatics*. 2011;12(1):99.

[16] Galvan A, Ioannidis JP a, Dragani T a. Beyond genome-wide association studies: genetic heterogeneity and individual predisposition to cancer. *Trends in genetics : TIG*. 2010;26(3):132-41.

[17] Oexle K, Meitinger T. Sampling GWAS subjects from risk populations. *Genetic epidemiology*. 2011;35(3):148-53.

[18] Chen L, Zhang L, Zhao Y, Xu L, Shang Y, Wang Q, Li W, Wang H, Li X. Prioritizing risk pathways: a novel association approach to searching for disease pathways fusing SNPs and pathways. *Bioinformatics (Oxford, England)*. 2009;25(2):237-42.

[19] http://www.sciencemag.org/content/318/5858/1842.full.pdf

[20] http://www.dailymail.co.uk/news/article-1355898/New-gene-based-tests-explain-childrens disabilities--highlight-cases-incest.html

[21] Hosking FJ, Dobbins SE, Houlston RS. Genome-wide association studies for detecting cancer susceptibility. *British medical bulletin*. 2011;97(1):27-46.

[22] http://bioinfo.cnio.es/files/training/Genome_Browsing_Course/variations.pdf

[23] Psychiatric GWAS Consortium Coordinating Committee, Cichon S, Craddock N, Daly M, Faraone SV, Gejman PV, Kelsoe J, Lehner T, Levinson DF, Moran A, Sklar P, Sullivan PF. Genomewide association studies: history, rationale, and prospects for psychiatric disorders. *The American journal of psychiatry*. 2009;166(5):540-56.

[24] Martin P, Barton A, Eyre S. ASSIMILATOR: a new tool to inform selection of associated genetic variants for functional studies. *Bioinformatics (Oxford, England)*. 2011;27(1):144-6.

[25] Zhao J, Gupta S, Seielstad M, Liu J, Thalamuthu A. Pathway-based analysis using reduced gene subsets in genome-wide association studies. *BMC bioinformatics*. 2011;12(1):17.

[26] Yuan H-Y, Chiou J-J, Tseng W-H, Liu CH, Liu CK, Lin YJ, Wang HH, Yao A, Chen YT, Hsu CN. FASTSNP: an always up-to-date and extendable service for SNP function analysis and prioritization. *Nucleic acids research*. 2006;34(Web Server issue):W635-41.

[27] Pico AR, Smirnov IV, Chang JS, Yeh RF, Wiemels JL, Wiencke JK, Tihan T, Conklin BR, Wrensch M. SNPLogic: an interactive single nucleotide polymorphism selection, annotation, and prioritization system. *Nucleic acids research*. 2009;37(Database issue):D803-9.

[28] Saccone SF, Bolze R, Thomas P, Quan J, Mehta G, Deelman E, Tischfield JA, Rice JP. SPOT: a web-based tool for using biological databases to prioritize SNPs after a genome wide association study. *Nucleic acids research*. 2010;38:201-209.

[29] Xu Z, Taylor J a. SNPinfo: integrating GWAS and candidate gene information into functional SNP selection for genetic association studies. *Nucleic acids research*. 2009;37(Web Server issue):W600-5.

[30] Terry H S, Christopher S C, and Tarczy-Hornoch P. SNPit: a federated data integration

system for the purpose of functional SNP annotation. *Comput Methods Programs Biomed*. 2009; 95(2): 181–189.

[31] http://people.revoledu.com/kardi/tutorial/AHP/AHP.htm

[32] http://thequalityportal.com/q_ahp.htm

[33] http://en.wikipedia.org/wiki/Analytic_Hierarchy_Process

[34] Üstünkar, G., 2011. An Integrative Approach to Structured SNP Prioritization and Representative SNP Selection for Genome-Wide Association Studies. Ph.D Thesis, Middle East Technical University.

[35] Üstünkar, G., Aydın Son, Y. METU-SNP: An Integrated Software System for SNP Complex Disease Association Analysis. *Journal of Integrative Bioinformatics* invited the paper to special Translational Bioinformatics Workshop Issue.

[36] Sensen C.W. *Essentials of Genornics and Bioinformatics*, 1st ed. Wiley-VCH; 2002.

[37] Burns LC, Minster RL, Demirci FY, Barmada MM, Ganguli M, Lopez OL, DeKosky ST, Kamboh MI. Replication study of genome-wide associated SNPs with late-onset Alzheimer's disease. *American journal of medical genetics. Part B, Neuropsychiatric genetics : the official publication of the International Society of Psychiatric Genetics*. 2011;156B(4):507-12.

[38] Hu X, Pickering E, Liu YC, Hall S, Fournier H, Katz E, Dechairo B, John S, Van Eerdewegh P, Soares H; Alzheimer's Disease Neuroimaging Initiative. Meta-Analysis for Genome-Wide Association Study Identifies Multiple Variants at the BIN1 Locus Associated with Late-Onset Alzheimer's Disease Bush A, ed. *PLoS ONE*. 2011;6(2):e16616.

[39] Ballard C, Gauthier S, Corbett A, Brayne C, Aarsland D, Jones E. Alzheimer's disease. *Lancet*. 2011;377(9770):1019-31.

[40] Ghiso J, Frangione B. Amyloidosis and Alzheimer's disease. *Advanced Drug Delivery Reviews*. 2002;54(12):1539-1551.

[41] Naj AC, Jun G, Beecham GW, Wang LS, Vardarajan BN, Buros J, Gallins PJ, Buxbaum JD, Jarvik GP, Crane PK, Larson EB, Bird TD, Boeve BF, Graff-Radford NR, De Jager PL, Evans D, Schneider JA, Carrasquillo MM, Ertekin-Taner N, Younkin SG, Cruchaga C, Kauwe JS, Nowotny P, Kramer P, Hardy J, Huentelman MJ, Myers AJ, Barmada MM, Demirci FY, Baldwin CT, Green RC, Rogaeva E, St George-Hyslop P, Arnold SE, Barber R, Beach T, Bigio EH, Bowen JD, Boxer A, Burke JR, Cairns NJ, Carlson CS, Carney RM, Carroll SL, Chui HC, Clark DG, Corneveaux J, Cotman CW, Cummings JL, DeCarli C, DeKosky ST, Diaz-Arrastia R, Dick M, Dickson DW, Ellis WG, Faber KM, Fallon KB, Farlow MR, Ferris S, Frosch MP, Galasko DR, Ganguli M, Gearing M, Geschwind DH, Ghetti B, Gilbert JR, Gilman S, Giordani B, Glass JD, Growdon JH, Hamilton RL, Harrell LE, Head E, Honig LS, Hulette CM, Hyman BT, Jicha GA, Jin LW, Johnson N, Karlawish J, Karydas A, Kaye JA, Kim R, Koo EH, Kowall NW, Lah JJ, Levey AI, Lieberman AP, Lopez OL, Mack WJ, Marson DC, Martiniuk F, Mash DC, Masliah E, McCormick WC, McCurry SM, McDavid AN, McKee AC, Mesulam M, Miller BL, Miller CA, Miller JW, Parisi JE, Perl DP, Peskind E, Petersen RC, Poon WW, Quinn JF, Rajbhandary RA, Raskind M, Reisberg B, Ringman JM, Roberson ED, Rosenberg RN, Sano M, Schneider LS, Seeley W, Shelanski ML, Slifer MA, Smith CD, Sonnen JA, Spina S, Stern RA, Tanzi RE, Trojanowski JQ, Troncoso JC, Van Deerlin VM, Vinters HV, Vonsattel JP, Weintraub

S, Welsh-Bohmer KA, Williamson J, Woltjer RL, Cantwell LB, Dombroski BA, Beekly D, Lunetta KL, Martin ER, Kamboh MI, Saykin AJ, Reiman EM, Bennett DA, Morris JC, Montine TJ, Goate AM, Blacker D, Tsuang DW, Hakonarson H, Kukull WA, Foroud TM, Haines JL, Mayeux R, Pericak-Vance MA, Farrer LA, Schellenberg GD. Common variants at MS4A4/MS4A6E, CD2AP, CD33 and EPHA1 are associated with late-onset Alzheimer's disease. *Nature Genetics*. 2011; 43(5):436-41.

[42] Bertram L, Tanzi RE. Genome-wide association studies in Alzheimer's disease. *Human molecular genetics*. 2009;18(R2):R137-45.

[43] Amouyel, P., Lambert, J.C. and the GWALZ Investigators. Genome wide association analysis of Alzheimer's disease. In 2009 International Conference on Alzheimer's Disease, 2009; Vol. S5-03-05, Alzheimer's Association, Vienna, Austria.

[44] Williams, J. Genome wide studies of Alzheimer's disease: the bigger the better. In 2009 International Conference on Alzheimer's Disease, Vol. PL-05.Plenary 5, Alzheimer's Association, Vienna, Austria.

[45] Zlokovic, B.V., Martel, C.L., Matsubara, E., McComb, J.G., Zheng, G., McCluskey, R.T., Frangione, B. and Ghiso, J. Glycoprotein 330/ megalin: probable role in receptor-mediated transport of apolipoprotein J alone and in a complex with Alzheimer disease amyloid beta at the blood-brain and blood-cerebrospinal fluid barriers. Proc. Natl. Acad. Sci. USA, 1996;93, 4229–4234.

[46] Tanzi, R.E., Moir, R.D. and Wagner, S.L. Clearance of Alzheimer's Abeta peptide: the many roads to perdition. *Neuron*, 2004; 43, 605–608.

[47] Ertekin-Taner N. Genetics of Alzheimer disease in the pre- and post-GWAS era. *Alzheimer's research & therapy*. 2010;2(1):3.

[48] Hardy J, Selkoe DJ: The amyloid hypothesis of Alzheimer's disease: progress and problems on the road to therapeutics. *Science* 2002; 297:353-356.

[49] Small SA, Duff K: Linking Abeta and tau in late-onset Alzheimer's disease: a dual pathway hypothesis. *Neuron* 2008, 60:534-542.

[50] Shen J, Kelleher RJ 3rd: The presenilin hypothesis of Alzheimer's disease: evidence for a loss-of-function pathogenic mechanism. *Proc Natl Acad Sci USA* 2007; 104:403-409.

[51] Ertekin-Taner N. Gene expression endophenotypes: a novel approach for gene discovery in Alzheimer's disease. *Molecular neurodegeneration*. 2011;6:31.

[52] Netzer WJ, Dou F, Cai D, Veach D, Jean S, Li Y, Bornmann WG, Clarkson B, Xu H, Greengard P. Gleevec inhibits beta- amyloid production but not Notch cleavage. *Proc Natl Acad Sci USA* 2003; 100: 12444–12449.

[53] He G, Luo W, Li P, Remmers C, Netzer WJ, Hendrick J, Bettayeb K, Flajolet M, Gorelick F, Wennogle LP, Greengard P. Gamma-secretase activating protein is a therapeutic target for Alzheimer's disease. *Nature* 2010; 467: 95–98.

[54] Harold D, Abraham R, Hollingworth P, Sims R, Gerrish A, Hamshere ML, Pahwa JS, Moskvina V, Dowzell K, Williams A, Jones N, Thomas C, Stretton A, Morgan AR, Lovestone S, Powell J, Proitsi P, Lupton MK, Brayne C, Rubinsztein DC, Gill M, Lawlor B, Lynch A, Morgan K, Brown KS, Passmore PA, Craig D, McGuinness B, Todd S, Holmes C, Mann D, Smith AD, Love S, Kehoe PG, Hardy J, Mead S, Fox N, Rossor M,

Collinge J, Maier W, Jessen F, Schürmann B, van den Bussche H, Heuser I, Kornhuber J, Wiltfang J, Dichgans M, Frölich L, Hampel H, Hüll M, Rujescu D, Goate AM, Kauwe JS, Cruchaga C, Nowotny P, Morris JC, Mayo K, Sleegers K, Bettens K, Engelborghs S, De Deyn PP, Van Broeckhoven C, Livingston G, Bass NJ, Gurling H, McQuillin A, Gwilliam R, Deloukas P, Al-Chalabi A, Shaw CE, Tsolaki M, Singleton AB, Guerreiro R, Mühleisen TW, Nöthen MM, Moebus S, Jöckel KH, Klopp N, Wichmann HE, Carrasquillo MM, Pankratz VS, Younkin SG, Holmans PA, O'Donovan M, Owen MJ, Williams J. Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease. *Nature Genetics* 2009; 41:1088–1093.

[55] Lambert J-C, Heath S, Even G, Campion D, Sleegers K, Hiltunen M, Combarros O, Zelenika D, Bullido MJ, Tavernier B, Letenneur L, Bettens K, Berr C, Pasquier F, Fiévet N, Barberger-Gateau P, Engelborghs S, De Deyn P, Mateo I, Franck A, Helisalmi S, Porcellini E, Hanon O; European Alzheimer's Disease Initiative Investigators, de Pancorbo MM, Lendon C, Dufouil C, Jaillard C, Leveillard T, Alvarez V, Bosco P, Mancuso M, Panza F, Nacmias B, Bossù P, Piccardi P, Annoni G, Seripa D, Galimberti D, Hannequin D, Licastro F, Soininen H, Ritchie K, Blanché H, Dartigues JF, Tzourio C, Gut I, Van Broeckhoven C, Alpérovitch A, Lathrop M, Amouyel P. Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease. *Nature Genetics* 2009; 41:1094-1099.

