Alzheimer's & Dementia 12 (2016) 121-129

Featured Article

Interaction between variants in *CLU* and *MS4A4E* modulates Alzheimer's disease risk

Mark T. W. Ebbert[a], Kevin L. Boehme[a], Mark E. Wadsworth[a], Lyndsay A. Staley[a], for the Alzheimer's Disease Neuroimaging Initiative[1], Alzheimer's Disease Genetics Consortium, Shubhabrata Mukherjee[b], Paul K. Crane[b], Perry G. Ridge[a], John S. K. Kauwe[a,*]

[a]*Department of Biology, Brigham Young University, Provo, UT, USA*
[b]*Department of Medicine, University of Washington, Seattle, WA, USA*

Abstract

Introduction: Ebbert et al. reported gene-gene interactions between rs11136000-rs670139 (*CLU-MS4A4E*) and rs3865444-rs670139 (*CD33-MS4A4E*). We evaluate these interactions in the largest data set for an epistasis study.

Methods: We tested interactions using 3837 cases and 4145 controls from Alzheimer's Disease Genetics Consortium using meta-analyses and permutation analyses. We repeated meta-analyses stratified by apolipoprotein E (*APOE*) ε4 status, estimated combined odds ratio (OR) and population attributable fraction (cPAF), and explored causal variants.

Results: Results support the *CLU-MS4A4E* interaction and a dominant effect. An association between *CLU-MS4A4E* and *APOE* ε4 negative status exists. The estimated synergy factor, OR, and cPAF for rs11136000-rs670139 are 2.23, 2.45, and 8.0, respectively. We identified potential causal variants.

Discussion: We replicated the *CLU-MS4A4E* interaction in a large case-control series and observed *APOE* ε4 and possible dominant effect. The *CLU-MS4A4E* OR is higher than any Alzheimer's disease locus except *APOE* ε4, *APP*, and *TREM2*. We estimated an 8% decrease in Alzheimer's disease incidence without *CLU-MS4A4E* risk alleles and identified potential causal variants.

Keywords: Alzheimer's disease; Epistasis; MS4A4E; CLU; CD33; Meta-analysis; ADGC; ADNI

1. Introduction

Alzheimer's disease (AD) is a complex neurodegenerative disease and is the third leading cause of death in the United States [1]. AD is characterized by the accumulation of amyloid plaques and neurofibrillary tangles in the brain. Many genetic loci exist that modify AD risk, but collectively, they explain only a fraction of AD's heritability [2] and are not diagnostically useful [3,4]. Rare variants with large effects and epistatic interactions may account for much of the unexplained AD heritability, but are largely unknown due to limitations in traditional genome-wide association studies. Although rare variants 1and epistatic effects on AD are poorly understood, recent studies suggest that gene-gene interactions play a critical role in AD etiology and progression [3,5–7].

A previous study [3] reported evidence of two gene-gene interactions that increase AD risk. Specifically,

The authors report no conflicts of interest, financial or otherwise.

[1]Data used in preparation of this article were obtained in part from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

*Corresponding author. Tel.: +1-801-422-2993; Fax: +801-422-0900.
E-mail address: kauwe@byu.edu

http://dx.doi.org/10.1016/j.jalz.2015.08.163
1552-5260/

Ebbert et al. reported interactions between rs11136000 C/C (*CLU*; minor allele = T, MAF = 0.38) and rs670139 G/G (*MS4A4E*; minor allele = T, MAF = 0.38) genotypes (synergy factor [SF] = 3.81; P = .016) and the rs3865444 C/C (*CD33*; minor allele = A, MAF = 0.21) and rs670139 G/G (*MS4A4E*) genotypes (SF = 5.31; P = .003). All three variants have been implicated in numerous AD GWAS studies [8–13] and are on the "AlzGene Top Results" list [14], which summarizes the most established genes associated with AD.

MS4A4E and *CLU* were recently replicated in a large meta-analysis of 74,046 individuals, but *CD33* did not replicate [15]. Despite *CD33* failing to replicate, several studies demonstrated that *CD33* is involved in AD-related pathways and pathology, giving convincing evidence that *CD33* is somehow involved in AD. Three specific studies demonstrated that *CD33* alters monocyte function, amyloid uptake, and that *CD33* expression is associated with clinical dementia ratings [16–18]. rs3865444 is located in the 5′ untranslated regions (UTR) of *CD33*.

