Genomic insights into the etiology of Alzheimer’s disease: a review

Christiane Reitz1–3

1Taub Institute for Research on Alzheimer’s Disease and the Aging Brain, 2Gertrude H Sergievsky Center, 3Department of Neurology, Columbia University, New York, NY, USA

Abstract: Over the past decade, studies capitalizing on high-throughput genome technologies have significantly advanced the knowledge on the genetic underpinnings of Alzheimer’s disease (AD) by identifying a wide set of pathophysiological mechanisms implicated in the disease in addition to amyloid precursor protein (APP) metabolism. These include: innate immune response and inflammation, lipid metabolism, endocytosis, cell migration, tau pathology, hippocampal synaptic function and axonal transport, regulation of gene expression and posttranslational modification of proteins, and microglial and myeloid cell function. The cumulative population attributable fraction associated with known genetic variants amounts now to ∼80%. High-throughput sequencing studies have started to map specific causative variants in these genes and have provided invaluable evidence for an involvement of rare variants in AD, overturning the “common disease–common variant” hypothesis. The ongoing and future large-scale translational studies and next generation whole genome or whole exome sequencing efforts hold the promise of mapping the specific causative variants in these genes; of identifying additional risk variants, including rare and structural variants; and of identifying novel targets for genetic testing, prevention, and treatment.

Keywords: genetics, gene, variation, polymorphism, genome-wide association study, sequencing

Introduction

Late-onset Alzheimer’s disease (LOAD) is the most frequent form of dementia in Western societies leading to a considerable health care burden. It is estimated that 17 million people worldwide have Alzheimer’s disease (AD). In the US alone, five million people have been affected, leading to a direct estimated health care cost of US$157–$215 billion dollars per year. The annual incidence rate of AD increases from 1% among people aged 65 years to approximately 8% for people aged 85 years and older. The duration of illness can be as long as 20 years, but the average is between 4 and 8 years. The aging of those born in the post-World War II era suggests that these numbers may triple by the year 2050, resulting in an increase of nearly 80% in total societal costs per adult.

AD typically begins with the onset of symptoms after age 60 and evolves slowly from mildly impaired memory function to severe cognitive loss, finally terminating inevitably in complete incapacity and death. Although in recent years there have been significant advances in biomarkers for the disease (plasma amyloid beta [Aβ], cerebrospinal fluid [CSF] Aβ and tau, and amyloid imaging) and prediction of cognitive decline, to date, there are no definitive diagnostic tests or biological markers of the
disease, and the diagnosis during life is based on a clinical examination. At death, the pathological manifestations in brain include deposits of extracellular β-amyloid protein (Aβ) in diffuse plaques and plaques containing elements of degenerating neurons (“neuritic plaques”). Intracellular changes include deposits of abnormally hyperphosphorylated tau protein, a microtubule assembly protein, in the form of neurofibrillary tangles. Activation of microglia and loss of neurons and synapses is also widespread.

Efforts to limit the impact of AD are being hindered by the lack of success of experimental drugs, which is attributable to incomplete characterization of the basic pathologic mechanisms. It is clear that there is a significant genetic component, with a heritability of 58% to 79%. Consequently, determining which genes and gene networks contribute to AD risk is expected to reveal major basic pathogenic mechanisms, highlighting key proteins and pathways for drug development (“druggable targets”); and to inform the development of genetic testing methods for identifying those at greatest risk for AD when preventive measures become available.

In recent years, the genetic analysis of LOAD has focused on the identification of common variants through genome-wide association studies (GWASs) and has identified several novel susceptibility genes, implicating specific pathways in the disease. This review article summarizes the current genomic insight into AD and provides suggestions for future research.

Data source and study selection

The primary sources of the studies addressed in this review article were abstracts and full-text articles published in English in the PubMed database between January 2008 and 2014. The key words used for searching PubMed were “Alzheimer’s disease”, “dementia”, “gene”, “genetics”, “genomics”, “epigenetics”, “endophenotype”, “genome-wide association study”, “rare variants”, “common variants”, and “sequencing”. The retrieved abstracts were read to identify studies addressing the topics included in this review. We also performed a manual search of references cited in published articles. The studies were read in their entirety to assess their appropriateness for inclusion in this article.