[56] Hollingworth P, Harold D, Sims R, Gerrish A, Lambert J-C, Carrasquillo MM, Abraham R, Hamshere ML, Pahwa JS, Moskvina V, Dowzell K, Jones N, Stretton A, Thomas C, Richards A, Ivanov D, Widdowson C, Chapman J, Lovestone S, Powell J, Proitsi P, Lupton MK, Brayne C, Rubinsztein DC, Gill M, Lawlor B, Lynch A, Brown KS, Passmore PA, Craig D, McGuinness B, Todd S, Holmes C, Mann D, Smith AD, Beaumont H, Warden D, Wilcock G, Love S, Kehoe PG, Hooper NM, Vardy ER, Hardy J, Mead S, Fox NC, Rossor M, Collinge J, Maier W, Jessen F, Rüther E, Schürmann B, Heun R, Kölsch H, van den Bussche H, Heuser I, Kornhuber J, Wiltfang J, Dichgans M, Frölich L, Hampel H, Gallacher J, Hüll M, Rujescu D, Giegling I, Goate AM, Kauwe JS, Cruchaga C, Nowotny P, Morris JC, Mayo K, Sleegers K, Bettens K, Engelborghs S, De Deyn PP, Van Broeckhoven C, Livingston G, Bass NJ, Gurling H, McQuillin A, Gwilliam R, Deloukas P, Al-Chalabi A, Shaw CE, Tsolaki M, Singleton AB, Guerreiro R, Mühleisen TW, Nöthen MM, Moebus S, Jöckel KH, Klopp N, Wichmann HE, Pankratz VS, Sando SB, Aasly JO, Barcikowska M, Wszolek ZK, Dickson DW, Graff-Radford NR, Petersen RC; Alzheimer's Disease Neuroimaging Initiative, van Duijn CM, Breteler MM, Ikram MA, DeStefano AL, Fitzpatrick AL, Lopez O, Launer LJ, Seshadri S; CHARGE consortium, Berr C, Campion D, Epelbaum J, Dartigues JF, Tzourio C, Alpérovitch A, Lathrop M; EADI1 consortium, Feulner TM, Friedrich P, Riehle C, Krawczak M, Schreiber S, Mayhaus M, Nicolhaus S, Wagenpfeil S, Steinberg S, Stefansson H, Stefansson K, Snaedal J, Björnsson S, Jonsson PV, Chouraki V, Genier-Boley B, Hiltunen M, Soininen H, Combarros O, Zelenika D, Delepine M, Bullido MJ, Pasquier F, Mateo I, Frank-Garcia A, Porcellini E, Hanon O, Coto E, Alvarez V, Bosco P, Siciliano G, Mancuso M, Panza F, Solfrizzi V, Nacmias B, Sorbi S, Bossù P, Piccardi P, Arosio B, Annoni G, Seripa D, Pilotto A, Scarpini E, Galimberti D, Brice A, Hannequin D, Licastro F, Jones L, Holmans PA, Jonsson T, Riemenschneider M, Morgan K, Younkin SG, Owen MJ, O'Donovan M, Amouyel P, Williams J. Common variants at ABCA7, MS4A6A/MS4A4E, EPHA1, CD33 and CD2AP are associated with Alzheimer's disease. *Naure. Genetics*. 2011; doi:10.1038/ng.803.

[57] Naj AC, Jun G, Beecham GW, Wang L-S,Vardarajan BN, Buros J, Gallins PJ, Buxbaum JD, Jarvik GP, Crane PK, Larson EB, Bird TD, Boeve BF, Graff-Radford NR, De Jager PL, Evans D, Schneider JA, Carrasquillo MM, Ertekin-Taner N, Younkin SG, Cruchaga C, Kauwe JS, Nowotny P, Kramer P, Hardy J, Huentelman MJ, Myers AJ, Barmada MM, Demirci FY, Baldwin CT, Green RC, Rogaeva E, St George-Hyslop P, Arnold SE, Barber

R, Beach T, Bigio EH, Bowen JD, Boxer A, Burke JR, Cairns NJ, Carlson CS, Carney RM, Carroll SL, Chui HC, Clark DG, Corneveaux J, Cotman CW, Cummings JL, DeCarli C, DeKosky ST, Diaz-Arrastia R, Dick M, Dickson DW, Ellis WG, Faber KM, Fallon KB, Farlow MR, Ferris S, Frosch MP, Galasko DR, Ganguli M, Gearing M, Geschwind DH, Ghetti B, Gilbert JR, Gilman S, Giordani B, Glass JD, Growdon JH, Hamilton RL, Harrell LE, Head E, Honig LS, Hulette CM, Hyman BT, Jicha GA, Jin LW, Johnson N, Karlawish J, Karydas A, Kaye JA, Kim R, Koo EH, Kowall NW, Lah JJ, Levey AI, Lieberman AP, Lopez OL, Mack WJ, Marson DC, Martiniuk F, Mash DC, Masliah E, McCormick WC, McCurry SM, McDavid AN, McKee AC, Mesulam M, Miller BL, Miller CA, Miller JW, Parisi JE, Perl DP, Peskind E, Petersen RC, Poon WW, Quinn JF, Rajbhandary RA, Raskind M, Reisberg B, Ringman JM, Roberson ED, Rosenberg RN, Sano M, Schneider LS, Seeley W, Shelanski ML, Slifer MA, Smith CD, Sonnen JA, Spina S, Stern RA, Tanzi RE, Trojanowski JQ, Troncoso JC, Van Deerlin VM, Vinters HV, Vonsattel JP, Weintraub S, Welsh-Bohmer KA, Williamson J, Woltjer RL, Cantwell LB, Dombroski BA, Beekly D, Lunetta KL, Martin ER, Kamboh MI, Saykin AJ, Reiman EM, Bennett DA, Morris JC, Montine TJ, Goate AM, Blacker D, Tsuang DW, Hakonarson H, Kukull WA, Foroud TM, Haines JL, Mayeux R, Pericak-Vance MA, Farrer LA, Schellenberg GD. Common variants at MS4A4/MS4A6E, CD2AP, CD33 and EPHA1 are associated with late-onset Alzheimer's disease. *Nature Genetics* 2011; doi:10.1038/ng.801.

[58] Morgan K. Commentary: The three new pathways leading to Alzheimer's disease. *Neuropathology and applied neurobiology*. 2011.

[59] Bertram L, McQueen MB, Mullin K, Blacker D, Tanzi RE. "Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database." *Nature Genetics* 2007; 39(1): 17-23.

[60] http://www.polygenicpathways.co.uk/alzpolys.html

[61] http://omim.org/entry/107741

[62] http://adni.loni.ucla.edu

[63] http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000219.v1.p1

[64] Peng G, Luo L, Siu H, Zhu Y, Hu P, Hong S, Zhao J, Zhou X, Reveille JD, Jin L, Amos CI, Xiong M. "Gene and pathway-based second-wave analysis of genome-wide association studies." *European Journal of Human Genetics: EJHG*, 2010; vol. 18, no. 1, pp. 111-117.

[65] Youn H, Jeoung M, Koo Y, Ji H, Markesbery WR, Ji I, Ji TH. "Kalirin is under-expressed in Alzheimer's disease hippocampus." *Journal of Alzheimer's Disease* 2007; 11(3): 385-397.

[66] Chaudhury AR, Gerecke KM, Wyss JM, Morgan DG, Gordon MN, Carroll SL. "Neuregulin-1 and erbB4 immunoreactivity is associated with neuritic plaques in Alzheimer disease brain and in a transgenic model of Alzheimer disease." *J Neuropathol Exp Neurol* 2003;62(1):42-54.

[67] Ertekin-Taner N, Graff-Radford N, Younkin LH, Eckman C, Baker M, Adamson J, Ronald J, Blangero J, Hutton M, Younkin SG. Linkage of plasma Abeta42 to a quantitative locus on chromosome 10 in late-onset Alzheimer's disease pedigrees. *Science (New York, N.Y.)*. 2000;290(5500):2303-4.

[68] Ertekin-Taner N, Ronald J, Asahara H, Younkin L, Hella M, Jain S, Gnida E, Younkin S,

Fadale D, Ohyagi Y, Singleton A, Scanlin L, de Andrade M, Petersen R, Graff-Radford N, Hutton M, Younkin S. Fine mapping of the α-T catenin gene to a quantitative trait locus on chromosome 10 in late-onset Alzheimer's disease pedigrees. *Human Molecular Genetics*. 2003;12(23): 3133–3143.

[69] Shafaati M, Marutle A, Pettersson H, Lövgren-Sandblom A, Olin M, Pikuleva I, Winblad B, Nordberg A, Björkhem I. Marked accumulation of 27-hydroxycholesterol in the brain of Alzheimer patients with the Swedish APP 670 / 671 mutation. *Laboratory Medicine*.1-27.

[70] Leoni V., and Caccia C. Oxysterols as biomarkers in neurodegenerative diseases. *Chem. Phys. Lipids* 2011 doi:10.1016/j.chemphyslip.2011.04.002.

[71] Yau JL, Rasmuson S, Andrew R, Graham M, Noble J, Olsson T, Fuchs E, Lathe R, Seckl JR. "Dehydroepiandrosterone 7-hydroxylase CYP7B: predominant expression in primate hippocampus and reduced expression in Alzheimer's disease." *Neuroscience* 2003;121(2):307-14.

[72] Mashayekhi F, Hadavi M, Vaziri HR, Naji M. "Increased acidic fibroblast growth factor concentrations in the serum and cerebrospinal fluid of patients with Alzheimer's disease." *J Clin Neurosci*. 2010; 17(3):357-9.

[73] Russo, R., Borghi, R., Markesbery, W., Tabaton, M., Piccini, A. Neprylisin (sic) decreases uniformly in Alzheimer's disease and in normal aging. *FEBS Letters* 2005; 579: 6027-6030.

[74] Wu Y-H, Fischer DF, Swaab DF. A promoter polymorphism in the monoamine oxidase A gene is associated with the pineal MAOA activity in Alzheimer's disease patients. *Brain research*. 2007;1167:13-9.

[75] Jacobsen E, Beach T, Shen Y, Li R, Chang Y. Deficiency of the Mre11 DNA repair complex in Alzheimer's disease brains. *Brain research. Molecular brain research*. 2004;128(1):1-7.

[76] Pirskanen M, Hiltunen M, Mannermaa A, Helisalmi S, Lehtovirta M, Hänninen T, Soininen H. Estrogen receptor beta gene variants are associated with increased risk of Alzheimer's disease in women. *European journal of human genetics : EJHG*. 2005;13(9):1000-6.

[77] Tan Z, Sun X, Hou F-S, Oh HW, Hilgenberg LG, Hol EM, van Leeuwen FW, Smith MA, O'Dowd DK, Schreiber SS. Mutant ubiquitin found in Alzheimer's disease causes neuritic beading of mitochondria in association with neuronal degeneration. *Cell death and differentiation*. 2007;14(10):1721-32.

[78] Li X, Alafuzoff I, Soininen H, Winblad B, Pei J-J. Levels of mTOR and its downstream targets 4E-BP1, eEF2, and eEF2 kinase in relationships with tau in Alzheimer's disease brain. *The FEBS journal*. 2005;272(16):4211-20.

[79] http://www.genatlas.org/

[80] Feld M, Kipp M, Ndiaye A, Heckmann D. Weka : Practical machine learning tools and techniques with Java implementations. *Seminar*. 2007.

# APPENDICES

# APPENDIX A: GWAS TERMINOLOGY

***Bayes' factors***: Alternative approach for hypothesis testing similar to likelihood ratio tests: prior and posterior probabilites are involved, strength of the evidence is measured in favour of one model rather than the other.

***Case–control design***: Primary comparison is performed between case and control subjects, former are observed to involve the phenotype of interest and predicted to have a high prevalence of susceptibility alleles for that trait, latter do not have the phenotype of interest and considered to have a lower prevalence of susceptibility alleles.

***Cochran–Armitage test***: A genotype-based contingency-table test for association suitable for the detection of trends across ordinal genotypes. It aims to assess for the presence of an association between a variable with two categories and a variable with *k* categories. It is often used as a genotype-based test for case-control genetic association studies.

***Common SNPs***: They involve a frequency greater than or equal to 5%. There are approximately 10 million common SNPs in the human genome and approximately 2.8 million on the current HapMap. Common SNPs are the main targets of GWA studies.

***Common variant-common disease hypothesis***: Common SNPs contribution of genetic risk to the formation and predisposition of diseases.

***Complex disease***: It occurs due to various genetic and environmental factors. Interaction of multiple factors is involved.

***Copy number variant (CNV)***: A type of a variation in a genome in which the result is a departure from the expected diploid makeup of a DNA sequence. Deletion or duplication in a chromosomal segment indicates a copy number variation. Other structural variants are inversions and translocations.

