Genomics 105 (2015) 237–241

Contents lists available at ScienceDirect

Genomics

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

Homozygosity mapping in an Irish ALS case–control cohort describes local demographic phenomena and points towards potential recessive risk loci

Russell L. McLaughlin [a,*], Kevin P. Kenna [a], Alice Vajda [b], Mark Heverin [b], Susan Byrne [b], Colette G. Donaghy [e], Simon Cronin [c,d], Daniel G. Bradley [a], Orla Hardiman [b]

[a] Population Genetics Laboratory, Smurfit Institute of Genetics, Trinity College Dublin, College Green, Dublin 2, Ireland
[b] Academic Unit of Neurology, Trinity Biomedical Sciences Institute, Trinity College Dublin, Pearse Street, Dublin 2, Ireland
[c] Beaumont Hospital, Dublin 9, Ireland
[d] Cork University Hospital, Cork, Ireland
[e] Department of Neurology, Royal Victoria Hospital, Belfast, Northern Ireland, United Kingdom

ARTICLE INFO

Article history:
Received 20 October 2014
Accepted 16 January 2015
Available online 23 January 2015

Keywords:
Amyotrophic lateral sclerosis
ALS
Runs of homozyosity
ROH
Irish population
Panmixia
Population density

ABSTRACT

Runs of homozygosity are common in European populations and are indicative of consanguinity, restricted population size and recessively inherited traits. Here, we map runs of homozygosity (ROHs) in an Irish case–control cohort for amyotrophic lateral sclerosis (ALS), a devastating neurological condition with high heritability yet only partially established genetic cause. We compare the extent of homozygosity in the Irish cohort with a large British cohort and observe that ROHs are longer and more frequent in the Irish population than in the British, and that extent of ROHs is correlated with demographic factors within the island of Ireland. ROHs are also longer and more frequent in ALS cases compared to population-matched controls, supporting the hypothesis that recessively inherited loci play a pathogenic role in ALS. Comparing homozygous haplotypes between cases and controls reveals several potential recessive risk loci for ALS, including a genomic interval spanning *ARHGEF1*, a compelling ALS candidate gene.

1. Background

Amyotrophic lateral sclerosis (ALS) is a rapidly-progressive, fatal neurodegenerative condition characterized by late-onset loss of upper and lower motor neurones resulting in paralysis and death from respiratory failure. In around 5–10% of cases there is a discernible family history of the disease [1]. The remainder of cases are termed sporadic ALS; however high heritability suggests that genetic risk factors play a role in the pathogenesis of the condition [2]. Nevertheless, the genetic cause of ALS in the majority of cases remains elusive, especially in Ireland [3–5]. Recessive inheritance of ALS could be one of many factors prohibiting the discovery of risk loci; in such cases the power of even the largest genome-wide association studies (GWAS) to date could still be low for the likely effect sizes in question. However, the argument that ALS may be inherited recessively in many cases is supported by evidence forwarded by Hemminki et al. [6], who observed higher risk for ALS between siblings with unaffected parents. Furthermore, several established ALS-causing mutations are recessive, including a subset of those in *SOD1* [7] and *OPTN* [8]. Alternative methods for the identification of recessive risk loci are therefore warranted in ALS research.

One such method is homozygosity mapping. If a variant is identical-by-descent in both parents, a signature of its recessive inheritance would be detectable in the offspring as a run of homozygosity (ROH)—an improbably long series of consecutive homozygous genotypes—surrounding the (unobserved) variant. This occurs when recombination has not disrupted the haplotype surrounding the recessive variant and mapping such regions is a powerful means to detect recessive disease loci as it increases the likelihood of detecting a genomic locus with a segregating recessive variant. This differs from single-marker association under a recessive model because such methods make fewer inferences about the haplotype on which the two alleles appear. Alternative haplotypes and extreme minor allele frequencies can therefore quench association signals at these loci.

