European Journal of Human Genetics (2013) 21, 102–108

www.nature.com/ejhg

ARTICLE

The *C9ORF72* expansion mutation is a common cause of ALS+/−FTD in Europe and has a single founder

Bradley N Smith[1,16], Stephen Newhouse[1,16], Aleksey Shatunov[1,16], Caroline Vance[1], Simon Topp[1], Lauren Johnson[1], Jack Miller[1], Younbok Lee[1], Claire Troakes[1], Kirsten M Scott[1], Ashley Jones[1], Ian Gray[1], Jamie Wright[1], Tibor Hortobágyi[1], Safa Al-Sarraj[1], Boris Rogelj[1], John Powell[1], Michelle Lupton[1], Simon Lovestone[1], Peter C Sapp[2], Markus Weber[3], Peter J Nestor[4], Helenius J Schelhaas[5], Anneloor ALM ten Asbroek[6], Vincenzo Silani[7], Cinzia Gellera[8], Franco Taroni[8], Nicola Ticozzi[7], Leonard Van den Berg[9], Jan Veldink[9], Phillip Van Damme[10], Wim Robberecht[10], Pamela J Shaw[11], Janine Kirby[11], Hardev Pall[12], Karen E Morrison[12], Alex Morris[13], Jacqueline de Belleroche[13], JMB Vianney de Jong[6], Frank Baas[6], Peter M Andersen[14], John Landers[2], Robert H Brown Jr[2], Michael E Weale[15], Ammar Al-Chalabi[1,16] and Christopher E Shaw*[,1,16]

A massive hexanucleotide repeat expansion mutation (HREM) in *C9ORF72* has recently been linked to amyotrophic lateral sclerosis (ALS) and frontotemporal dementia (FTD). Here we describe the frequency, origin and stability of this mutation in ALS+/−FTD from five European cohorts (total $n=1347$). Single-nucleotide polymorphisms defining the risk haplotype in linked kindreds were genotyped in cases ($n=434$) and controls ($n=856$). Haplotypes were analysed using PLINK and aged using DMLE+. In a London clinic cohort, the HREM was the most common mutation in familial ALS+/−FTD: *C9ORF72* 29/112 (26%), *SOD1* 27/112 (24%), *TARDBP* 1/112 (1%) and *FUS* 4/112 (4%) and detected in 13/216 (6%) of unselected sporadic ALS cases but was rare in controls (3/856, 0.3%). HREM prevalence was high for familial ALS+/−FTD throughout Europe: Belgium 19/22 (86%), Sweden 30/41 (73%), the Netherlands 10/27 (37%) and Italy 4/20 (20%). The HREM did not affect the age at onset or survival of ALS patients. Haplotype analysis identified a common founder in all 137 HREM carriers that arose around 6300 years ago. The haplotype from which the HREM arose is intrinsically unstable with an increased number of repeats (average 8, compared with 2 for controls, $P<10^{-8}$). We conclude that the HREM has a single founder and is the most common mutation in familial and sporadic ALS in Europe.
European Journal of Human Genetics (2013) **21,** 102–108; doi:10.1038/ejhg.2012.98; published online 13 June 2012

Keywords: ALS; common founder; C9ORF72

INTRODUCTION

Despite apparent differences in the clinical phenotype of amyotrophic lateral sclerosis (ALS) and frontotemporal dementia (FTD), evidence of an etiopathological link between these disorders is irrefutable. ALS due to motor neuron degeneration usually presents with focal weakness in a limb or mouth/throat muscles (bulbar) and spreads relentlessly causing widespread paralysis.[1] FTD presents with changes in behaviour, personality and language due to degeneration of neurons in the frontal and temporal lobes.[2] Both disorders can be familial and in a subset of these kindreds, individuals can present with either ALS or FTD, or features of both. In 2006, we reported linkage to a 11-Mb locus on chromosome 9p13.2–21.3 in Dutch and Scandanavian kindreds with autosomal-dominant ALS-FTD.[3,4] Linkage was subsequently confirmed in eight other dominant kindreds defining a minimal overlapping region of ~3.6 Mb.[5,6] Genome-wide association studies in sporadic and familial ALS demonstrated highly significant association with single-nucleotide polymorphisms (SNPs) across a 170-Kb region at 9p21.2.[7–11]