***Cryptic relatedness***: The residual degrees of relatedness among GWAS samples can violate the independence assumptions of standard statistical techniques.

***DNA pooling approaches***: Estimates of allele frequencies originated from pools of DNA acquired from multiple subjects rather than individual DNA samples are used in this approach.

***False-positive report probability***: The probability that a reported association between a trait of interest and a genetic variant is not true.

***Family-based association methods***: Association studies is performed within families, this approach provides a protection from population substructure with a cost of reduced sensitivity.

***Frequentist***: This approach uses p-values and combines them with hypothesis testing to make statistical inferences. It aims to draw conclusions from statistical samples.

***Genome-wide association studies***: Genetic markers capturing a substantial proportion of common variation is typed in a set of DNA samples that are informative for a trait or disease of interest. Major aim is to map susceptibility variants through the associations between genotype frequency and trait status.

***Global allele frequency***: Its distribution is essential for GWAS for two main reasons. First, the frequency of a trait-associated allele determines the degree it can contribute to variability in its phenotype in a given population. This is particularly true for SNPs that contribute directly to phenotypic variation rather than tagging causative variants. Second, large allele frequency differences between populations for trait-associated SNPs may indicate that selection has acted upon the trait.

***Haplotype***: a set of SNPs on a single chromatid that are statistically associated

***Haplotype-based methods***: They rely on the relationship between the distribution of estimated haplotype and trait status.

***Hardy–Weinberg equilibrium (HWE)***: It represents the relationship between genotype and allele frequencies dependent on random mating in a stable population in the absence of selection, new mutations and gene flow.Departures from equilibrium can emphasize genotyping errors.

***Heritability***: It can be defined as the proportion of the phenotype variance that is due to genes, estimated from risks to twins and other relatives.

***Imputation methods***: They focus on filling in missing genotype data using a sparse set of genotypes and a scaffold of linkage disequilibrium relationships. Imputation involves the use of additional information to predict missing values in a sample.

***Informative missingness***: Nonrandom missing data pattern with repsect to both genotype and trait status might lead to misleading associations in the analysis of the available genotypes.

***Linkage disequilibrium (LD)***: Alleles at nearby variants might be allocated to individual chromosomes which can be caused by recent mutation, genetic drift or selection, leading to correlations between genotypes at closely linked markers. Some combinations of SNP alleles occur more or less frequently in a population than would be expected from a random formation of haplotypes. SNPs that are in high LD with their neighbors tag larger regions than do SNPs that are not in LD with surrounding variants. The former are more likely to mark a genomic region contaning a causative variant for a particular trait.

***Mendelian disease***: Type of disease that is caused by a usually rare mutation in DNA sequence on one (dominant) or both (recessive) of an individual's pair of chromosomes.

***Mendelian randomization***: It makes it possible to test for a causal relationship between two phenotypes involving observational associations, but are subject to confounding. Random segregation of susceptibility alleles at meiosis is used to explore causality in a model that is freed from most sources of confounding.

***Minor allele frequency***: It is the frequency of the less common allele. SNPs with MAF greater than 5% are the targets of HapMap project. Trait-SNP associations that involve SNPs with high MAF are more prone to be detected by GWA studies.

***Misclassification bias***: It is due to incorrectly assignment of individuals to the relevant group in a case-control study. For instance, some individuals in the control group might meet the criterias to be case subjects or vice versa.

***Multiple rare variant hypothesis***: Some of the rare SNPs might involve genetic risk to diseases, especially if they are located in protein coding or gene regulatory regions.

***Pleiotropy***: A single allele might affect several distinct aspects of the phenotype of an organism, often traits not previously thought to be related.

***Population stratification***: Different ancestral and demographic histories among the samples might lead to a misclassification that, markers which are informative for them might be confounded with disease status and lead to artificial associations.

***Quantile-quantile plot (Q-Q plot)***: A diagnostic plot comparing the distribution of observed test statistics with the distribution expected under the null.

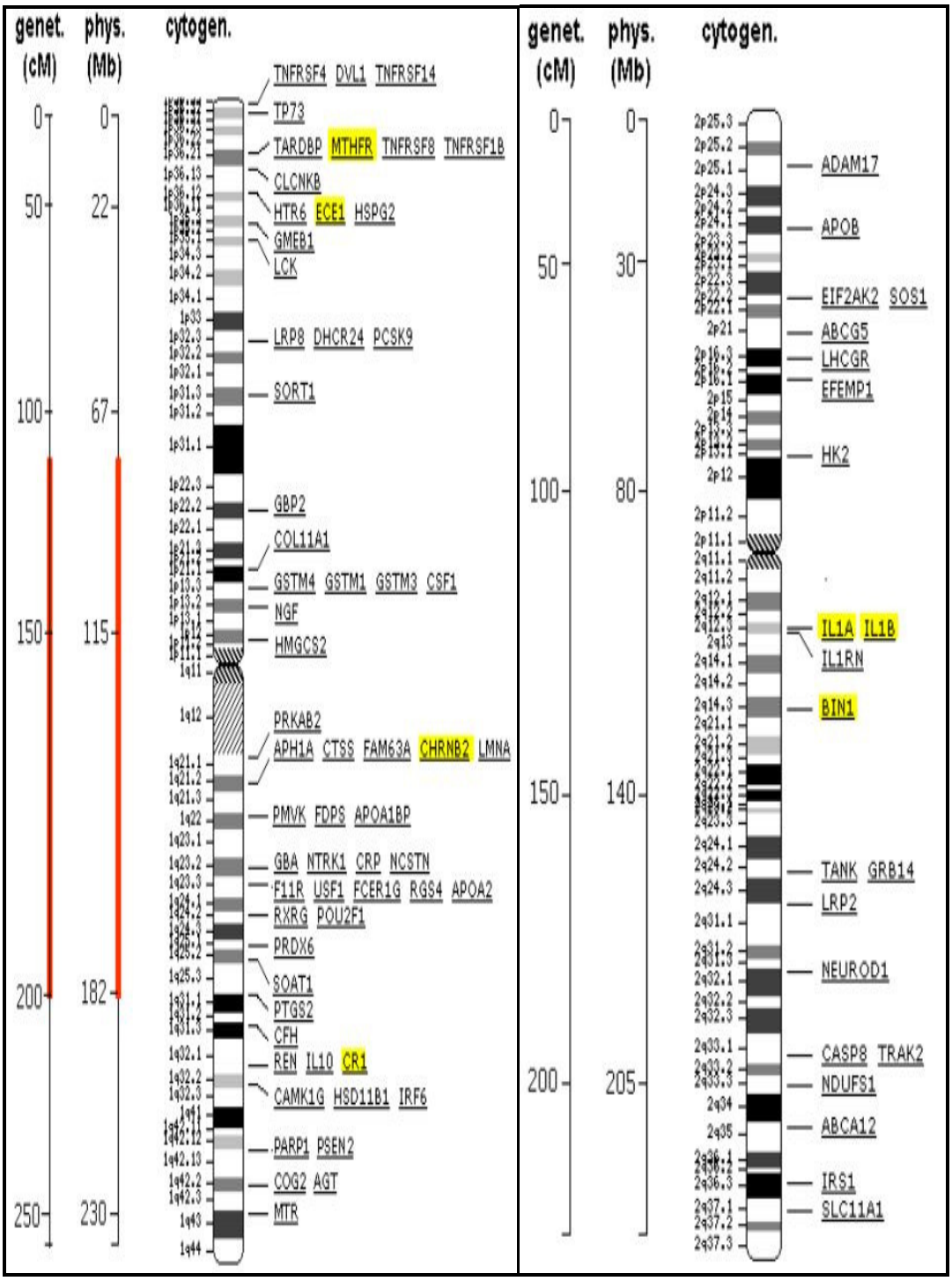
***Rare SNPs***: They are in a frequency of less than 1%. The ones at the coding regions are more harmful than those at the other regions of the genome.

***Selection bias***: It arises due to the fact that the samples ascertained for the study, particularly controls might not represent the wider population that they are expected to represent.

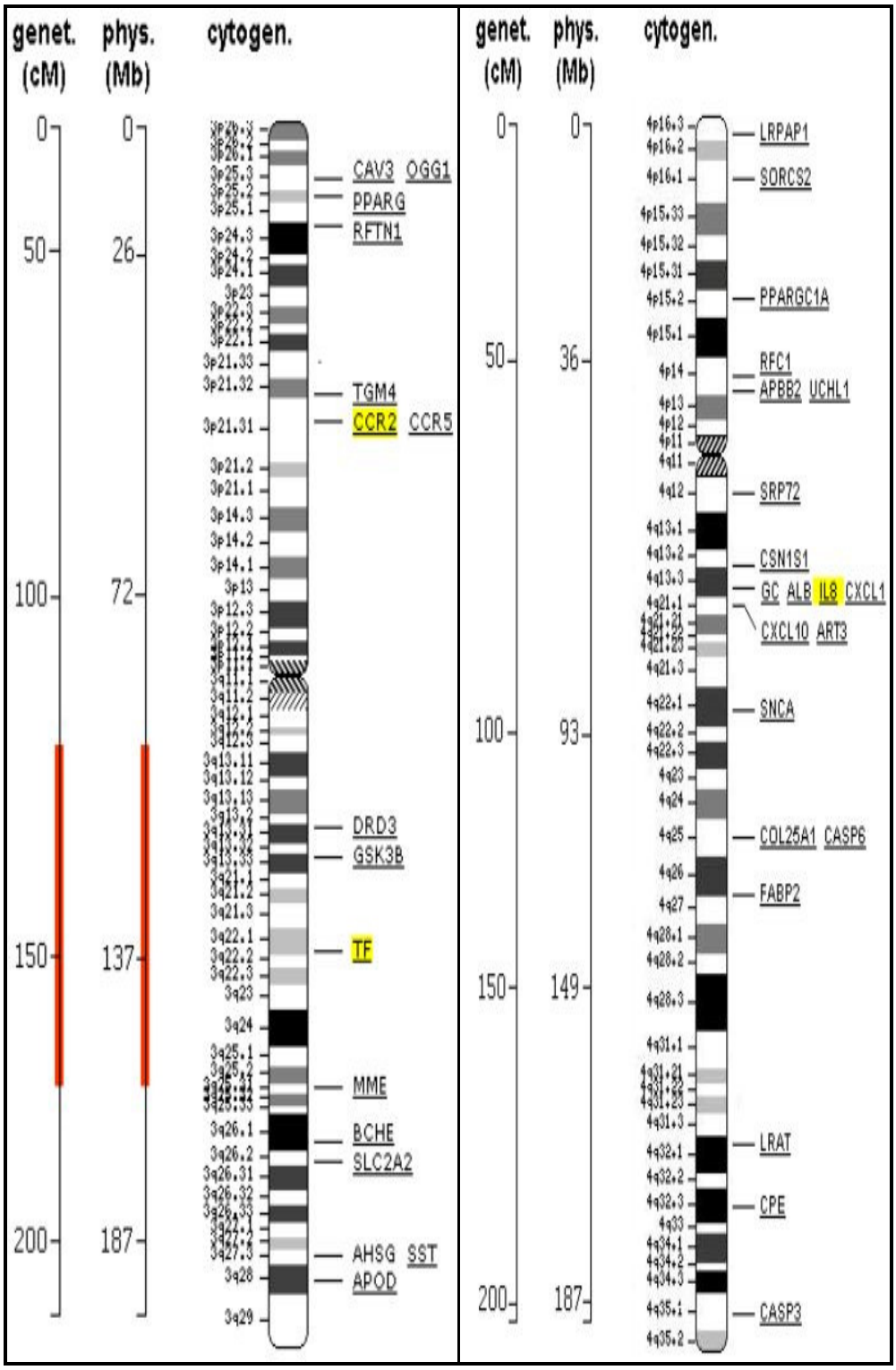
***Signal intensity (cluster) plots***: Raw intensity data plots for individual variants generated by the genotyping platform. They provide a useful visual diagnostic to estimate the data quality of the genotyping data since those plots represent the extent to which the various genotypes can be discriminated.

***SNP***: Specific position on the genome where chromosomes carry different nucleic acids. There are approximately 15 million SNPs with frequency ≥1%. HapMap project involves about 4 million of them.