Recently, Mok *et al.* described a study that mapped ROHs in a large, pooled multinational cohort of ALS cases and controls [9]. They used a haplotype-naïve approach to assess the association of runs of homozygosity with ALS risk on a per-locus basis, finding that a region approximately 5 Mb upstream of *TARDBP* was nominally associated with the

* Corresponding author.
E-mail addresses: mclaugr@tcd.ie (R.L. McLaughlin), kennak@tcd.ie (K.P. Kenna), vajdaa@tcd.ie (A. Vajda), mark.heverin@tcd.ie (M. Heverin), suabyrne@gmail.com (S. Byrne), donaghy1a@hotmail.com (C.G. Donaghy), scronin@rcsi.ie (S. Cronin), dbradley@tcd.ie (D.G. Bradley), hardimao@tcd.ie (O. Hardiman).

http://dx.doi.org/10.1016/j.ygeno.2015.01.002
0888-7543/

disease. Multiple populations were used in this study, with imbalanced assignment of cases and controls to each study population. It is therefore possible that some of the study's findings could have been driven by population stratification, especially since runs of homozygosity are highly predictive of population of origin [10]. Furthermore, by simply grouping overlapping ROHs and not considering the actual homozygous haplotypes that were mapped, it is possible that some subtle effects were missed. In the current study, we map runs of homozygosity in a single large ALS cohort drawn from the Irish population. We demonstrate that ROHs are longer and more numerous in Irish individuals than in the British, illustrating the utility of the Irish as a model population for studies of recessive disease. We also identify potential recessive ALS risk loci by mapping specific allelically-matching homozygous haplotypes that are overrepresented in cases compared to controls.

2. Methods

2.1. Participants

The final Irish study cohort consisted of a total of 1784 individuals, 605 of which had clinically definite, probable or possible ALS [11] and 1179 control subjects being neurologically normal at time of sampling. All participants provided written informed consent and the study was approved by the Beaumont Hospital Research Ethics Committee. ALS patients were diagnosed between 1999 and 2012 by a neurologist with expertise in the condition at the Beaumont Hospital motor neurone disease clinic (Dublin, Ireland; $n = 537$) or by neurologists at specialist centres across Northern Ireland ($n = 68$). All patients were of self-reported ancestry from the island of Ireland. Patients had a mean age of ALS onset ($\pm$ SD) of 61.3 $\pm$ 12.5 years. 27.4% of patients had bulbar onset ALS and 43.8% were female. Internal and external Irish control cohorts were used with mean age at sampling of 58.7 $\pm$ 13.6 years and 36.1 $\pm$ 12.4 years, respectively. Any ALS patient carrying an established pathogenic ALS mutation was excluded [3,4].

2.2. Genotype data

Study participants were genotyped at 561,466 to 949,974 genome-wide polymorphic sites using four different genotyping platforms, yielding four datasets designated ALS1 (201 cases, 198 controls, Illumina HumanHap550v3.0 [12]), ALS2 (106 cases, 127 controls, Illumina Human610-Quadv1.0), ALS3 (298 cases, Illumina HumanOmniExpressExome-8v1) and ISC (854 controls, Affymetrix Genome-Wide Human SNP Array 6.0; provided by the Trinity Biobank on behalf of the International Schizophrenia Consortium). The union of all four datasets (prior to quality control and filtering) contained 1,726,772 single nucleotide polymorphisms (SNPs) with just 89,905 SNPs in the intersection. A further dataset, WTCCC, representing genotype data from 2718 British individuals (Illumina Human1M-Duo BeadChip), was provided by the Wellcome Trust Case–Control Consortium for comparison of ROHs between the Irish and British populations.