Genetic epidemiology of Alzheimer’s disease

Consistent with the role of a significant genetic contribution, the heritability of AD is 58% to 79%, and a positive family history is the strongest risk factor. Families that are multiply affected by AD are at increased risk for dementia, but the distribution of secondary cases is not consistent with Mendelian inheritance. AD is more frequent among monozygotic than dizygotic twins, and first-degree relatives of patients with AD have approximately twice the expected lifetime risk of developing the disease.

Studies of numerous large pedigrees with early-onset AD (onset age: 30–50 years) led to the discovery of autosomal dominant mutations in the amyloid precursor protein (APP), presenilin 1 (PSEN1), and presenilin 2 (PSEN2) genes. These studies suggested a common pathogenic mechanism involving enhanced generation and aggregation of Aβ. According to this “amyloid hypothesis”, β-secretase cleaves APP near the N terminus of the Aβ peptide; then, the membrane-bound C-terminal APP fragment is cleaved by γ-secretase, leading to accumulation of Aβ40 and Aβ42.

Role of rare and common variants in Alzheimer’s disease

For several decades, the main hypothesis assumed to underlie genetically complex diseases including AD was the “common disease–common variant” hypothesis. According to this hypothesis, the genetic factors underlying common diseases will be alleles that are themselves quite common in the population at large. In line with this notion, over the past several years, the most common strategy for finding novel AD gene candidates has been the GWAS. In GWASs, as many as several million genetic markers (single nucleotide polymorphisms [SNPs]) are tested for genetic association with disease risk and/or phenotypic endophenotypes, such as age of onset, biomarkers, imaging results, and neuropathological end points. These studies, however, have identified loci accounting for only part of the heritability of most complex diseases. Although some of this “missing heritability” may be ascribable to a large number of SNPs with weak effect, there is increasing evidence that there is a substantial contribution from rare variants with large effect that are not readily identifiable by SNP-based methods unless very large sample sizes (for example, >250,000 unselected individuals) are assessed. Thus, in recent work, resequencing of known risk loci and whole genome and whole exome sequencing has been pursued to reveal rare point mutations that may have an appreciable impact on disease risk or severity and explain part of this missing heritability.

Apolipoprotein (APO)E region

For more than a decade, only one genetic risk factor, the APOE-ε4 allele, located on chromosome 19q13, was...
an unequivocally established “susceptibility” gene in non-Hispanic Whites of European ancestry. APOE is a lipid-binding protein and is expressed in humans as three common isoforms coded for by three alleles, APOE-e2, -e3, and -e4. A single APOE-e4 allele is associated with a two- to threefold increased risk, and having two copies is associated with a fivefold or more increase.\(^\text{20}\) In addition, each inherited APOE-e4 allele lowers the age at onset by 6–7 years.\(^\text{21–28}\) APOE-e4 is also associated with lower cognitive performance, in particular the memory domain, and is associated with mild cognitive impairment\(^\text{29–32}\) and with progression from mild cognitive impairment to dementia.\(^\text{29–39}\) While the population attributable risk for APOE-e4 is estimated at 20%–50%,\(^\text{40}\) the presence of -e4 is neither necessary nor sufficient for developing the disease.\(^\text{41}\) In ethnic groups other than non-Hispanic Whites, the association between APOE and LOAD is largely inconsistent across studies.\(^\text{42,43}\)