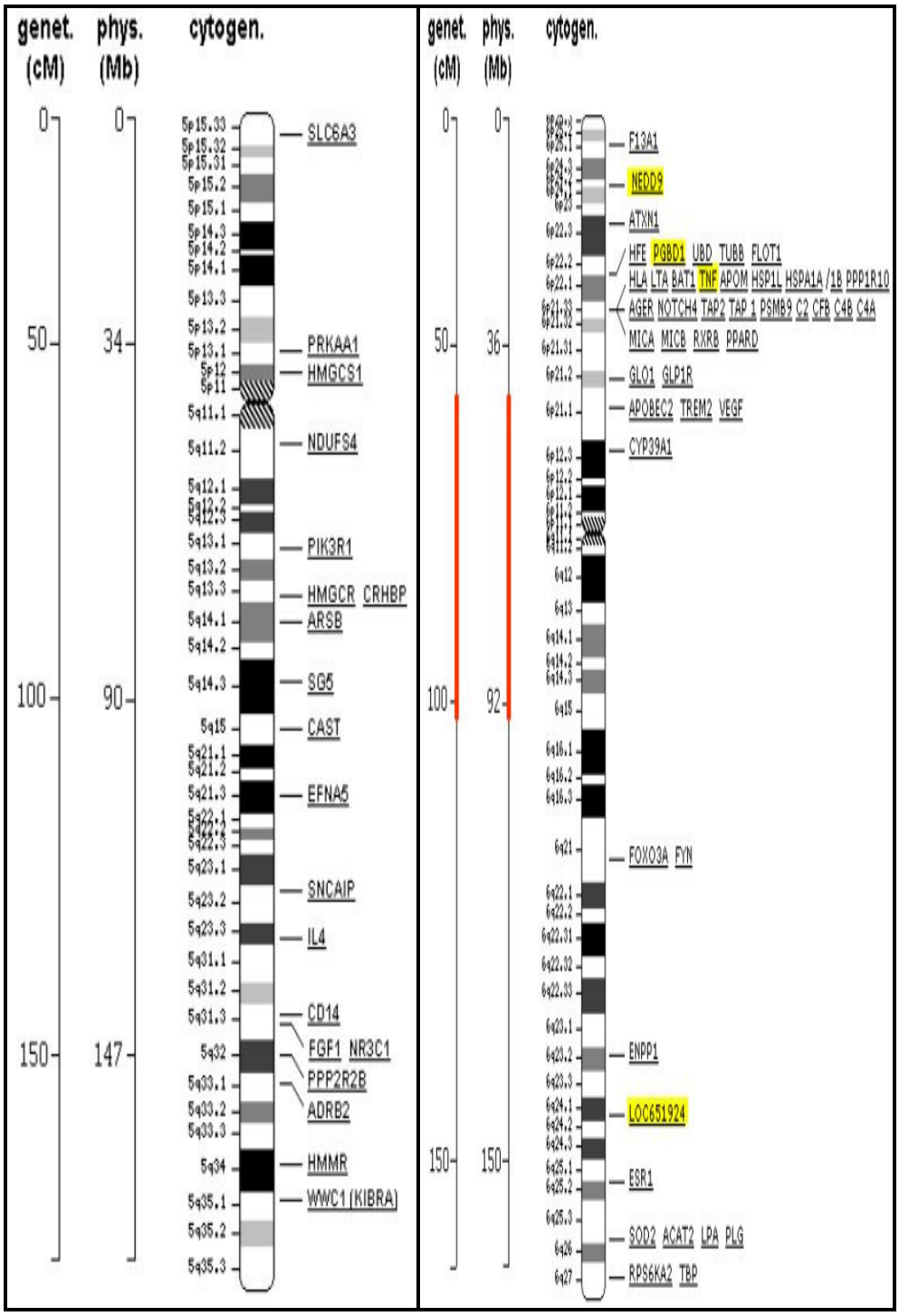
# APPPENDIX B: GENES AND LOCUS ON HUMAN CHROMOSOMES FOUND TO BE POTENTIALLY ASSOCIATED WITH AD