The association between *CLU* and AD status has been strongly established by both genetic and biological data. Recent studies demonstrated that rs11136000—an intronic single nucleotide polymorphism within *CLU*—is associated with AD-related pathology in healthy individuals including neural inefficiency [19] and decreased white matter integrity [20].

MS4A4E is a member of the membrane-spanning 4-domains subfamily A, but little else is known about the gene. However, rs670139—located in the *MS4A4E* 3′UTR according to gene model XM_011545416.1—is consistently associated with AD [15,18,21].

In this study, we attempted to replicate these gene-gene interactions using the largest data set used in an epistasis study, to date [22]. We performed an independent meta-analysis of data sets from the Alzheimer's Disease Genetics Consortium (ADGC) using 3837 cases and 4145 controls, followed by a combined meta-analysis that included the original Cache County results [3] with an additional 326 cases and 2093 controls. We also tested for dosage or dominant effects and an apolipoprotein E (*APOE*) ε4 effect. Finally, we explored possible causal variants using whole-genome sequence data from the Alzheimer's Disease Neuroimaging Initiative (ADNI).

2. Methods

2.1. Data description

We used SNP data from the ADGC, which consists of 32 studies collected over two phases and includes 16,000 cases and 17,000 controls. All subjects are self-reported as being of European-American ancestry. More information about this data set can be found in the study by Naj et al. [8] and the ADGC data preparation description [23].

Genotype data from 2419 individuals from the Cache County Study on Memory Health and Aging were also used in this study. The full cohort of 5092 individuals represented approximately 90% of the Cache County population aged ≥65 years when the study began in 1994 [24]. The Cache County data consist exclusively of individuals of European-American ancestry. Exactly 2673 individuals were excluded from the original Cache County analysis because of incomplete genotype or clinical data [3]. Additional information on this data set can be found in previous reports [3,24].

Whole-genome data from 747 individuals (223 controls, 195 cases, and 329 mild cognitive impairment [MCI]) were used in this article and were obtained from the ADNI database (adni.loni.usc.edu). ADNI is a large collaboration from several academic and private institutions, and subjects have been recruited from over 50 sites across the United States and Canada. Currently, over 1500 adults (ages 55–90) participate, consisting of cognitively normal older individuals, people with early or late MCI, and people with early stage AD. For up-to-date information, see www.adni-info.org.

2.2. SNP data preparation and statistical analysis

As gene-gene interactions are challenging to identify and replicate, we used the highest quality data possible. For each ADGC data set, we filtered SNPs imputed with low information (info <0.5) and converted the IMPUTE2/SNPTEST format files to PLINK format, using PLINK v1.90b2i [25,26]. We used the default PLINK uncertainty cutoff of 0.1, meaning any imputed call with uncertainty greater than 0.1 was treated as missing. We included SNPs with a missing genotype rate less than 0.05 and individuals with a missing rate less than 0.01. We then extracted the SNPs of interest: rs3865444 (CD33), rs670139 (*MS4A4E*), and rs11136000 (*CLU*) and tested Hardy-Weinberg equilibrium [27,28]. Using R version 3.1.1 [29], we excluded samples without complete data for all covariates including age, gender, case-control status, *APOE* ε4 dose, and the two SNPs being tested in the corresponding interaction. Entire data sets missing in the respective SNPs or covariates after data cleaning were excluded from further analysis. The requirement of complete data for both SNPs and all covariates is necessary for this analysis. Unfortunately, this requirement led to the exclusion of 23 and 24 entire data sets for the *CD33-MS4A4E* and *CLU-MS4A4E* interactions, respectively. We also excluded the ADC1 data set because it contained only one AD case, likely making it biased.

After data preparation, we tested the individual interactions in each data set using logistic regression. We defined the R models as "case_control ~ rs3865444 + rs670139 + rs3865444:rs670139 + apoe4dose + age + sex" and "case_control ~ rs11136000 + rs670139 + rs11136000:rs670139 + apoe4dose + age + sex" for the *CD33-MS4A4E* and *CLU-MS4A4E* interactions, respectively.