**Findings from candidate gene and genome-wide association studies**

Table 1 lists the major GWAS studies conducted. Due to a paucity of data in other ethnic groups, most genetic association studies have restricted their efforts to non-Hispanic White populations. In addition, there are differences in linkage disequilibrium and allele frequencies between ethnic groups, which can lead to genetic background noise and the likelihood of false-positive findings due to confounding in combined analyses. Consequently, the first two sets of large-scale GWASs were performed in individuals of European ancestry and identified CLU, PICALM, CR1, BIN1, MS4A4A, ABCA7, CD2AP, CD33, and EPHA1 as AD susceptibility loci.\(^\text{44–47}\) CLU, also known as APOJ, is a lipoprotein highly expressed in both the periphery and the brain.\(^\text{48}\) Like APOE, it is involved in lipid transport.\(^\text{49}\) CLU is also hypothesized to act as an extracellular chaperone that influences Aβ aggregation and receptor-mediated Aβ clearance by endocytosis.\(^\text{50}\) Unlike APOE, there are no known coding variants that account for the observed genetic association to CLU, suggesting that genetic variation in expression levels may be responsible for the altered risk for LOAD.\(^\text{50}\) BIN1 (amphiphysin II) is a member of the Bin1/amphiphysin/RVS167 (BAR) family of genes, which are involved in diverse cellular processes, including actin dynamics, membrane trafficking, and clathrin-mediated endocytosis,\(^\text{51}\) affecting APP-processing and Aβ-production or Aβ-clearance from brain. PICALM is also involved in clathrin-mediated endocytosis and recruits clathrin and adaptor protein complex 2 to sites of vesicle assembly.\(^\text{52}\) Complement receptor 1 (CR1) is a cell surface receptor that is part of the complement system. It has binding sites for the complement factors C3b and C4b and is involved in clearing immune complexes containing these two proteins. Since Aβ oligomers can bind C3b, CR1 may participate in the clearance of Aβ. CR1 may also play a role in neuroinflammation, which is a prominent feature

<table>
<thead>
<tr>
<th>Study</th>
<th>Ethnic group</th>
<th>Sample size</th>
<th>Genes identified outside APOE region</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lambert et al (2009)(^\text{47})</td>
<td>Caucasian</td>
<td>Stage 1: 2,032 AD cases; 5,328 controls</td>
<td>CLU, CR1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Stage 2: 3,978 AD cases; 3,297 controls</td>
<td></td>
</tr>
<tr>
<td>Harold et al (2009)(^\text{44})</td>
<td>Caucasian</td>
<td>Stage 1: 3,941 AD cases; 7,848 controls</td>
<td>CLU, PICALM</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Stage 2: 2,023 AD cases; 2,340 controls</td>
<td></td>
</tr>
<tr>
<td>Seshadri et al (2010)(^\text{46})</td>
<td>Caucasian</td>
<td>Stage 1: 3,006 AD cases; 4,642 controls</td>
<td>BIN1, XOC3L2/BLOC1S3/MARK4, CLU, PICALM</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Stage 2: 2,032 AD cases; 5,328 controls</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Stage 3: 3,333 AD cases; 6,995 controls</td>
<td></td>
</tr>
<tr>
<td>Naj et al (2011)(^\text{45})</td>
<td>Caucasian</td>
<td>Stage 1: 8,309 AD cases; 7,366 controls</td>
<td>MS4A4A, CD2AP, CD33 and EPHA1, CR1,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Stage 2: 3,531 AD cases; 3,565 controls</td>
<td>BIN1, CLU, PICALM</td>
</tr>
<tr>
<td>Hollingworth et al (2011)(^\text{55})</td>
<td>Caucasian</td>
<td>Stage 1: 6,688 AD cases; 13,685 controls</td>
<td>ABCA7, MS4A6A/MS4A4E, EPHA1, CD33,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Stage 2: 4,896 AD cases; 4,903 controls</td>
<td>CD2AP</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Stage 3: 8,286 AD cases; 21,258 controls</td>
<td></td>
</tr>
<tr>
<td>Lambert et al (2013)(^\text{56})</td>
<td>Caucasian</td>
<td>Stage 1: 17,008 AD cases; 37,646 controls</td>
<td>CR1, CD33, BIN1, CD2AP, CLU, EPHA1,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Stage 2: 8,572 AD cases; 11,312 controls</td>
<td>PICALM, MS4, ABCA7, HLA-DRB5/HLA-DRB1,</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>PTK2B, SORL1, SL24A4/RIN3, DSG2</td>
</tr>
<tr>
<td>Lee et al (2011)(^\text{57})</td>
<td>Caribbean Hispanic</td>
<td>549 AD cases; 544 controls</td>
<td>CLU, PICALM, BIN1, CUGBP2, loci on</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2p25.1; 3q25.2; 7p21.1; 10q23.1</td>
</tr>
<tr>
<td>Reitz et al (2013)(^\text{58})</td>
<td>African American</td>
<td>1,968 AD cases; 3,928 controls</td>
<td>ABCA7, intergenic locus on 5q35.2</td>
</tr>
<tr>
<td>Miyasaka et al (2013)(^\text{53})</td>
<td>Japanese, Korean, Caucasian</td>
<td>14,072 AD cases; 14,061 controls</td>
<td>SORLI</td>
</tr>
</tbody>
</table>