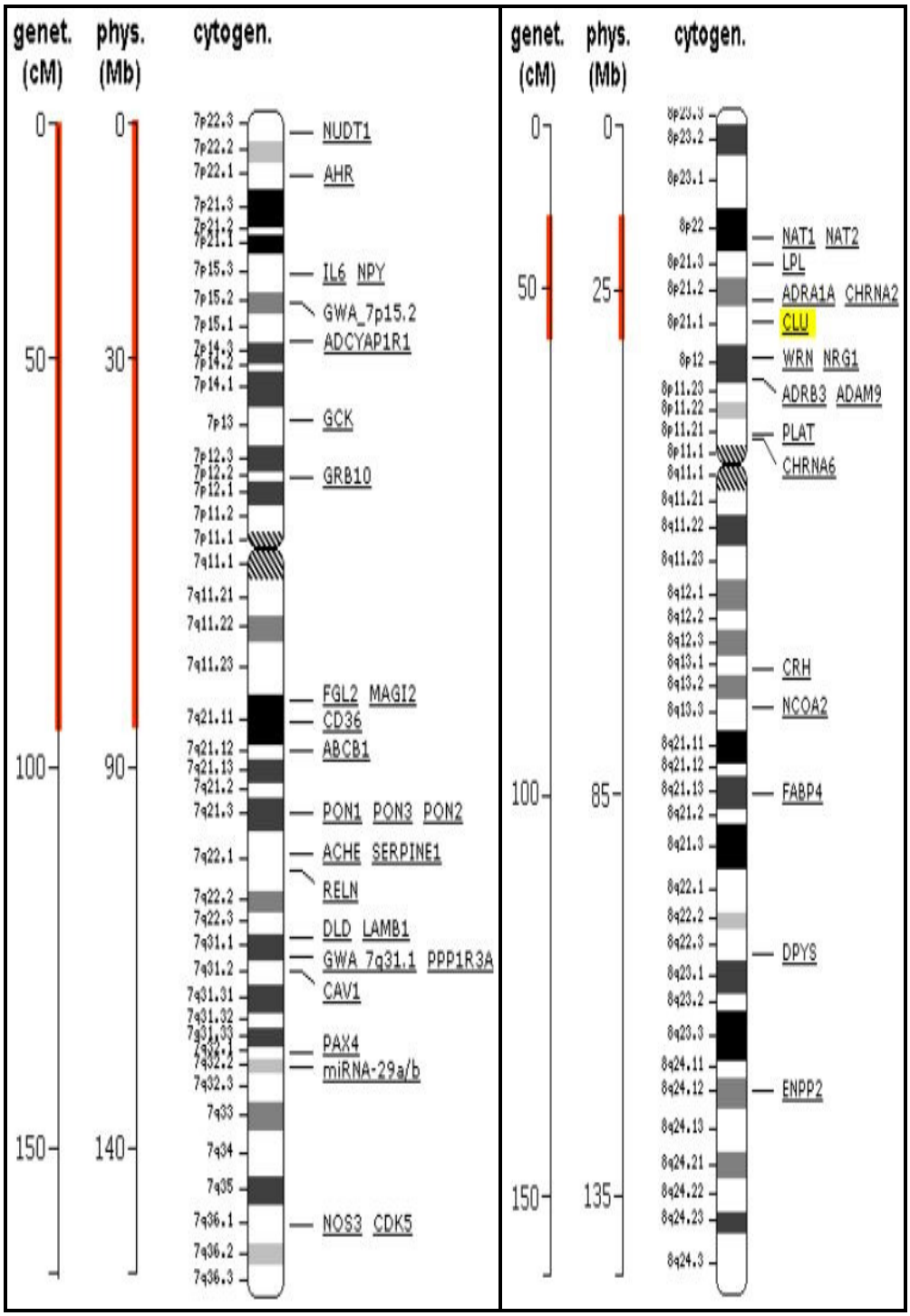

**Published AD candidate genes and locus (chromosomes 1 and 2)**



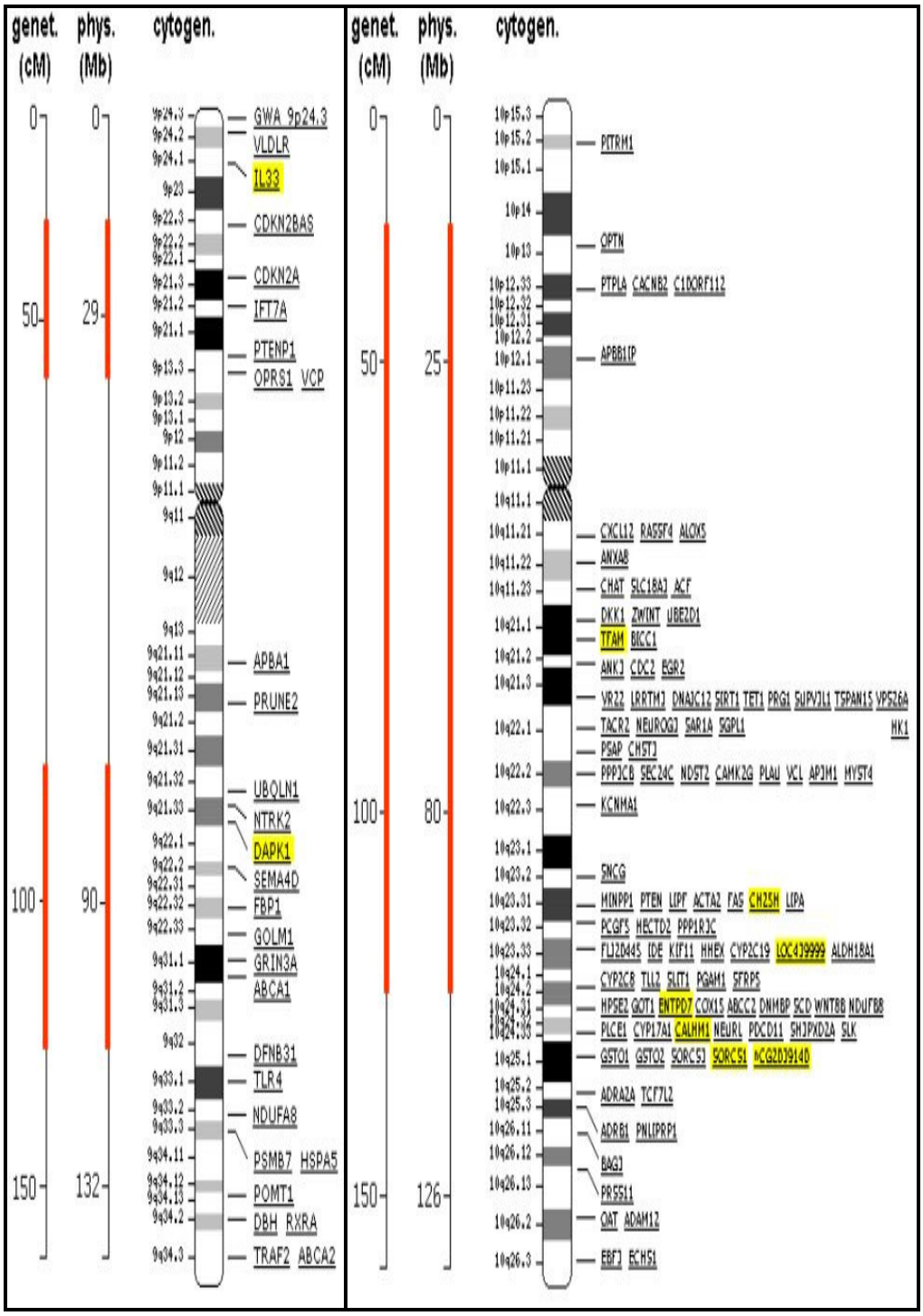

**Published AD candidate genes and locus (chromosomes 3 and 4)**



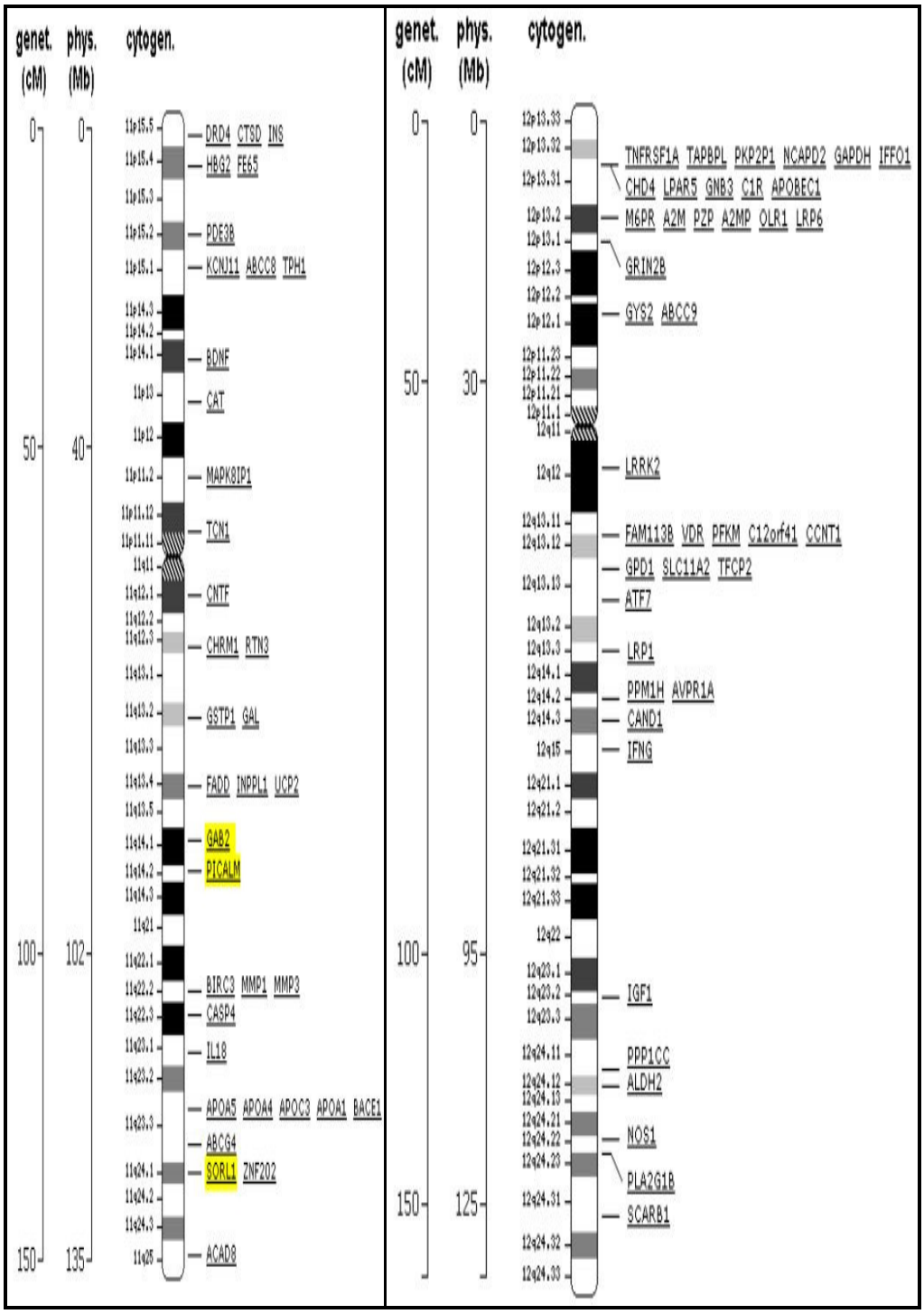

**Published AD candidate genes and locus (chromosomes 5 and 6)**



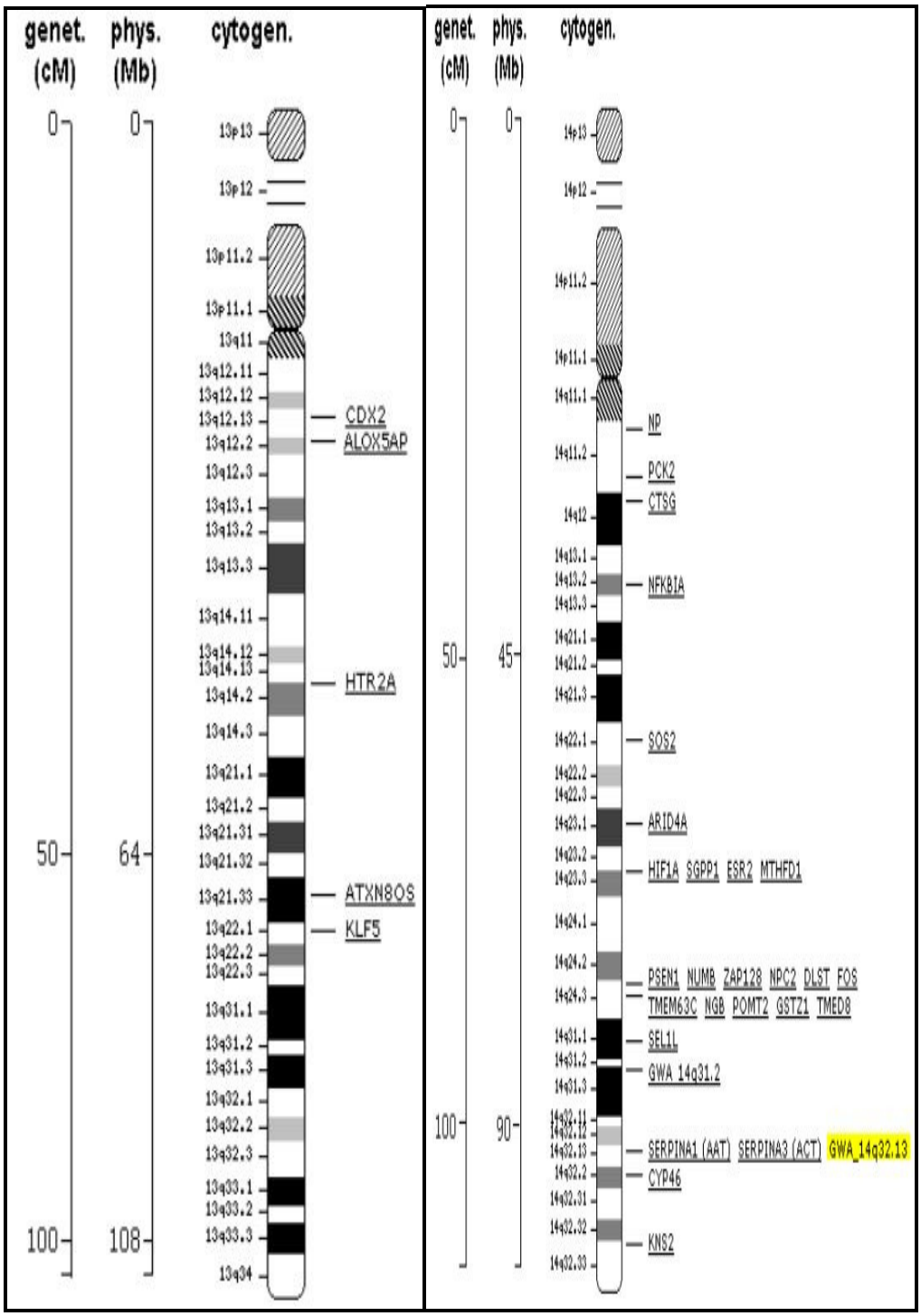

**Published AD candidate genes and locus (chromosomes 7 and 8)**



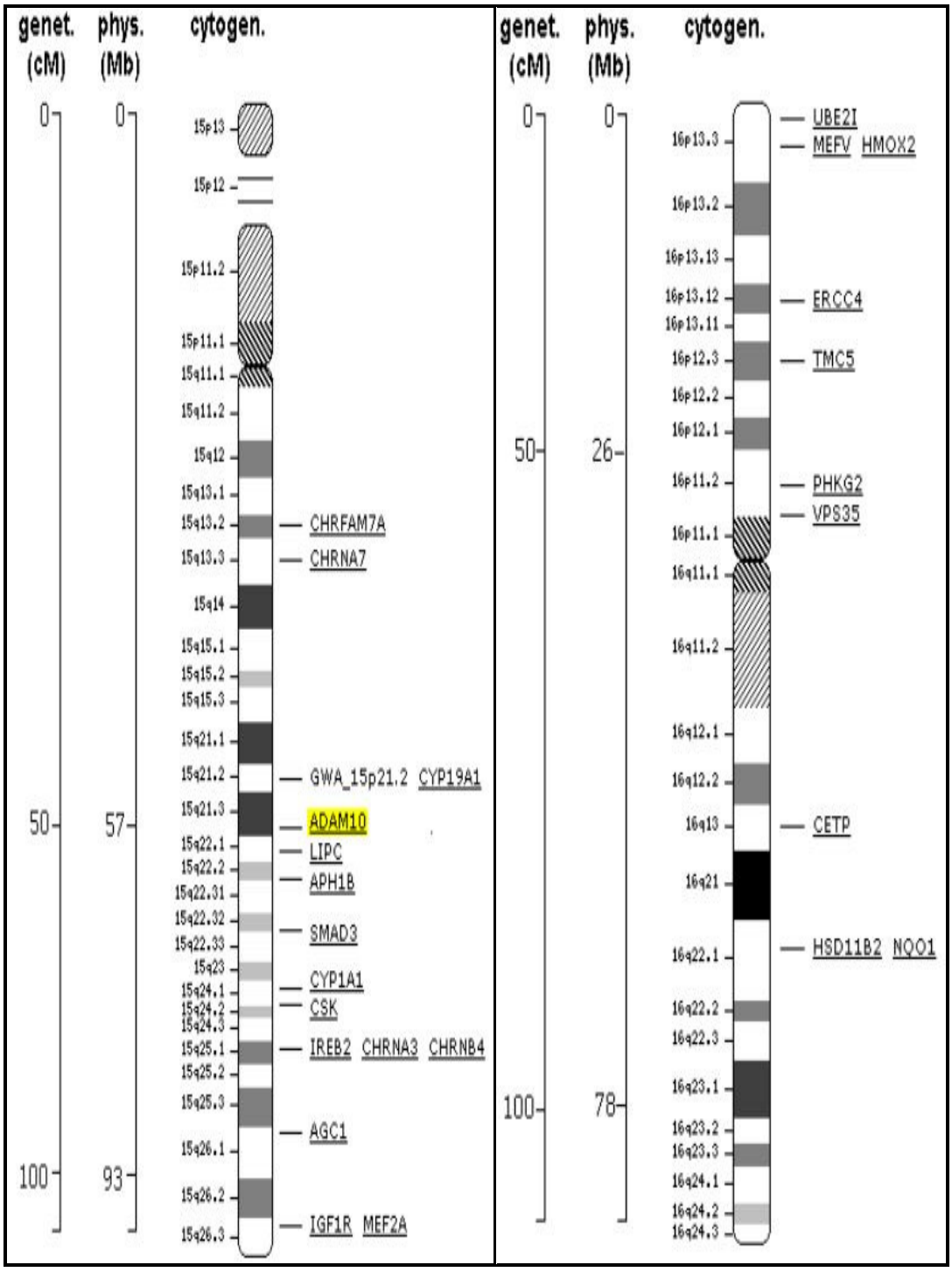

**Published AD candidate genes and locus (chromosomes 9 and 10)**



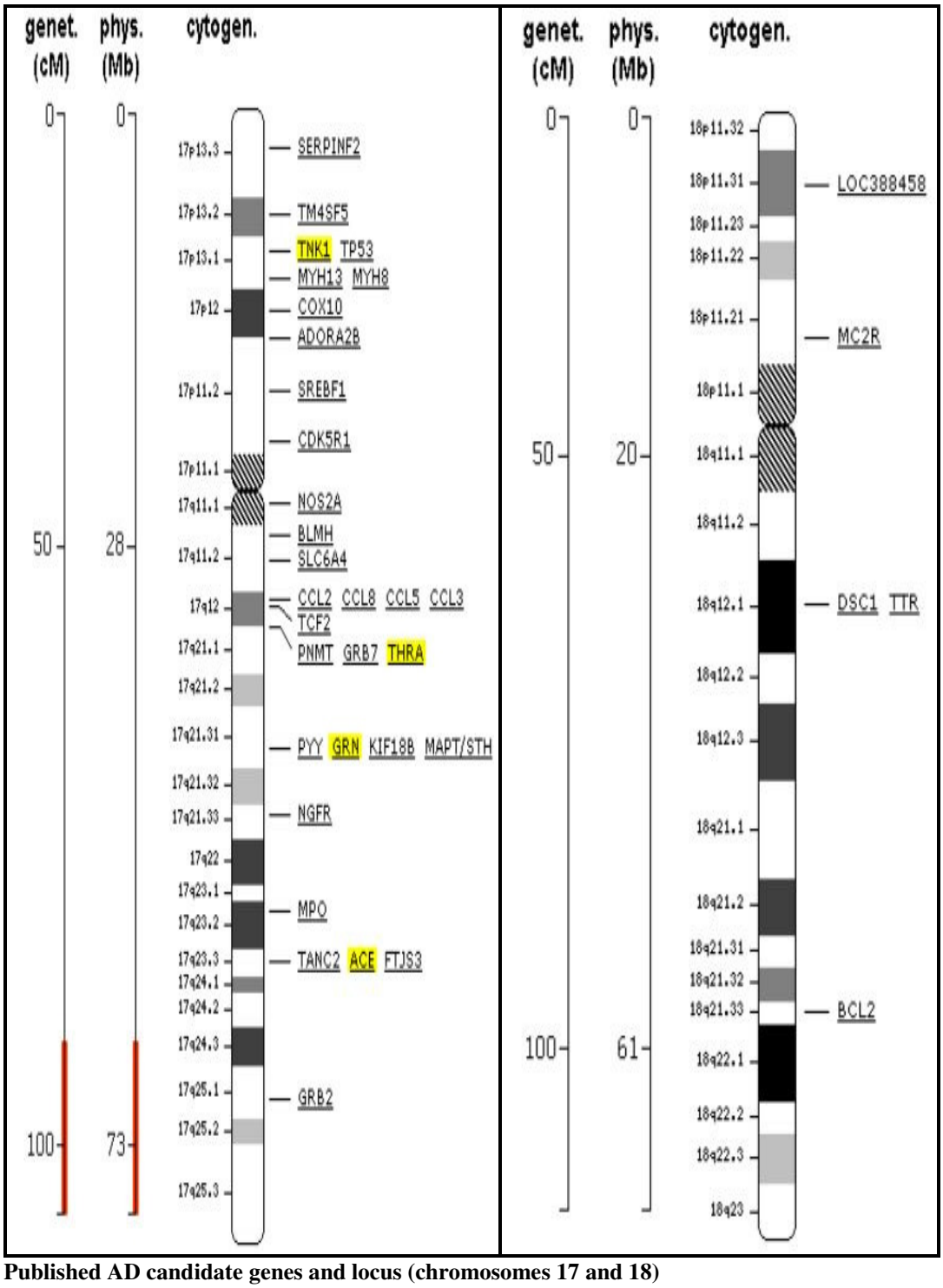

**Published AD candidate genes and locus (chromosomes 11 and 12)**



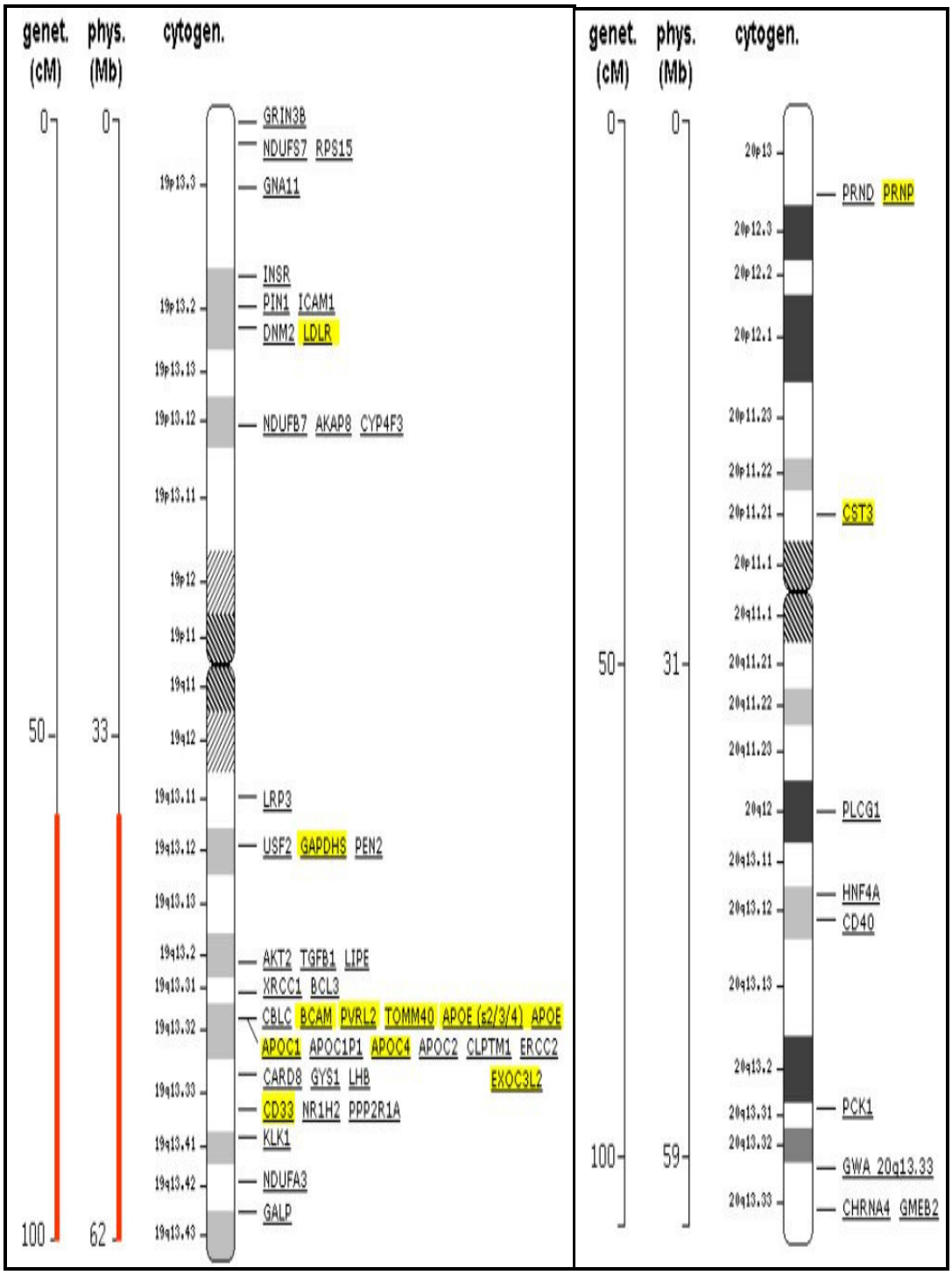

**Published AD candidate genes and locus (chromosomes 13 and 14)**



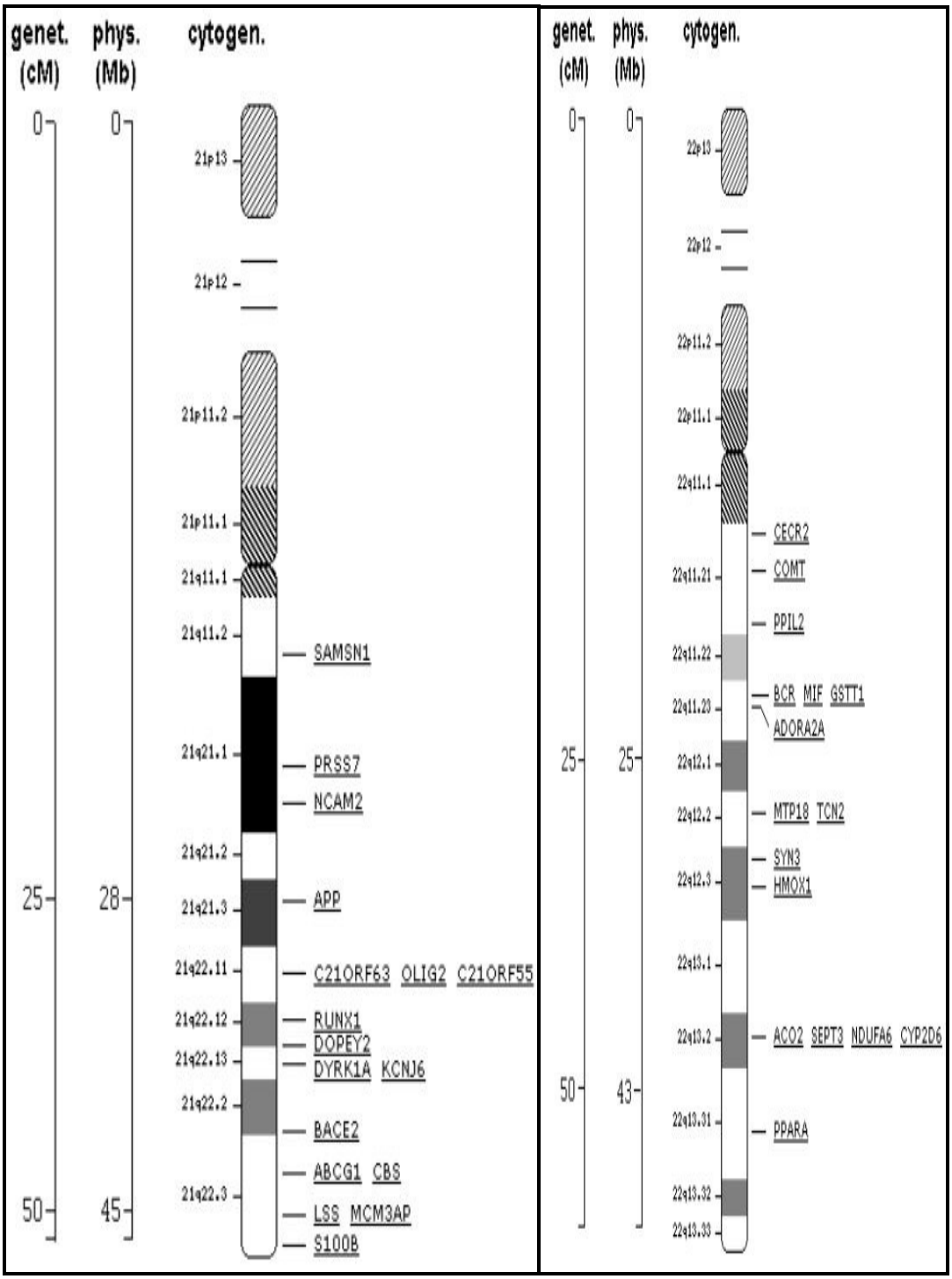

**Published AD candidate genes and locus (chromosomes 15 and 16)**



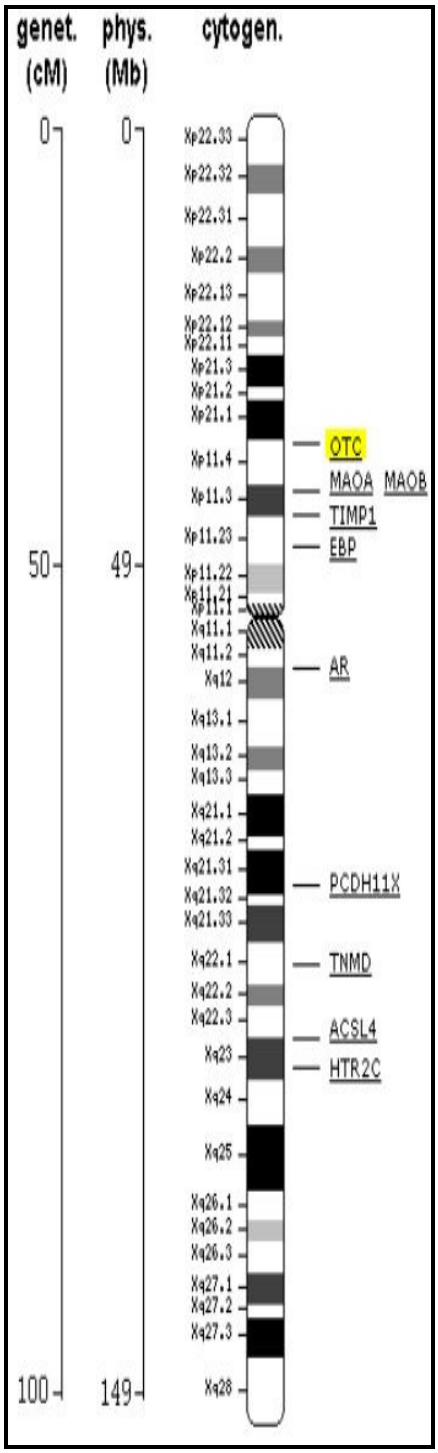

**Published AD candidate genes and locus (chromosomes 17 and 18)**




**Published AD candidate genes and locus (chromosomes 19 and 20)**




**Published AD candidate genes and locus (chromosomes 21 and 22)**




**Published AD candidate genes and locus (chromosome X)**

# APPENDIX C: TOP 100 GENES DEPENDING ON THE COMBINED P-VALUES AND THEIR OMIM ASSOCIATIONS FOR ADNI DATA

| Ranking | Gene | Combined p-value | OMIM association |
