Neurobiology of Aging 36 (2015) 1221.e7–1221.e13

Contents lists available at ScienceDirect

Neurobiology of Aging

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

A second-generation Irish genome-wide association study for amyotrophic lateral sclerosis

Russell L. McLaughlin [a,*], Kevin P. Kenna [a], Alice Vajda [b], Peter Bede [b], Marwa Elamin [b], Simon Cronin [c,d], Colette G. Donaghy [e], Daniel G. Bradley [a], Orla Hardiman [b]

[a] Population Genetics Laboratory, Smurfit Institute of Genetics, Trinity College Dublin, College Green, Dublin, Republic of Ireland
[b] Academic Unit of Neurology, Trinity Biomedical Sciences Institute, Trinity College Dublin, Dublin, Republic of Ireland
[c] Beaumont Hospital, Dublin, Republic of Ireland
[d] Cork University Hospital, Cork, Republic of Ireland
[e] Department of Neurology, Royal Victoria Hospital, Belfast, Northern Ireland

ARTICLE INFO

Article history:
Received 15 May 2014
Received in revised form 22 August 2014
Accepted 28 August 2014
Available online 6 September 2014

Keywords:
Amyotrophic lateral sclerosis
Genome-wide association study
Imputation
Irish population

ABSTRACT

Amyotrophic lateral sclerosis (ALS) is a heritable neurological disease for which the underlying genetic etiology is only partially understood. In Ireland, 83%–90% of cases are currently unexplained. Through large international collaborations, genome-wide association studies (GWASs) have succeeded in identifying a number of genomic loci that contribute toward ALS risk and age at onset. However, for the large proportion of risk that remains unexplained, population specificity of pathogenic variants could interfere with the detection of disease-associated loci. Single-population studies are therefore an important complement to larger international collaborations. In this study, we conduct a GWAS for ALS risk and age at onset in a large Irish ALS case-control cohort, using genome-wide imputation to increase marker density. Despite being adequately powered to detect associations of modest effect size, the study did not identify any locus associated with ALS risk or age at onset above the genome-wide significance threshold. Several speculative associations were, however, identified at loci that have been previously implicated in ALS. The lack of any clear association supports the conclusion that ALS is likely to be caused by multiple rare genetic risk factors. The findings of the present study highlight the importance of ongoing genetic research into the cause of ALS and its likely future challenges.

1. Introduction

Amyotrophic lateral sclerosis (ALS) is an incurable disease of the upper and lower motor neurones characterized by muscle wasting, paralysis, and death from respiratory failure within 3–5 years of symptom onset. In ~15% of cases, there is a discernible family history of the disease (Byrne et al., 2013); the remainder are termed sporadic ALS. However, with heritability estimated at 0.61 (0.38–0.78, Al-Chalabi et al., 2010) and no well-established environmental contribution, it is clear that genetic factors play a role in the etiology of the disease. Although several genes have been implicated in ALS, only 10%–17% of cases in Ireland can be explained by known genetic factors (Kenna et al., 2013), the most common of which is a hexanucleotide repeat expansion in *C9orf72*, accounting for 9% of cases (Byrne et al., 2012).

A number of genome-wide association studies (GWASs) have contributed to a deeper understanding of the genetic basis for ALS etiology (ALSGEN Consortium et al., 2013; Fogh et al., 2014; Shatunov et al., 2010; van Es et al., 2009). In European populations, the most robust results have stemmed from collaborative efforts that pool data from several countries to increase statistical power; such studies have led to the confirmation of chromosome 9p21 (*C9orf72*) as a major contributor to ALS pathogenesis (Shatunov et al., 2010), along with loci on chromosomes 19p13 (van Es et al., 2009), 1p34 (ALSGEN Consortium et al., 2013), and 17q11 (Fogh et al., 2014). Although these studies benefited from large sample sizes to increase power to detect loci harboring single pathogenic variants of reasonable effect size, the methodology of combining data from a number of different populations potentially precluded the detection of disease loci demonstrating

DGB and OH of this manuscript contributed in equal part to the study and wish to be considered as joint authors.

* Corresponding author at: Trinity College Dublin, Population Genetics Laboratory, Smurfit Institute of Genetics, College Green, Dublin 2, Ireland. Tel.: +353 1896 1265; fax: +353 1679 8558.
E-mail address: mclaugr@tcd.ie (R.L. McLaughlin).

0197-4580/$ – see front matter © 2015 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.neurobiolaging.2014.08.030

population-specific allelic heterogeneity. Extensive population specificity has been observed in the frequency profiles of established ALS-causing variants (Kenna et al., 2013), and it is possible that undiscovered ALS genes have eluded detection in multi-population GWASs by harboring multiple disease-causing variants on alternately tagged haplotypes.

It is therefore valid to complement large international GWASs with single-population studies. Recently, Deng et al. (2013) identified 2 novel ALS-associated loci (chromosomes 1q32 and 22p11) in a large GWAS for ALS in the Han Chinese population, exemplifying the continued utility of large single-population GWASs for ALS in parallel with the efforts of international consortia. In the present study, we conduct a GWAS in an Irish ALS case-control cohort of 605 cases negative for any established ALS-causing mutation and 1,179 controls. We improve power to detect disease-associated loci by increasing marker density through genome-wide imputation, which also addresses poor intersection of marker sets between genotyping platforms and results in our assessment of an order of magnitude of more markers than a typical intersection-based international GWAS. Our approach eliminates the effects of platform stratification observed after imputation, which can confound results when multiple genotyping platforms are used.