**Abbreviations:** AD, Alzheimer’s disease; APOE, apolipoprotein E; GWASs, genome-wide association studies; LOAD, late-onset Alzheimer’s disease.
The second set of large GWASs identified additional susceptibility genes (CD33, MS4A4A/MS4A4E/MS4A6E cluster, ABCA7, CD2AP, and EPHA1). In line with the pathways identified by the first set of GWASs, all of these five loci are likely involved in the immune system, while ABCA7 is in addition involved in lipid metabolism and APP-processing (Table 2). The CD33 gene encodes a protein that is a member of a family of cell surface immune receptors that bind extracellular sialylated glycans and signal via a cytoplasmic domain called the immunoreceptor tyrosine inhibitory motif. CD33 has primarily been studied in the peripheral immune system, where it is expressed on myeloid progenitors and monocytes and also in the brain. In the periphery, CD33 appears to inhibit proliferation of myeloid cells. The MS4A4A/MS4A4E/MS4A6E locus is part of a cluster of 15 MS4A genes on chromosome 11 and encodes proteins with multiple membrane-spanning domains that were initially identified by their homology to CD20, a B-lymphocyte cell surface molecule. Little is known about the function of MS4A4A gene products; however, like CD33, MS4A4A is expressed on myeloid cells and monocytes and likely has an immune-related function. EPHA1 encodes a member of the ephrin family of cell surface receptors, which interact with ephrin ligands on adjacent cells to modulate cell adhesion, migration, and axon guidance and synapse formation and plasticity. While there is a substantial body of research on the function of ephrin receptors in general, little is known about the EPHA1 gene product. Like other ephrin receptors, it regulates cell morphology and motility, and early work implicated this receptor in regulating vascular morphogenesis and angiogenesis. Knocking out the EPHA1 gene in mice results in abnormal tail and reproductive tract development but no effects on the brain. Consistent with this notion, in mice, expression is restricted to epithelial tissue. In humans, EPHA1 is expressed by CD4-positive T lymphocytes, monocytes, intestinal epithelium, including colon epithelium. Combined with the lack of evidence for brain expression, this may suggest that, like CD33, CR1, and MS4A4/MS4A6E, the role of the EPHA1 gene product in AD may be mediated through the immune system. The CD2-associated protein gene (CD2AP) encodes a scaffolding protein that binds directly to actin, nephrin, and other proteins involved in cytoskeletal organization. In the immune system, CD2AP is required for synapse formation in a process that involves clathrin-dependent actin polymerization. ABCA7 is an integral transmembrane adenosine triphosphate (ATP)-binding cassette transporter belonging to the ABC family proteins that mediate the biogenesis of high-density lipoprotein with cellular lipid and helical apolipoproteins. It binds APOA-1 and functions in APO-mediated phospholipid and cholesterol efflux from cells. In addition, ABCA7 affects the transport of other important proteins (including Ab protein) through the cell membrane and is involved in host defense through effects on phagocytosis of apoptotic cells by macrophages.