|---|---|---|---|
| 1 | SLC16A9 | ~0.0 | |
| 2 | RBFOX1 | ~0.0 | |
| 3 | CNTN5 | ~0.0 | Yes |
| 4 | C6orf10 | ~0.0 | |
| 5 | KDM4C | ~0.0 | Yes |
| 6 | RBMS3 | ~0.0 | Yes |
| 7 | SLC9A7 | ~0.0 | Yes |
| 8 | DSCAM | ~0.0 | Yes |
| 9 | FHIT | ~0.0 | Yes |
| 10 | LRP1B | ~0.0 | Yes |
| 11 | ANKRD44 | ~0.0 | |
| 12 | DACH1 | ~0.0 | Yes |
| 13 | SNX25 | ~0.0 | |
| 14 | FAM188B | ~0.0 | |
| 15 | CSMD1 | ~0.0 | Yes |
| 16 | HS3ST4 | ~0.0 | Yes |
| 17 | FER1L6 | ~0.0 | |
| 18 | NELL1 | ~0.0 | Yes |
| 19 | ALDH1A2 | ~0.0 | Yes |
| 20 | SUCLG2 | ~0.0 | Yes |
| 21 | LINGO2 | ~0.0 | Yes |
| 22 | KHDRBS2 | ~0.0 | Yes |
| 23 | ADAMTS12 | ~0.0 | Yes |
| 24 | MYO16 | 1.0559729726876105E-13 | |
| 25 | PSD3 | 1.1478770545894033E-13 | |
| 26 | NAALADL2 | 2.8086915382032135E-13 | Yes |
| 27 | KCNC2 | 5.097764790237947E-13 | Yes |
| 28 | SYT16 | 5.635802661215643E-13 | Yes |
| 29 | ODZ4 | 8.235055126059465E-13 | Yes |
| 30 | UBE2F | 9.52179225019045E-13 | |
| 31 | THSD7B | 1.220787352428633E-12 | |
| 32 | IL2RA | 1.7212225243400452E-12 | Yes |
| 33 | SLC9A11 | 2.295709207450806E-12 | |
| 34 | C6orf105 | 2.3931693941333888E-12 | |
| 35 | PTGDR | 2.5250913485194445E-12 | Yes |