Genotyping platform was found to have a significant effect on the outcome of ROH analyses unless only the intersection of SNPs between datasets was used; however this intersection was small. For this reason, for analyses concerning identification of potential recessive risk loci, a single Irish dataset was generated by merging and imputing genotypes, as described in detail elsewhere [13]. Briefly, SNPs were remapped to the GRCh37 build of the human genome and genotypes were reassigned to the forward strand, removing A/T and C/G SNPs. SNPs were filtered based on genotyping rate, systematic missingness and Hardy–Weinberg disequilibrium and individuals were filtered by gender discordance, genotyping rate and extreme heterozygosity. Datasets were then separately prephased [14] using Shape-IT [15], then imputed to the density of variant calls (SNPs and indels) in the 1000 Genomes Project [16] using IMPUTE2 [17]. Following imputation, data were separately re-filtered based on allele frequency, genotyping rate, Hardy–Weinberg disequilibrium, systematic missingness and complete linkage disequilibrium and variants were filtered if they showed high differentiation in genotype counts between case/case and control/control pairs of datasets (eg ALS1 controls vs ALS2 controls), thus removing bias introduced by imputation. Resulting datasets were merged and subjected to a second round of imputation to recover variant density, followed again by the extensive QC described above, as well as exclusions based on cryptic relatedness and population outliers determined using the smartpca programme implemented within EIGENSOFT [18].

2.3. Homozygosity mapping

ROH analyses were carried out for two general purposes. The first was to describe broad demographic phenomena and global case–control differences and the second, to search for genomic loci that may harbour recessive ALS risk variants. For the demographic analyses, a dataset containing the intersection of ALS1, ALS2, ALS3 and WTCCC was used, without any imputation of markers (223,483 SNPs in the intersection). For the identification of recessive risk loci, the imputed dataset derived from ALS1, ALS2, ALS3 and ISC was used (2,086,843 variants). ROHs were mapped using the –homozyg option in PLINK version 1.07 [19]. Parameters were set according to their default values, with some exceptions. Only one missing genotype and no heterozygotes were permitted within a ROH, which was defined as a run of consecutive homozygous genotypes greater than 1 Mb in length and containing more than 100 variants. Hemizygous deletions in ALS1, ALS2 and ALS3 misidentified as ROHs were excluded if a ROH overlapped with a region defined by the intersection of deletions identified using QuantiSNP [20] and PennCNV [21] (supplementary methods S1). To sort identified ROHs into allelically-matching groups of overlapping ROHs, PLINK's –homozyg-group option was invoked, with the –homozyg-match parameter set to 1, requiring a 100% match between alleles in overlapping ROH segments.

2.4. Demographic assessments

As well as comparing ROHs between the Irish and British cohorts, comparisons were made within Irish individuals based on geographical locale. For 508 patients, the population density was calculated for all electoral divisions (a system of administrative division in the Republic of Ireland) whose centroid points fell within a 100 km^2 circle surrounding the home address of the patient. This was used to assess the correlation between genomic runs of homozygosity and local population density. Furthermore, for 321 individuals (cases and controls), the Irish county of birth for all four grandparents was known, permitting assessment of the relationship between runs of homozygosity and possible reduced panmixia in the pedigrees of study subjects. Both geographical metrics were assessed by linear regression; population density was also assessed using the Mann–Whitney–Wilcoxon (MWW) test between patients in areas of low (<500 people/km^2) and high (>500 people/km^2) population density.

2.5. Identification of potential recessive risk loci

The procedure used to parse allelically-matching, overlapping groups of ROHs and identify loci that were overrepresented in cases is depicted in Fig. 1. Allelically-matching groups within pools of overlapping ROHs identified by PLINK were parsed out and removed if the group size was less than three segments or the group was duplicated elsewhere in the dataset. Segments were removed from groups if their overlap score was less than 0.1 (supplementary methods S2). Membership of an allelically-matching group was used as a proxy for carrying a homozygous haplotype at a locus defined by the consensus interval of all ROHs within the group. The association of the homozygous haplotype with disease risk was assessed using Fisher's exact test on counts of individuals carrying versus individuals not carrying the homozygous haplotype. Results were considered significant if the p-value was