Neurobiology of Aging xxx (2017) 1.e1–1.e8

Contents lists available at ScienceDirect

Neurobiology of Aging

journal homepage: www.elsevier.com/locate/neuaging

Polygenic risk score in postmortem diagnosed sporadic early-onset Alzheimer's disease

Sultan Chaudhury [a], Tulsi Patel [a], Imelda S. Barber [a], Tamar Guetta-Baranes [a], Keeley J. Brookes [a], Sally Chappell [a], James Turton [a], Rita Guerreiro [b,c,d], Jose Bras [b,c,d], Dena Hernandez [e], Andrew Singleton [e], John Hardy [b,d], David Mann [f], ARUK Consortium, Kevin Morgan [a,*]

[a] Human Genetics Group, University of Nottingham, Nottingham, UK
[b] Department of Molecular Neuroscience, Institute of Neurology, University College London, London, UK
[c] Department of Medical Sciences, Institute of Biomedicine-iBiMED, University of Aveiro, Aveiro, Portugal
[d] UK Dementia Research Institute at UCL (UK DRI), London, UK
[e] Laboratory of Neurogenetics, National Institute of Aging, National Institute of Health, Bethesda, MD, USA
[f] Faculty of Medical and Human Sciences, Institute of Brain, Behaviour and Mental Health, University of Manchester, Manchester, UK

ARTICLE INFO

Article history:
Received 2 August 2017
Received in revised form 27 September 2017
Accepted 29 September 2017

Keywords:
Polygenic risk score (PRS)
Sporadic early-onset Alzheimer's disease (sEOAD)
Genotyping
NeuroX
NeuroChip

ABSTRACT

Sporadic early-onset Alzheimer's disease (sEOAD) exhibits the symptoms of late-onset Alzheimer's disease but lacks the familial aspect of the early-onset familial form. The genetics of Alzheimer's disease (AD) identifies *APOEε4* to be the greatest risk factor; however, it is a complex disease involving both environmental risk factors and multiple genetic loci. Polygenic risk scores (PRSs) accumulate the total risk of a phenotype in an individual based on variants present in their genome. We determined whether sEOAD cases had a higher PRS compared to controls. A cohort of sEOAD cases was genotyped on the NeuroX array, and PRSs were generated using PRSice. The target data set consisted of 408 sEOAD cases and 436 controls. The base data set was collated by the International Genomics of Alzheimer's Project consortium, with association data from 17,008 late-onset Alzheimer's disease cases and 37,154 controls, which can be used for identifying sEOAD cases due to having shared phenotype. PRSs were generated using all common single nucleotide polymorphisms between the base and target data set, PRS were also generated using only single nucleotide polymorphisms within a 500 kb region surrounding the *APOE* gene. Sex and number of *APOE* ε2 or ε4 alleles were used as variables for logistic regression and combined with PRS. The results show that PRS is higher on average in sEOAD cases than controls, although there is still overlap among the whole cohort. Predictive ability of identifying cases and controls using PRSice was calculated with 72.9% accuracy, greater than the *APOE* locus alone (65.2%). Predictive ability was further improved with logistic regression, identifying cases and controls with 75.5% accuracy.

1. Introduction

Alzheimer's disease (AD) is the most common form of dementia, characterized by the deterioration of memory, language, visuo-spatial skills, and behavior (Budson and Kowall, 2011). Dementia currently affects an estimated 46.8 million people globally (Prince et al., 2015). Hallmarks of AD were originally identified postmortem from histopathological signs of neuritic plaques, composed of amyloid-β, and neurofibrillary tangle formation; postmortem examination of brain tissue for these hallmarks remains the most definitive diagnosis of AD. Clinical diagnosis is accurately verified in more than 85% of cases (Naj and Schellenberg, 2016).