| | | | |
|---|---|---|---|
| 36 | MDGA2 | 2.8231664353972757E-12 | Yes |
| 37 | CYP7B1 | 2.9844193338157977E-12 | Yes |
| 38 | WDR70 | 5.677298720896958E-12 | |
| 39 | DCC | 5.712138177223898E-12 | Yes |
| 40 | SPOCK3 | 7.429956512262358E-12 | Yes |
| 41 | UBASH3B | 9.016076010394048E-12 | Yes |
| 42 | CEP350 | 1.3695540446557348E-11 | |
| 43 | LRRC16A | 1.4549263666909511E-11 | Yes |
| 44 | MSI2 | 1.8786553402355918E-11 | Yes |
| 45 | HECW1 | 2.086217427143857E-11 | Yes |
| 46 | CDH13 | 2.1040073589353478E-11 | Yes |
| 47 | ARMCX4 | 2.271542742004075E-11 | |
| 48 | DLC1 | 5.016841702736779E-11 | Yes |
| 49 | PTGER3 | 5.4892373319763484E-11 | Yes |
| 50 | NKAIN3 | 6.147873394576069E-11 | Yes |
| 51 | ATRN | 7.307505066989754E-11 | Yes |
| 52 | KDM2B | 1.426071364037515E-10 | Yes |
| 53 | STXBP6 | 1.4839718820136236E-10 | Yes |
| 54 | ETV1 | 1.8723044307087347E-10 | Yes |
| 55 | HNRNPA1P4 | 1.892663136483239E-10 | |
| 56 | ZNF804B | 1.8961181441612422E-10 | |
| 57 | PDZRN4 | 3.8780411444008007E-10 | Yes |
| 58 | TRHDE | 4.933298594402748E-10 | Yes |
| 59 | BANK1 | 5.85007022672771E-10 | Yes |
| 60 | SDK1 | 6.401213279486818E-10 | Yes |
| 61 | CADM2 | 8.31451862930459E-10 | Yes |
| 62 | IL1RAP | 8.917987290688169E-10 | Yes |
| 63 | ADRA1A | 9.574293099790778E-10 | Yes |
| 64 | DLG2 | 1.4783508070275882E-9 | Yes |
| 65 | SMC6 | 1.4989003339035996E-9 | Yes |
| 66 | ZNF385D | 1.6180339125834562E-9 | |
| 67 | PLB1 | 1.6730178379192263E-9 | Yes |
| 68 | SLC44A4 | 1.7779500384792024E-9 | Yes |
| 69 | ABCA4 | 1.7805123088534963E-9 | Yes |
| 70 | PDZD2 | 1.852501676202066E-9 | Yes |
| 71 | SCLY | 2.0998204078561198E-9 | Yes |
| 72 | SMYD3 | 2.355594164391871E-9 | Yes |
| 73 | SLC8A1 | 2.46402357928276E-9 | Yes |
| 74 | IFT140 | 2.835825071094462E-9 | |
| 75 | MYO3B | 3.115662085606353E-9 | Yes |

| | | | |
|---|---|---|---|
| 76 | TNR | 3.5111797911669363E-9 | Yes |
| 77 | P2RY12 | 4.824727258543983E-9 | Yes |
| 78 | TMEM132D | 5.439923406189135E-9 | Yes |
| 79 | KIAA1328 | 5.85506142710832E-9 | |
| 80 | NSUN6 | 6.77429694563221E-9 | |
| 81 | ALK | 7.471823757149835E-9 | Yes |
| 82 | KAZN | 7.678629411879916E-9 | |
| 83 | CLEC16A | 8.362899841095292E-9 | Yes |
| 84 | LRPPRC | 8.631067530760875E-9 | Yes |
| 85 | NINL | 8.687370370675846E-9 | Yes |
| 86 | PDE4B | 9.752840005719736E-9 | Yes |
| 87 | SPAG16 | 1.008547610587095E-8 | Yes |
| 88 | ROCK2 | 1.0753607555369714E-8 | Yes |
| 89 | BACE1 | 1.0807974976229822E-8 | Yes |
| 90 | NT5E | 1.0918997077251672E-8 | Yes |
| 91 | BTBD9 | 1.1375441188542542E-8 | Yes |
| 92 | MCTP1 | 1.3326185274486867E-8 | |
| 93 | GEN1 | 1.4688014332860152E-8 | Yes |
| 94 | SEL1L2 | 1.5040966592610165E-8 | |
| 95 | AGAP1 | 1.5743466936395407E-8 | Yes |
| 96 | FRMD1 | 1.6442845925245657E-8 | |
| 97 | COG3 | 1.6865028111553102E-8 | Yes |
| 98 | GCOM1 | 1.7491279078454985E-8 | |
| 99 | C2CD3 | 1.9331230510838086E-8 | |
| 100 | FERMT2 | 2.0835665975214104E-8 | Yes |

# APPENDIX D: TOP 100 GENES DEPENDING ON THE COMBINED P-VALUES AND THEIR OMIM ASSOCIATIONS FOR GenADA DATA

| Ranking | Gene | Combined p-value | OMIM association |
|---|---|---|---|
| 1 | AGAP11 | 1,089181E-08 | |
| 2 | C4orf18 | 1,187423E-08 | |
| 3 | CPT2 | 2,232619E-08 | Yes |
| 4 | OR1L1 | 1,773873E-07 | |
| 5 | NACC2 | 1,974952E-07 | |
| 6 | MCPH1 | 2,484007E-07 | Yes |
| 7 | C9orf150 | 4,419189E-07 | |
| 8 | PRSS1 | 4,765822E-07 | Yes |
| 9 | ESR2 | 7,396964E-07 | Yes |
| 10 | ARHGEF19 | 8,743196E-07 | Yes |
| 11 | STC1 | 8,987053E-07 | Yes |
| 12 | SMURF1 | 1,272100E-06 | Yes |
| 13 | CPZ | 1,276079E-06 | Yes |
| 14 | MYPN | 1,451700E-06 | Yes |
| 15 | IFI30 | 1,649066E-06 | Yes |
| 16 | DAP | 1,937083E-06 | Yes |
| 17 | GNAZ | 1,997541E-06 | Yes |
| 18 | MRPS10 | 2,519949E-06 | Yes |
| 19 | PPT1 | 2,546921E-06 | Yes |
| 20 | HNRNPU | 2,780977E-06 | Yes |
| 21 | SEMA3E | 2,933871E-06 | Yes |
| 22 | EZH1 | 3,297509E-06 | Yes |
| 23 | SCCPDH | 3,529015E-06 | |
| 24 | MPDZ | 5,170320E-06 | Yes |
| 25 | CYP11A1 | 5,383462E-06 | Yes |
| 26 | CD46 | 6,608089E-06 | Yes |
| 27 | AEBP1 | 8,547281E-06 | Yes |
| 28 | SLC27A6 | 1,133377E-05 | Yes |
| 29 | PITHD1 | 1,276816E-05 | |
| 30 | OR51E2 | 1,395784E-05 | Yes |
| 31 | JAZF1 | 1,661085E-05 | Yes |
| 32 | BCO2 | 1,985933E-05 | Yes |
| 33 | FAM160B2 | 2,082886E-05 | |
| 34 | SUN1 | 2,532301E-05 | Yes |
| 35 | GAB2 | 2,658490E-05 | Yes |
| 36 | ACVR1B | 2,826451E-05 | Yes |

| | | | |
|---|---|---|---|
| 37 | MGC2752 | 2,837561E-05 | |
| 38 | DHPS | 2,885398E-05 | Yes |
| 39 | CD59 | 3,050824E-05 | Yes |
| 40 | NFATC4 | 3,272083E-05 | Yes |
| 41 | GPATCH4 | 3,284696E-05 | |
| 42 | SLC20A2 | 3,579173E-05 | Yes |
| 43 | EAPP | 3,699946E-05 | Yes |
| 44 | SC4MOL | 4,189537E-05 | Yes |
| 45 | DMPK | 4,263547E-05 | Yes |
| 46 | RPL9P25 | 4,506857E-05 | |
| 47 | NUP54 | 5,543072E-05 | Yes |
| 48 | WSB1 | 5,782227E-05 | Yes |
| 49 | LOC649930 | 6,378632E-05 | |
| 50 | LOC727744 | 6,505067E-05 | |
| 51 | C20orf111 | 7,225648E-05 | |
| 52 | TUBB2C | 7,510362E-05 | Yes |
| 53 | CCBL1 | 7,860632E-05 | Yes |
| 54 | FN3KRP | 8,217669E-05 | Yes |
| 55 | PPL | 8,296717E-05 | Yes |
| 56 | USP21 | 8,574191E-05 | Yes |
| 57 | SH3GL1 | 9,594797E-05 | Yes |
| 58 | VWF | 9,761645E-05 | Yes |
| 59 | CTSH | 1,176471E-04 | Yes |
| 60 | TRAK1 | 1,241100E-04 | Yes |
| 61 | DHCR7 | 1,274716E-04 | Yes |
| 62 | SDHA | 1,301780E-04 | Yes |
| 63 | SH3BP5 | 1,402876E-04 | Yes |
| 64 | PKN1 | 1,432841E-04 | Yes |
| 65 | C1QTNF3 | 1,582540E-04 | Yes |
| 66 | ATMIN | 1,593690E-04 | |
| 67 | SETD8 | 1,625813E-04 | Yes |
| 68 | LOC731709 | 1,657844E-04 | |
| 69 | LMCD1 | 1,720101E-04 | Yes |
| 70 | CTR9 | 1,803722E-04 | Yes |
| 71 | **CD36** | 1,825279E-04 | Yes |
| 72 | KCTD15 | 1,880624E-04 | Yes |
| 73 | ATP6V0E1 | 1,946420E-04 | Yes |
| 74 | CITED2 | 2,024721E-04 | Yes |
| 75 | EBNA1BP2 | 2,212395E-04 | |
| 76 | COPS7A | 2,360458E-04 | |

| | | | |
|---|---|---|---|
| 77 | KLHDC8B | 2,650786E-04 | Yes |
| 78 | DDOST | 2,698274E-04 | Yes |
| 79 | SDF2L1 | 2,718846E-04 | Yes |
| 80 | AKR1A1 | 2,735476E-04 | Yes |
| 81 | SMAD1 | 2,895042E-04 | Yes |
| 82 | LOC389458 | 3,163264E-04 | |
| 83 | CLDN4 | 3,167976E-04 | Yes |
| 84 | LUM | 3,220273E-04 | Yes |
| 85 | RAB7A | 3,354747E-04 | |
| 86 | WNT2 | 3,390472E-04 | Yes |
| 87 | CPPED1 | 3,414755E-04 | |
| 88 | PALLD | 3,462945E-04 | Yes |
| 89 | HEBP1 | 3,522131E-04 | Yes |
| 90 | TBCK | 3,539400E-04 | |
| 91 | CDK6 | 3,717210E-04 | Yes |
| 92 | LOC400499 | 3,746483E-04 | |
| 93 | ATP2A3 | 3,889549E-04 | Yes |
| 94 | DERL2 | 4,042802E-04 | Yes |
| 95 | LGMN | 4,044587E-04 | Yes |
| 96 | LOC644538 | 4,110027E-04 | |
| 97 | IGF2R | 4,125967E-04 | Yes |
| 98 | CHD1L | 4,397654E-04 | Yes |
| 99 | SEP15 | 4,493724E-04 | Yes |
| 100 | FAM19A4 | 4,669621E-04 | |