2. Methods

2.1. Participants

All patients included in the study had clinically definite, probable, or possible ALS (Brooks, 1994), diagnosed by specialist neurologists in the Republic of Ireland and Northern Ireland. All subjects provided informed written consent to participate in the study. Mean age (±standard deviation) of ALS onset was 61.3 ± 12.5 years; 27.4% of patients had bulbar-onset ALS and 43.8% were female. Patients were screened for established ALS-causing mutations in a wide panel of genes (Byrne et al., 2012; Kenna et al., 2013; McLaughlin et al., 2014). Controls were neurologically normal at time of sampling, with a mean age of 58.7 ± 13.6 years. An additional cohort of 866 healthy controls was provided by the Trinity Biobank (mean age 36.1 ± 12.4 years).

2.2. Genotype datasets

Genome-wide single-nucleotide polymorphism (SNP) data from 4 origins were used in the study. Data for 221 cases and 216 controls generated using the Illumina HumanHap550v3.0 BeadChip were obtained from the 2008 Irish ALS GWAS of Cronin et al. (2008) and were designated as ALS1 (dbGaP study accession number phs000127.v1.p1). These were supplemented with 131 cases and 139 controls genotyped using the Illumina Human610-Quadv1.0 BeadChip (ALS2); these data have been used in a number of previous studies (ALSGEN Consortium et al., 2013; Blauw et al., 2010). Three hundred twenty-three further cases were genotyped using the Illumina HumanOmniExpressExome-8v1 BeadChip (ALS3), and a fourth dataset of 866 Irish controls genotyped on the Affymetrix Genome-Wide Human SNP Array 6.0 was provided by the Trinity Biobank as part of a larger dataset generated by the International Schizophrenia Consortium (ISC). Before quality control (QC) and filtering, the union of all data represented genotypes for 1,726,772 SNPs in 1,901 individuals, with 89,905 SNPs intersecting all 4 datasets. Twelve samples were replicated across platforms to assess genotyping and imputation accuracy.

Before imputation, each dataset was remapped to GRCh37 genome coordinates for SNPs with dbSNP rs identifiers. A-T and C-G SNPs were discarded, and alleles for remaining SNPs were reassigned to match the forward strand of the GRCh37 reference genome. Using PLINK version 1.07 (Purcell et al., 2007), datasets were then subjected to a number of QC measures, including per-SNP and per-individual checks for low genotyping rate, systematic missingness, sex mismatches, Hardy-Weinberg disequilibrium, and extreme heterozygosity (Supplementary Materials 1). Data were then split per chromosome for imputation. During imputation and QC, datasets were frequently assessed for evidence of bias with quantile-quantile plots for association statistics, examination of intermarker distances, and principal component analysis using the smartpca program implemented within EIGENSOFT (Patterson et al., 2006).

2.3. Imputation

Genotype datasets were separately prephased (Howie et al., 2012) using SHAPE-IT v2.r644 (Delaneau et al., 2008) and then imputed using IMPUTE v2.3.0 (Howie et al., 2009) at 30,072,760 autosomal variant sites (SNPs and indels) with a minor allele count >1 in the phase 1 integrated release of the 1,000 Genomes Project (1000 Genomes Project Consortium et al., 2012). Preliminary analysis suggested that stratification of cases and controls between genotyping platforms affected association statistics after imputation; to address this, a second round of extensive QC and imputation on merged data was carried out as follows (detailed in Supplementary Materials 1). Taking frequent (minor allele frequency >1%) variant genotypes that either were present in the pre-imputed data or had been imputed in >98% of individuals at a probability >90%, Fisher exact tests were used to assess the association of genotype counts with platform of origin for all possible case-case or control-control pairs of datasets (e.g., ALS1 controls vs. ALS2 controls). Any variant that was associated ($p < 0.01$) with platform of origin was removed, and 1,836,219 variants with low evidence of bias and high genotyping rate remained after imputation and QC. To recover additional variant density without stratification, these genotypes were then merged across all the 4 platforms and passed back into the prephasing and imputation pipeline for a second round of imputation.

After the second round of imputation, genotypes with a probability >99% were retained for further analysis; other genotypes were set as missing. Imputed datasets were subjected to further QC measures using PLINK, excluding variants and individuals based on genotyping rate, complete linkage disequilibrium, systematic missingness, Hardy-Weinberg disequilibrium, allele frequency, cryptic relatedness and non-Irish ancestry, or possible technical artifact based on principal component analysis (Supplementary Materials 2). Patients that carried an established pathogenic mutation in an ALS gene were excluded. Samples that had been genotyped on more than one platform were used to assess concordance between imputed and array genotypes. The final dataset contained 2,092,686 variants at a frequency of >1% for 1,784 individuals (605 cases and 1,179 controls).

2.4. Association testing

Using PLINK, variants were tested for association with ALS risk under dominant and recessive inheritance models and using the allelic, genotypic, and Cochrane-Armitage trend tests. Clumping of linkage disequilibrium–based results was used to determine associated intervals based on an index variant threshold of $p < 5 \times 10^{-5}$ and a clumped variant threshold of 1×10^{-2}. All analyses were also carried out on a dataset with the ISC cohort removed. Genome-wide variants were also tested for association with age at onset of ALS using linear regression under additive, dominant, and recessive models.