AD can be categorized based on the age of onset, where presentation of symptoms in individuals before the age of 65 years are classified as early-onset Alzheimer's disease (EOAD), whereas late-onset Alzheimer's disease (LOAD) classifies individuals with onset over 65 years (Barber et al., 2017; Wingo et al., 2012). LOAD has a heritability estimated to be around 70%, lower than estimates of heritability for EOAD, which vary between 80% and 100% (Barber et al., 2017; Wingo et al., 2012). An estimated 10% of EOAD cases have a familial aspect and are subsequently classified as early-onset familial AD (EOFAD). Autosomal dominant variants in the genes amyloid precursor protein (*APP*), presenilin 1 (*PSEN1*), and

* Corresponding author at: Human Genetics Group, School of Life Sciences, University of Nottingham, Nottingham NG7 2RD, UK. Tel.: +44 115 82 30724; fax: +44 115 970 9167.
E-mail address: kevin.morgan@nottingham.ac.uk (K. Morgan).

0197-4580/$ – see front matter © 2017 Elsevier Inc. All rights reserved.
https://doi.org/10.1016/j.neurobiolaging.2017.09.035

presenilin 2 (*PSEN2*) have been discovered to increase amyloid-β production, increasing the risk of EOFAD (Liu et al., 2013; Wingo et al., 2012). The remaining early-onset cases, classed as sporadic (sEOAD), are thought to be predominantly polygenic. The accumulation of variants which independently increase the risk of LOAD may lead to sEOAD at an earlier stage of life (Barber et al., 2017; Wingo et al., 2012).

The association of the *APOE* gene has been the most consistent observation in AD genetics with the presence of an *APOE* ε4 allele significantly more common among individuals diagnosed with AD, whereas the ε2 allele is considered protective (Liu et al., 2013; Naj and Schellenberg, 2016). Through genome-wide association studies (GWASs), around 20 genetic loci had been discovered, which affect risk of LOAD (Lambert et al., 2013). Follow-up studies based on the GWAS have identified other potential candidate AD risk genes not previously identified, including the *TRIP4*, *SPPL2A* (Ruiz et al., 2014), and *ABI3* genes (Sims et al., 2017). Next-generation sequencing has also enabled the identification of rare variants, one of the most consistent being the R47H variant in the *TREM2* gene locus (Guerreiro et al., 2013; Jonsson et al., 2013) which affect the risk of AD previously not identified in GWAS (Giri et al., 2016; Lambert et al., 2013; Naj and Schellenberg, 2016). Although most studies utilize Caucasian populations, further risk variants have been identified through next-generation sequencing in African-American individuals within the gene *AKAP9* (Giri et al., 2016; Logue et al., 2014). Conversely, protective variants have also been identified including a small coding deletion (rs10553596) within the *CASP7* gene associated with reduced incidence of AD among individuals with the *APOE* ε4ε4 genotype in 4 independent imputed data sets (Ayers et al., 2016). Further protective rare variants have also been identified by imputation of previous data sets such as the *PLCG2* gene (Sims et al., 2017).

Following on from these studies, Marden and colleagues (Marden et al., 2014, 2016) sought to determine if a summative analysis of GWAS variants would be able to predict a dementia probability score. An AD genetic risk score was calculated by multiplying each individual GWAS allele effect size using the beta coefficients obtained from a previous data set. This type of analysis demonstrated that AD genetic risk score could predict LOAD phenotype (Chouraki et al., 2016; Desikan et al., 2017; Sleegers et al., 2015; Verhaaren et al., 2013; Xiao et al., 2015; Yokoyama et al., 2015), mild cognitive impairment conversion to LOAD (Adams et al., 2015; Rodriguez-Rodriguez et al., 2013), hippocampal cortical thickness (Harrison et al., 2016; Sabuncu et al., 2012), hippocampal volume (Lupton et al., 2016), cerebrospinal fluid biomarkers (Martiskainen et al., 2015), and plasma inflammatory biomarkers (Morgan et al., 2017). This approach has been expanded to include further polymorphisms of smaller but important effect sizes to develop a polygenic risk score (PRS) (Euesden et al., 2015). This is an improvement on previous tests as they do not perform well when nonassociated single nucleotide polymorphisms (SNPs) are included (Basu et al., 2011; Chapman and Whittaker, 2008) and is considered to find SNPs of disease relevance that have too small an effect size to be identified conventionally (Pan et al., 2015).