The largest GWAS in persons of European ancestry to date was recently performed by the International Genomics of Alzheimer’s Project. This study, by Lambert et al., consisted of a large, two-stage meta-analysis of the aforementioned major GWASs of individuals of European ancestry and included in total 74,046 subjects. In addition to the APOE locus, 19 loci (CR1, BIN1, CD2AP, EPHA1, CLU, MS4A4A, PICALM, ABCA7, HLA-DRB5/HLA-DRB1, PTK2B, SORL1, SLC24A4/RIN3, INPP5D, MEF2C, NME8, ZCWPW1, CELF1, FERMT2, CASS4) reached genome-wide significance, defined as \( P < 5 \times 10^{-8} \), in the combined stage 1 and stage 2 analysis. Out of these, eleven (HLA-DRB5/HLA-DRB1, PTK2B, SORL1, SLC24A4/RIN3, INPP5D, MEF2C, NME8, ZCWPW1, CELF1, FERMT2, CASS4) were novel. Out of the 12 novel loci reaching genome-wide significance, several cluster in the specific pathways previously identified, ie, immune response (HLA-DRB5/DRB1, INPP5D, MEF2C), APP-processing (SORL1, CASS4), tau pathology (CASS4, FERMT2), cell migration (PTK2B), and lipid transport and endocytosis (SORL1). The SORL1 (sortilin-related receptor;
L[DLR class] 1) gene has previously been demonstrated to modulate trafficking and processing of APP, in a candidate gene approach.69,70 The results of this study further suggested the existence of additional pathways, such as hippocampal synaptic function (MEF2C, PTK2B), cytoskeletal function and axonal transport (CELF1, NME8, CASS4), regulation of gene expression and posttranslational modification of proteins, and microglial and myeloid cell function (INPP5D).

Consistent with the notion of a genetically complex disorder, the odds ratios (ORs) of all disease-associated loci other than the APOE-ε4 allele range from 1.08 to 1.29. Examining the genetic effect attributable to all the associated loci, the most strongly associated SNPs at each locus other than APOE show population attributable fractions (PAFs) or preventive fractions between 1.0%–8.0%.68 The cumulative PAF has been estimated at 89.4%.68 Finally, the results from the combined stage 1 and stage 2 data sets also identified 13 suggestive loci (an intergenic locus at chr1q31.2, HS3ST1, SQSTM1, TREML2, NDUFAF6, ECHDC3, AP2A2, ADAMTS20, IGH, SPPL2A, TRIP4, SCIMP, ACE) with association P-values <10−6.

In the largest GWAS performed to date in Caribbean Hispanics,71 associations in CLU, PICALM, and BIN1 were replicated, and several additional loci on 2p25.1, 3q25.2, 7p21.1 and 10q23.1 – which could be replicated in an independent cohort of non-Hispanic Whites of European ancestry from the National Institute on Aging Late-Onset Alzheimer’s Disease Family Study – were observed. Finally, in the largest GWAS of African Americans performed, Reitz et al72 identified ABCA7 as a major susceptibility locus in this ethnic group. Interestingly, in contrast to all GWAS loci identified in Caucasians, the ABCA7 locus had, in African Americans, an effect size as strong as that of APOE-ε4 (ie, a 70%–80% increase in risk compared with a 10%–20% increase in risk through the GWAS loci observed in Whites). This finding may represent a winner’s curse (ie, inflation of the estimated effect in a discovery set, in relation to follow-up studies) and needs to be confirmed by independent studies in African Americans and functional methods; however, this finding may have major implications for developing targets for genetic testing, prevention, and treatment in this ethnic group if proven true. In addition, this study confirmed APOE as a susceptibility gene in this ethnic group, which had been, prior to this study, inconsistently seen across studies. This study also replicated CR1, BIN1, EPHA1, and CD33. A large-scale GWAS study in Asian (Japanese and Korean) populations identified SORL1 as a susceptibility gene,73 and candidate gene studies performed in Asian populations identified MS4A6A, CD33, PICALM, CR1, CLU.74–79 The effect of the APOE-ε4 allele was universally observed in all GWASs performed across ethnic groups. (These findings suggest that there is, at least to some degree, overlap in causative genetic loci across different ethnic groups, although the causative variants within the disease-associated genes may differ.)