# APPENDIX E: COMPARISON OF TOP 100 SNPs AFTER AHP PRIORITIZATION VS SPOT PRIORITIZATION FOR ADNI DATA IN TERMS OF BIOLOGICAL RELEVANCE

| **METU-SNP** | | | | **SPOT** | | |
|---|---|---|---|---|---|---|
| **Ranking** | **SNP ID** | **AD linked gene** | | **Ranking** | **SNP ID** | **AD linked gene** |
| 1 | rs4651138 | | | 1 | rs4795895 | |
| 2 | rs2070045 | **SORL1** | | 2 | rs17365991 | |
| 3 | rs2230806 | **ABCA1** | | 3 | rs3795263 | |
| 4 | rs4652769 | | | 4 | rs4426564 | |
| 5 | rs3779870 | | | 5 | rs2075650 | |
| 6 | rs10808738 | | | 6 | rs12605132 | |
| 7 | rs4395923 | | | 7 | rs9268368 | |
| 8 | rs4936637 | **SORL1** | | 8 | rs10941091 | |
| 9 | rs6424883 | | | 9 | rs667782 | |
| 10 | rs10752893 | | | 10 | rs885691 | |
| 11 | rs9832203 | | | 11 | rs1233651 | |
| 12 | rs895286 | | | 12 | rs5442 | **GNB3** |
| 13 | rs4358067 | | | 13 | rs12489170 | |
| 14 | rs10857526 | | | 14 | rs6729218 | |
| 15 | rs603634 | | | 15 | rs13006848 | |
| 16 | rs10934675 | | | 16 | rs12457258 | |
| 17 | rs1800464 | | | 17 | rs6020624 | |
| 18 | rs13390226 | | | 18 | rs4935801 | |
| 19 | rs1799898 | **LDLR** | | 19 | rs3735080 | |
| 20 | rs688 | **LDLR** | | 20 | rs3862683 | |
| 21 | rs4234221 | | | 21 | rs1055207 | |
| 22 | rs8027035 | **CHRNA7** | | 22 | rs2548032 | |
| 23 | rs644511 | | | 23 | rs2235573 | |
| 24 | rs1355920 | **CHRNA7** | | 24 | rs756847 | |
| 25 | rs333309 | | | 25 | rs2228527 | |
| 26 | rs10932400 | | | 26 | rs10941100 | |
| 27 | rs12621088 | | | 27 | rs2302674 | |
| 28 | rs1917810 | | | 28 | rs1542604 | |
| 29 | rs589663 | | | 29 | rs1043261 | |
| 30 | rs13012677 | | | 30 | rs2199619 | |
| 31 | rs11892696 | | | 31 | rs1801591 | |
| 32 | rs12618478 | | | 32 | rs9935113 | |
| 33 | rs10932398 | | | 33 | rs4814111 | |
| 34 | rs1373928 | | | 34 | rs2286472 | |
| 35 | rs9813330 | | | 35 | rs2121473 | |
| 36 | rs4634050 | | | 36 | rs2326007 | |
| 37 | rs1373930 | | | 37 | rs2280511 | |
| 38 | rs1606659 | **CHRNA7** | | 38 | rs2484180 | |
| 39 | rs10997263 | | | 39 | rs6505403 | |
| 40 | rs7904053 | | | 40 | rs3849994 | |
| 41 | rs4746654 | | | 41 | rs11606296 | |

| | | | | | | |
|---|---|---|---|---|---|---|
| 42 | rs7074696 | | | 42 | rs4757268 | |
| 43 | rs7099157 | | | 43 | rs652888 | |
| 44 | rs6480128 | | | 44 | rs17488241 | |
| 45 | rs2619656 | | | 45 | rs821480 | |
| 46 | rs8178990 | | | 46 | rs2262425 | |
| 47 | rs20563 | | | 47 | rs2297781 | |
| 48 | rs2071421 | | | 48 | rs7443549 | |
| 49 | rs2298813 | **SORL1** | | 49 | rs6671527 | |
| 50 | rs2066718 | **ABCA1** | | 50 | rs2304977 | |
| 51 | rs4912868 | | | 51 | rs7911085 | |
| 52 | rs17416172 | | | 52 | rs1060242 | |
| 53 | rs3773892 | | | 53 | rs2969 | |
| 54 | rs7419259 | | | 54 | rs2257906 | |
| 55 | rs895061 | | | 55 | rs6138650 | |
| 56 | rs9840301 | | | 56 | rs9790 | |
| 57 | rs3772744 | | | 57 | rs4891524 | |
| 58 | rs1857796 | | | 58 | rs1461707 | |
| 59 | rs10497953 | | | 59 | rs972984 | |
| 60 | rs6706010 | | | 60 | rs536141 | |
| 61 | rs7626419 | | | 61 | rs2045191 | |
| 62 | rs13071977 | | | 62 | rs1044730 | |
| 63 | rs921001 | | | 63 | rs4851287 | |
| 64 | rs17259208 | | | 64 | rs17831682 | |
| 65 | rs1427281 | | | 65 | rs12125245 | |
| 66 | rs535801 | | | 66 | rs13360277 | |
| 67 | rs2043888 | | | 67 | rs1254929 | |
| 68 | rs6772915 | | | 68 | rs6713132 | |
| 69 | rs10207020 | | | 69 | rs473210 | |
| 70 | rs17347530 | | | 70 | rs1048101 | |
| 71 | rs1357139 | | | 71 | rs17055498 | |
| 72 | rs2204853 | | | 72 | rs4899065 | |
| 73 | rs16846100 | | | 73 | rs1073276 | |
| 74 | rs2777799 | **ABCA1** | | 74 | rs4862792 | |
| 75 | rs12695438 | | | 75 | rs10498817 | |
| 76 | rs1950091 | | | 76 | rs16900602 | |
| 77 | rs3793791 | | | 77 | rs2063979 | |
| 78 | rs3793792 | | | 78 | rs7615865 | |
| 79 | rs6757140 | | | 79 | rs2387976 | |
| 80 | rs12995889 | | | 80 | rs6076364 | |
| 81 | rs4073245 | | | 81 | rs11685766 | |
| 82 | rs6779362 | | | 82 | rs2305397 | |
| 83 | rs1699102 | **SORL1** | | 83 | rs676210 | |
| 84 | rs2289837 | | | 84 | rs3826007 | |
| 85 | rs8178992 | | | 85 | rs2235197 | |
| 86 | rs1800977 | **ABCA1** | | 86 | rs2537830 | |
| 87 | rs2289839 | | | 87 | rs11660401 | |
| 88 | rs2274873 | **ABCA1** | | 88 | rs815470 | |
| 89 | rs17561 | **IL1A** | | 89 | rs2273816 | |
| 90 | rs2280294 | | | 90 | rs681751 | |
| 91 | rs1986181 | | | 91 | rs2403088 | |
| 92 | rs9881879 | | | 92 | rs3811515 | |
| 93 | rs1010158 | **SORL1** | | 93 | rs1928565 | |

| | | | | | | |
|---|---|---|---|---|---|---|
| 94 | rs2830052 | **APP** | | 94 | rs8362 | |
| 95 | rs839519 | | | 95 | rs11071341 | |
| 96 | rs9843963 | | | 96 | rs5951332 | |
| 97 | rs4486246 | | | 97 | rs6471482 | |
| 98 | rs1620003 | **SORL1** | | 98 | rs2634974 | |
| 99 | rs2271446 | | | 99 | rs11948306 | |
| 100 | rs4327886 | | | 100 | rs35195 | |

# APPENDIX F: COMPARISON OF TOP 100 SNPs AFTER AHP PRIORITIZATION VS SPOT PRIORITIZATION FOR GenADA DATA IN TERMS OF BIOLOGICAL RELEVANCE

| **METU-SNP** | | | | **SPOT** | | |
|---|---|---|---|---|---|---|
| **Ranking** | **SNP ID** | **AD linked gene** | | **Ranking** | **SNP ID** | **AD linked gene** |
| 1 | rs7229 | | | 1 | rs1062683 | |
| 2 | rs14531 | | | 2 | rs3733472 | |
| 3 | rs4947 | | | 3 | rs10252253 | |
| 4 | rs2759 | **MPO** | | 4 | rs11166412 | |
| 5 | rs6304 | | | 5 | rs3812205 | |
| 6 | rs6324 | | | 6 | rs17074644 | |
| 7 | rs6323 | | | 7 | rs3739407 | |
| 8 | rs6305 | | | 8 | rs6746030 | |
| 9 | rs13136 | | | 9 | rs4687319 | |
| 10 | rs6314 | | | 10 | rs3742261 | |
| 11 | rs6308 | | | 11 | rs17593271 | |
| 12 | rs6313 | | | 12 | rs1065035 | |
| 13 | rs3010 | | | 13 | rs560659 | |
| 14 | rs8951 | | | 14 | rs16966703 | |
| 15 | rs1146 | **APP** | | 15 | rs4545143 | |
| 16 | rs761388 | | | 16 | rs304230 | |
| 17 | rs15966 | | | 17 | rs572846 | |
| 18 | rs669 | **A2M** | | 18 | rs8041254 | |
| 19 | rs12222 | | | 19 | rs6677080 | |
| 20 | rs3761 | | | 20 | rs17067596 | |
| 21 | rs216779 | **APP** | | 21 | rs4310078 | |
| 22 | rs191962 | **APP** | | 22 | rs1476970 | |
| 23 | rs57126 | **APP** | | 23 | rs12090877 | |
| 24 | rs214493 | **APP** | | 24 | rs4756055 | |
| 25 | rs928902 | **APP** | | 25 | rs14976 | |
| 26 | rs184200 | **APP** | | 26 | rs602668 | |
| 27 | rs15486 | | | 27 | rs8014810 | |
| 28 | rs6307 | | | 28 | rs1032471 | |
| 29 | rs8077 | | | 29 | rs17752628 | |
| 30 | rs3125 | | | 30 | rs7206841 | |
| 31 | rs7049 | | | 31 | rs352222 | |
| 32 | rs920812 | | | 32 | rs2978023 | |
| 33 | rs3991 | **APP** | | 33 | rs2150820 | |
| 34 | rs1145 | **APP** | | 34 | rs17804446 | |
| 35 | rs622337 | | | 35 | rs10132580 | |
| 36 | rs4175 | **APP** | | 36 | rs17793957 | |

| | | | | | | |
|---|---|---|---|---|---|---|
| 37 | rs621494 | | | 37 | rs12456284 | |
| 38 | rs469420 | **APP** | | 38 | rs17116710 | |
| 39 | rs13396 | | | 39 | rs936160 | |
| 40 | rs214494 | **APP** | | 40 | rs7518943 | |
| 41 | rs216767 | **APP** | | 41 | rs10031148 | |
| 42 | rs216758 | **APP** | | 42 | rs10132954 | |
| 43 | rs867442 | | | 43 | rs11810899 | |
| 44 | rs434844 | | | 44 | rs1060743 | |
| 45 | rs912127 | | | 45 | rs3825569 | |
| 46 | rs226407 | **A2M** | | 46 | rs429419 | **BIN1** |
| 47 | rs2477 | **A2M** | | 47 | rs10775471 | |
| 48 | rs226402 | **A2M** | | 48 | rs16893388 | |
| 49 | rs226403 | **A2M** | | 49 | rs11206955 | |
| 50 | rs169963 | **APP** | | 50 | rs3813487 | |
| 51 | rs979605 | | | 51 | rs3793511 | |
| 52 | rs448116 | | | 52 | rs4735627 | |
| 53 | rs226389 | **A2M** | | 53 | rs3802428 | |
| 54 | rs944047 | | | 54 | rs10142154 | |
| 55 | rs216769 | **APP** | | 55 | rs10024098 | |
| 56 | rs985933 | | | 56 | rs1886811 | |
| 57 | rs768040 | **APP** | | 57 | rs1201559 | |
| 58 | rs216765 | **APP** | | 58 | rs10736889 | |
| 59 | rs226384 | **A2M** | | 59 | rs10489622 | |
| 60 | rs6317 | | | 60 | rs16949276 | |
| 61 | rs6310 | | | 61 | rs3829799 | |
| 62 | rs6309 | | | 62 | rs7210713 | |
| 63 | rs13558 | | | 63 | rs6571727 | |
| 64 | rs14660 | | | 64 | rs839511 | |
| 65 | rs706149 | | | 65 | rs7614 | |
| 66 | rs6174 | | | 66 | rs2999061 | |
| 67 | rs5272 | **PTGS2** | | 67 | rs2955091 | |
| 68 | rs5273 | **PTGS2** | | 68 | rs52911 | |
| 69 | rs268 | **LPL** | | 69 | rs1808529 | |
| 70 | rs6265 | | | 70 | rs1653586 | |
| 71 | rs12505 | | | 71 | rs17377379 | |
| 72 | rs622397 | | | 72 | rs8026464 | |
| 73 | rs238740 | | | 73 | rs606114 | |
| 74 | rs7330 | | | 74 | rs3795685 | |
| 75 | rs10237 | | | 75 | rs6573270 | |
| 76 | rs16942 | | | 76 | rs714705 | |
| 77 | rs15997 | | | 77 | rs1451392 | |
| 78 | rs7720 | | | 78 | rs10084692 | |

| | | | | | | |
|---|---|---|---|---|---|---|
| 79 | rs12515 | | | 79 | rs2742424 | |
| 80 | rs14261 | | | 80 | rs12503735 | |
| 81 | rs16339 | | | 81 | rs9357738 | |
| 82 | rs180515 | | | 82 | rs12146894 | |
| 83 | rs385981 | | | 83 | rs2227127 | |
| 84 | rs551115 | | | 84 | rs2230742 | |
| 85 | rs5274 | **PTGS2** | | 85 | rs17732290 | |
| 86 | rs2953 | | | 86 | rs7825723 | |
| 87 | rs688 | **LDLR** | | 87 | rs4905897 | |
| 88 | rs5224 | | | 88 | rs12432214 | |
| 89 | rs4902 | | | 89 | rs2999081 | |
| 90 | rs36526 | | | 90 | rs6582406 | |
| 91 | rs36527 | | | 91 | rs2722278 | |
| 92 | rs16940 | | | 92 | rs2057116 | |
| 93 | rs5077 | **APOA1** | | 93 | rs4706990 | |
| 94 | rs4882 | **APOA1** | | 94 | rs2567982 | |
| 95 | rs20558 | | | 95 | rs13382811 | |
| 96 | rs4977 | **ACE** | | 96 | rs1045493 | |
| 97 | rs4314 | **ACE** | | 97 | rs2415306 | |
| 98 | rs4981 | **ACE** | | 98 | rs4789161 | |
| 99 | rs4976 | **ACE** | | 99 | rs2234971 | |
| 100 | rs5388 | | | 100 | rs10183045 | |