In a recent study, polygenic scores were calculated for a cohort of LOAD cases and controls: the study used genotype information of the cohort to identify common variants that affect the risk of developing AD and used polygenic scores to form a risk prediction model (Escott-Price et al., 2015). By producing a model which identifies individuals with a high PRS, the potential for early screening, diagnosis, and determination of disease severity becomes possible (Euesden et al., 2015).

In this study, we have used genotype information generated on the NeuroX chip to generate a PRS in sEOAD. The NeuroX is a customized genotyping array built on the foundation of the Infinium HumanExome BeadChip v1.1, with additional custom content (Illumina, 2012). The array is designed to collect genotype information at markers across the entire genome. The HumanExome BeadChip foundation is made up of 242,901 markers, identifying variants in a series of metabolic, cancerous, diabetic, and psychiatric disorders (Barber et al., 2017; Nalls et al., 2015). The custom content includes 24,706 markers from candidate loci associated with neurological diseases such as AD, frontotemporal dementia, Parkinson's disease, multiple system atrophy, amyotrophic lateral sclerosis, myasthenia gravis, Charcot-Marie-Tooth, and progressive supranuclear palsy (Nalls et al., 2015).

To calculate a PRS, we have used the software package, PRSice, which utilizes genotype information from individuals in a target data set based on the effect scores of SNPs from a second data set, termed the base data set. The program uses R to define parameters and PLINK for the computational analysis (Purcell et al., 2007; R Core Team, 2013). PRSice is a command line program that allows specific parameters to be considered when generating PRS. The output files of the analysis include a list of individuals' scores at the best-fit threshold for predicting disease risk and a list of each tested threshold with its corresponding Nagelkerke's R^2 value, quantifying the level of predictability using that threshold (Eusden et al., 2015).

Linkage disequilibrium (LD) is a common problem when SNPs are scored based on their weighted effect and frequency when comparing cases and controls of a disease. The alleles of 2 SNPs present on the same chromosome can be commonly inherited together, and the recurrence of particular alleles at loci is an indicator of the degree of LD between SNPs (Bush and Moore, 2012). Given 2 SNPs in tight LD, both could be perceived as contributing to the disease risk in a functional haplotype; however, it may be that only 1 polymorphism is responsible for the phenotypic effect.

The aim of this study was to genotype sEOAD cases and controls to generate a PRS based on the genotype information of SNPs identified, and then using the estimated cumulative effect size the SNPs have on disease risk, to determine the predictability of the PRS at predicting cases versus controls.

2. Methods

2.1. Samples

The cohort genotyped consisted of 451 sEOAD cases (48.6% female) and 528 controls (51.3% female). sEOAD cases were screened for known disease causing variants within exons 16 and 17 of *APP* as well as variants in genes *PSEN1* and *PSEN2* to minimize inclusion of EOFAD cases. The diseased individuals had a documented or predicted age of onset of ≤65 years. Diagnosis of definite or probable sEOAD had met guidelines set by the National Institute of Neurological and Communicative Disorders and Stroke, the Alzheimer's disease and Related Disorders Association, and the Consortium to Establish a Registry for Alzheimer's disease. APOE ε status was determined for all individuals. At least 1 *APOE* ε4 allele was present in 57.6% of cases, with 22.3% of which being homozygotes (n = 58); 22.7% of controls harbored at least 1 ε4 allele, 9 control samples were ε4 homozygotes. These samples are described in greater detail in the article by Barber et al. (2017). Full details of the samples used in this study are outlined in Table 1. Experimental procedures were completed with informed consent, with approval from local ethics committee (Nottingham Research Ethics Committee 2 (REC reference 04/Q2404/130) and completed in accordance with approved guidelines. A standard phenol chloroform DNA extraction method was used on 2 mL of blood or 100 mg of brain tissue. DNA quality was assessed using gel electrophoresis, and quantity was determined by NanoDrop 3300 spectrometry (Barber et al., 2017).