**Rare variants in late-onset Alzheimer’s disease**

Several loci containing rare, causative sequencing variants have been identified, some of which overlap with the aforementioned common variant loci identified by GWASs. In large multiplex pedigrees with LOAD, targeted exome sequencing of APP, PSEN1, PSEN2, APOE, GRN, and MAPT has identified 33 missense, nonsense, and splice site variants in 60 families (13.7%), 18 of which were novel.80 A recent study of whole-genome sequence data from 1,795 Icelanders identified a rare, coding mutation (A673T) in the APP gene protecting against LOAD and cognitive decline in the elderly without cognitive impairment.51 A673T is adjacent to the aspartyl protease β-site in APP and results in an approximately 40% reduction in the formation of Aβ in vitro. This strong protective effect of the A673T substitution against LOAD provides proof of principle for the hypothesis that reducing the β-cleavage of APP may protect against the disease. The TREM2/TREML2 locus associated with LOAD in a case-control GWAS had previously been reported by two independent multistage studies that combined sequencing analyses followed by a meta-analysis of independent imputed data sets (see below), with additional direct genotyping in other independent samples and functional analyses. These studies identified a rare exonic variant (R47H) associated with a three- to fourfold increased risk of developing AD.82,83 An exon sequencing study of PICALM in 48 AD cases and 48 controls identified an exonic splicing enhancer variation in exon 5 in linkage disequilibrium with GWAS SNP rs3851179.84 Other genes with rare variants for LOAD include SORL1,85 ADAM10,86 and PLD3.87 PLD3 is a poorly characterized member of the PLD superfamily of phospholipases but seems to affect APP-processing.80

**Discussion and significance**

In conclusion, over the past decade, studies capitalizing on high-throughput genome technologies have significantly advanced the knowledge on the genetic underpinnings of AD, implicating a wide set of pathophysiological mechanisms.
in addition to APP metabolism, including: innate immune response and inflammation, lipid metabolism, endocytosis, cell migration, tau pathology, hippocampal synaptic function and axonal transport, regulation of gene expression and posttranslational modification of proteins, and microglial and myeloid cell function. As described above, the cumulative PAF of the common GWAS loci has been estimated at 89.4%.\(^8\) However, it remains possible that this estimation is partly explained by the winner’s curse (ie, an overestimation of genetic effect size in initial studies). Consistent with this notion are 1) the fact that high-throughput sequencing studies have started to map additional causative rare variants in these genes, providing invaluable evidence for an involvement of rare variants in this complex disease and clearly overturning the common disease–common variant hypothesis; and 2) the increasing evidence for a significant role of noncoding ribonucleic acid (RNA) in complex disease.\(^8\)

While the human genome contains about 20,000 protein-encoding genes, 98.8% consists of non-protein-coding DNA sequence.\(^9\) The majority of these sequences are dynamically transcribed, mainly into non-protein-coding RNAs, with tens if not hundreds of thousands that show specific expression patterns and subcellular locations, as well as many classes of small regulatory RNAs. The emerging evidence indicates that these RNAs control the epigenetic states that underpin development and that many are dysregulated in cancer and other complex diseases, including AD.\(^8\) In addition, there is evidence that animals, particularly primates, have evolved plasticity in these RNA regulatory systems, especially in the brain. Thus, it appears that what was dismissed as “junk”, because it was not understood, may hold a significant part of the key to understanding human evolution, development, and cognition.

Overall, the specification of the pathways involved in AD etiology bears a strong potential for the development of therapeutically targets. The current basis for diagnostic biomarkers and therapies in clinical trials are the identified variants in APP, PSEN1, PSEN2, and APOE in AD. However, before this novel information can be used in clinical settings and safely considered for pharmaceutical intervention, ongoing and future large-scale next-generation sequencing approaches (both hypothesis-driven and hypothesis-free) as well as extensive functional studies are needed, to identify the specific causative variants and the specific factors with which they interact. In addition, a significant part of the missing heritability of AD, which is expected to be explained by several additional loci with small effect sizes each, needs to be identified. Such studies, including the Alzheimer’s Disease Sequencing Project, a major collaborative whole exome and whole genome sequencing effort by the National Institutes of Health (NIH), are now underway.

## Acknowledgments
Dr Reitz was supported by a Paul B Beeson Career Development Award (number K23AG034550).

## Disclosure
The author reports no conflicts of interest in this work.

## References