[1]Department of Clinical Neurosciences, MRC Centre for Neurodegeneration Research, Institute of Psychiatry, Kings College London, London, UK; [2]Department of Neurology, University of Massachusetts Medical Center, Worcester, MA, USA; [3]Kantonsspital St Gallen and University Hospital Basel, Basel, Switzerland; [4]Department of Clinical Neurosciences, University of Cambridge, Cambridge, UK; [5]Department of Neurology, Radboud University Nijmegen Medical Centre, Donders Institute for Brain, Cognition and Behaviour, Centre for Neuroscience, Nijmegen, The Netherlands; [6]Department of Neurogenetics and Neurology, Academic Medical Centre, Amsterdam, The Netherlands; [7]Department of Neurology and Laboratory of Neuroscience, 'Dino Ferrari' Center, Universita' degli Studi di Milano, IRCCS Istituto Auxologico Italiano, Milan, Italy; [8]SOSD Genetics of Neurodegenerative and Metabolic Diseases, Fondazione-IRCCS, Istituto Neurologico 'Carlo Besta', Milan, Italy; [9]Department of Neurology, Rudolf Magnus Institute of Neuroscience, University Medical Center Utrecht, Utrecht, The Netherlands; [10]Laboratory of Neurobiology, Department of Neurology, K.U. Leuven, Leuven, Belgium; [11]Academic Neurology Unit, Sheffield Institute for Translational Neuroscience, Department of Neuroscience, School of Medicine and Biomedical Sciences, University of Sheffield, Sheffield, UK; [12]School of Clinical and Experimental Medicine, College of Medicine and Dentistry, University of Birmingham, and Neurosciences Division, University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK; [13]Neurogenetics Group, Centre for Neuroscience, Division of Experimental Medicine, Hammersmith Hospital Campus, London, UK; [14]Department of Pharmacology and Clinical Neuroscience, Umeå University, Umeå, Sweden; [15]King's College London, Department of Medical and Molecular Genetics, London, UK
[16]These authors contributed equally to this work.
*Correspondence: Professor CE Shaw, Department of Clinical Neurosciences, MRC Centre for Neurodegeneration Research, Institute of Psychiatry, Kings College London, 1 Windsor Walk, Denmark Hill, PO43, London SE5 8AF, UK. Tel: +44 20 7848 5180; Fax: +44 20 7848 0988; E-mail: Christopher.shaw@kcl.ac.uk
Received 15 February 2012; revised 12 April 2012; accepted 24 April 2012; published online 13 June 2012

A massive GGGGCC hexanucleotide repeat expansion mutation (HREM) has recently been identified within intron 1 of *C9ORF72* as the pathogenic mutation responsible for familial and sporadic ALS and FTD in these cases.[12,13]

Here we describe HREM mutation frequencies in ALS in five European populations. We have generated a detailed map of genetic variation across the locus that provides evidence of genomic instability, which on one occasion gave rise to a massive insertion, generating a single common founder for all the European HREM cases.

METHODS

Samples

DNA was extracted from blood and post-mortem brain frontal cortex, using the standard procedures in patients diagnosed with ALS by the revised El Escorial criteria.[14] All the ALS samples were of Northern European Caucasian origin and collected in specialist regional centres following informed consent. ALS cases were designated as familial if one or more first- or second-degree relatives developed ALS or FTD. A person was classified as having ALS + FTD if they presented with major cognitive or behavioural change at any stage during the course of their illness. Patients provided consent conforming to local and national ethics committee guidelines. In the London Clinic samples, mutations in *SOD1, VCP, OPTN* and *UBQLN2* (all exons), *TDP43* (exon 6) and *FUS* (exons 14 and 15) were screened and any positive samples were excluded from further analysis. Familial samples from the other European cohorts were screened and excluded for *SOD1*, *FUS* and *TARDBP* mutations.

