Alzheimer's & Dementia 11 (2015) 658-671

Convergent genetic and expression data implicate immunity in Alzheimer's disease

International Genomics of Alzheimer's Disease Consortium (IGAP)[†]

Abstract

Background: Late-onset Alzheimer's disease (AD) is heritable with 20 genes showing genome-wide association in the International Genomics of Alzheimer's Project (IGAP). To identify the biology underlying the disease, we extended these genetic data in a pathway analysis.

Methods: The ALIGATOR and GSEA algorithms were used in the IGAP data to identify associated functional pathways and correlated gene expression networks in human brain.

Results: ALIGATOR identified an excess of curated biological pathways showing enrichment of association. Enriched areas of biology included the immune response ($P = 3.27 \times 10^{-12}$ after multiple testing correction for pathways), regulation of endocytosis ($P = 1.31 \times 10^{-11}$), cholesterol transport ($P = 2.96 \times 10^{-9}$), and proteasome-ubiquitin activity ($P = 1.34 \times 10^{-6}$). Correlated gene expression analysis identified four significant network modules, all related to the immune response (corrected $P = .002$–$.05$).

Conclusions: The immune response, regulation of endocytosis, cholesterol transport, and protein ubiquitination represent prime targets for AD therapeutics.

Keywords: Alzheimer's disease; Dementia; Neurodegeneration; Immune response; Endocytosis; Cholesterol metabolism; Ubiquitination; Pathway analysis; ALIGATOR; Weighted gene co-expression network analysis

1. Background

Alzheimer's disease (AD) affects more than five million Americans: one in eight at the age of >65 years and represents >60% of the six million dementia cases in Europe [1–3]. It is the commonest cause of dementia and imposes a large socioeconomic burden on individuals, their families, and society. Prevalence is estimated to treble by 2050; thus, understanding the mechanisms underlying this disease and developing treatments for it are essential. This study uses the largest genome-wide association study (GWAS) sample yet assembled for late-onset AD [4] and is the first to combine GWAS and expression data in a systematic search for the biological pathways underlying the genetic susceptibility to this disorder.

Much of our current understanding of the mechanisms that contribute to AD derives from the genetics of Mendelian forms of the disease: mutations in *APP*, *PSEN1*, and *PSEN2* cause early-onset forms of AD and underpin the amyloid cascade hypothesis [5]. Although amyloid deposition is diagnostic of AD, its etiologic contribution to the majority of common late-onset AD (LOAD) is unclear, and therapeutic strategies addressing the amyloid cascade hypothesis have been unsuccessful [6]. Therefore, other therapeutic avenues must be identified and targeted.

LOAD is genetically complex with 56% to 79% heritability [7]. In the Genetic and Environmental Risk in Alzheimer's Disease data set [8], approximately 20% of the total trait variance was accounted for by single-nucleotide polymorphisms (SNPs) on the GWAS chip outside the *APOE* region [9], with the ε4 allele of the *APOE* gene [10] accounting for a similar amount [9,11]. However, a substantial proportion of the genetic variance of late-onset AD is not accounted for by the 20 susceptibility genes currently identified [11]. The remaining genetic

[†]A list of members of the International Genomics of Alzheimer's Disease Consortium (IGAP) can be found in the Appendix at the end of this article.

*Corresponding authors: williamsj@cardiff.ac.uk and philippe.amouyel@pasteur-lille.fr

1552-5260/$ - see front matter © 2015 Published by Elsevier Inc. on behalf of The Alzheimer's Association.
http://dx.doi.org/10.1016/j.jalz.2014.05.1757

variance is likely to be due to both susceptibility genes of small effect that current sample sizes are insufficient to detect and rare variants, such as the coding variants in *TREM2* [12], that are poorly tagged by common variants in GWAS panels. In addition, individual genome-wide significant (GWS) genes identified in such studies may themselves not form good therapeutic targets, and the areas of biology that they highlight may only give a partial view of the potential therapeutic landscape. To gain the maximum useful information about causative pathways that may underpin LOAD and be prime targets for pharmaceutical intervention, we performed a robust pathway and integrated gene expression analysis using the largest available GWAS for AD [4].