9p21.2 locus-capture and sequencing

DNA from 12 individuals carrying the disease haplotype and 4 without from our previously linked kindred,[3] 2 affected members from the previously published linked Scandanavian family,[4] 14 cases with suggested linkage to ch9p and 21 other individuals with familial ALS + / −FTD were processed for DNA capture using custom-designed overlapping probes (Roche Nimblegen, Madison, WI, USA) across the 3.6-Mb locus between D9S169 (27 238 617)[5] and D9S251 (30 819 382).[6] A total of 5 μg of DNA was fragmented with a Bioruptor (Wolf Laboratories Ltd, York, UK) at 30 s on/off bursts for 45 mins to sizes of 200–300 bp. End repair was followed by addition of adenine ends and ligation of adaptors (Illumina, Little Chesterford, Essex, UK) and peak sizes checked using a Bioanlayzer (Agilent Technologies, Wokingham, UK). Purification steps were conducted with SPRI beads according to the manufacturer's instructions (Beckman Coulter Genomics, High Wycombe, UK). Libraries were hybridized with 4.5 μl of locus probes for 72 h at 47 °C, washed and bound to streptavidin beads, followed by PCR of 10 separate reactions per library. Individual reactions were pooled, cleaned using QIAquick (Qiagen, Crawley, UK), quantified by chip (DNA 1000, Agilent Technologies) and sequenced with 76 or 100 bp paired-end reads on GAII and Hiseq Analyzers (Illumina).

Sequencing data processing and analysis

Raw sequencing data were mapped to the human reference genome (Build hg19) using Novoalign (http://www.novocraft.com/) and processed using Picard tools v1.35 and the Genome Analysis Toolkit (GATK, version V1.1) to produce a 'clean' BAM file.[15] SNP and Indel calling was performed using the Unified Genotyper module in GATK in batch mode. The resulting Variant Call Format (VCF, version 4.0) file was annotated using Variant Filtration in GATK set as follows: QUAL < 30.0||QD < 5.0||HRun > 10||SB > −5.00||DP < 10 and cluster size 10. The VCF file was converted to 'pedigree' format using vcftools v1.3.1 allowing us to phase all SNPs on the risk allele (http://vcftools.sourceforge.net).[16] A PERL script was written to identify sequencing reads from the fastq files overlapping the HREM, and to count the numbers of repeats found within each.

SNP genotyping and haplotype analysis

A total of 82 SNPs spanning the locus that were shared among the affected individuals from the ch9p-linked families and highly represented in familial ALS/FTD cases were Kaspar genotyped (Kbiosciences Ltd, Hoddesdon, UK) in a cohort of 434 cases and 856 controls of European ancestry from Sweden, Belgium, England and Italy. In all, 16 cases from the locus-capture set were included to validate next-generation genotypes. Haplotypes were generated and frequencies were determined using PLINK v1.07,[17] and phasing was corroborated using Snphap (D. Clayton; http://www-gene.cimr.cam.ac.uk/clayton/software). Hardy–Weinberg equilibrium was assessed by a χ^2 test for quality control. DMLE + v2.3[18] was used to estimate the age of the HREM via the expected relationship between HREM allele frequency, local linkage disequilibrium and population growth rate. For comparison, we also used a decay in linkage disequilibrium method (Equation 1),[19] averaging the age estimates across all 82 SNPS. Between-SNP genetic distances were estimated using LDhat applied to HapMap Phase II (The International HapMap Consortium, 2007). The 82 SNPs span 110 186 bases, with the HREM estimated to be 105 131 bases from the telomeric SNP. We assumed that a random population sample of 39 091 would be expected to yield 137 HREM-bearing individuals, based on our observed frequency of 3 in 856 controls, and we assumed that such a sample would form a fraction of 7.8×10^{-5} of all Europeans, based on a current population size of 500 million. We performed DMLE + using a burn-in of 20 000 iterations followed by runs of 100 000 iterations, with population growth rates in Europe of 5%,[20] with a lower limit of 2.5%[21] and an upper limit of 8.5%,[22] and with a 25-year inter-generation time, with lower and upper limits of 20 and 30 years respectively.