2. Methods

2.1. Samples and genetic data

The sample comprised 17,008 AD cases and 37,646 control subjects in the primary GWAS analysis, with 8752 AD cases and 11,312 control subjects in the replication/extension sample and is described in detail elsewhere [4]. Only selected SNPs were genotyped in the replication/extension sample (see Online Methods).

2.2. Pathway analyses

We explored whether particular biological pathways were enriched for genetic associations [13,14] in the International Genomics of Alzheimer's Project (IGAP) data [4]. We used ALIGATOR [13,14] to test whether genes containing signals below the genome-wide significance threshold contribute to a pathway signal. ALIGATOR defines significant genes as having a best single-SNP P value less than a preset threshold. The resulting list of significant genes is compared with replicate gene sets generated by sampling SNPs randomly (thereby correcting for gene size). The method also controls for linkage disequilibrium (LD) between genes and multiple testing of nonindependent pathways (see Online Methods). Brown's method [15] was used to test pathway enrichment in the replication data. This method combines multiple SNPs together, explicitly correcting for both linkage disequilibrium (LD) between SNPs and the number of SNPs per gene (see Online Methods). Thus, correction for gene size was applied at both stages of the analysis. We interrogated the externally curated gene ontology and KEGG and MSigDB functional pathway collections (see Online Methods).

2.3. Expression correlation analyses

We used the expression data from Gibbs et al. [16] and performed weighted gene correlation network analysis (WGCNA) using the WGCNA package [17], separately on each tissue type to identify clusters of highly correlated genes called "modules." These modules were then tested for enrichment of genome-wide association signal in ALIGATOR.

3. Results

The sub-GWS variation in the IGAP data contains genetic signal, manifest by a significant excess of SNPs at all significance threshold up to $P = .05$ (Supplementary Table 1). This signal is unlikely to be due to uncorrected stratification because each of the individual Caucasian GWAS samples in the IGAP meta-analysis was corrected for ethnic variation using principal components [18].

We first identified a significant excess of biological pathways enriched for association signal in the IGAP data (Table 1 and Supplementary Table 2). Using the most significant 18,472 SNPs ($P < 8.32 \times 10^{-4}$) from IGAP [4], covering the top 5% of genes, 177 significantly enriched ($P < .01$) curated pathways were identified by ALIGATOR. To ensure that the excess of pathways was not an artifact of LD with genes of strong effect, we performed secondary enrichment analyses removing all genes that lay in the LD region of *APOE* or any of the GWS genes from the IGAP [4] study. A significant excess of enriched pathways remained (Table 1), showing that the pathways showed significant enrichment independent of the "known" AD genes. Likewise, a significant excess of enriched pathways was observed when the P-value criterion for defining significant SNPs and genes was varied (Supplementary Table 3).

Many of the 177 pathways with $P < .01$ in ALIGATOR are still significantly enriched after removing the *APOE*

Table 1
Significant excess of enriched pathways remain after removing *APOE* and the genome-wide significant genes

	Enrichment $P < .05$		Enrichment $P < .01$		Enrichment $P < .001$	
Genes removed (number of genes)	Number of pathways	P	Number of pathways	P	Number of pathways	P
None	542	<.0002	177	<.0002	40	<.0002
APOE + 1 Mb (77)	446	.0002	131	.0006	28	.0008
APOE + 1 Mb + GWS (98)	402	.0020	116	.0008	23	<.0002
APOE + 1 Mb + GWS+1 Mb (552)	336	.0094	93	.0066	22	.0018

Abbreviations: GWS, genome-wide significant; SNP, single-nucleotide polymorphism.

NOTE. Genes containing a SNP with $P < 8.32 \times 10^{-4}$ were counted as significant. This corresponds to the top 5% of genes (ranked by most significant SNP) when no genes are removed. The zero-kilobase window was used to assign SNPs to genes.