Evolutionary conservation of the repeat region

Exons 1A, 1B and the intervening intron were aligned from human, chimpanzee, gorilla, orangutan, mouse and rat reference genomes using ClustalW and manually edited in GeneDoc.

Genotyping and sequencing across the GGGGCC repeat

A total of 1347 ALS + / −FTD patients, including the 434 cases and 856 controls used in the haplotyping study, were screened for the GGGGCC HREM, using repeat primer PCR[13] with a final concentration of 7% DMSO, 1 M betaine, 0.17 mM of 7-deaza-2-deoxy GTP, 0.7–1.4 μM of primer mix, 0.85 mM of $MgCl_2$, 50% Applied Biosystems True Allele PCR Premix (Applied Biosystems, Warrington, UK) and 100 ng of genomic DNA. Primers included a FAM-labelled reverse primer, one repeat-specific forward primer with an attached anchor sequence and the same anchor sequence as an independent forward primer. Cycling conditions were denaturation 95 °C for 15 mins and touchdown from 70 to 56 °C with 3 min extension. Fragment analysis was conducted on an ABI 3130 genetic Analyser and peaks visualized using Genemapper 4.0 (Applied Biosystems). Chromatograms were scored as mutant (sawtooth pattern) or wild type (< 30 repeats). Direct sequencing of 48 cases and controls without the HREM was performed using Big Dye V1.1 chemistry and an ABI 3130 genetic analyser to validate repeat primer PCR genotypes, forward primer 5′-GGTTTAGGAGGTGTGTGTTTTTGT-3′, reverse primer 5′-CCAGCTTCGGTCAGAGAAAT-3′ and identical cycling conditions with two extra cycles at each stage of the touchdown protocol.

Association analysis

Unless otherwise stated, all calculations were performed in IBM SPSS v19 (SPSS Inc., Chicago, IL, USA) with two-sided significance tests. Independence of categorical variables was tested using the χ^2 distribution. For small cell counts, the Fisher's exact test was used. Alleles of the highly polymorphic non-expanded hexanucleotide repeat were tested for association using Monte-Carlo simulation in the program CLUMP,[23] which generates empirical *P*-values for observed χ^2 tables, accounting for the multiple testing inherent in having multiple alleles at a locus. Age of onset and disease duration was tested for association with the HREM using Kaplan–Meier product limit estimate and the log rank test.

RESULTS

Mutation frequencies by phenotype and country

Mutations in the familial ALS + / −FTD cohort from the King's College Hospital, London clinic, were identified in 55% of all familial

cases with the following frequency: *C9ORF72* 29/112 (26%), *SOD1* 27/112 (24%), *FUS* 4/112 (4%) and *TARDBP* 1/112 (1%). HREM mutations were also detected in 13/216 (6%) of unselected sporadic ALS cases from the same clinic. No mutations were identified in *VCP*, *OPTN* or *UBQLN2*.

Combining data from five European populations, (detailed individually in Table 1), the HREM in *C9ORF72* was detected in 226/1347 (17%) of all ALS+/−FTD cases, in whom known ALS genes had been excluded, and 3/856 (0.3%) controls (Fisher's exact test *P*-value for allelic association $= 4.12 \times 10^{-47}$; OR = 57, 95% CI = 17.7–224.6). The highest frequency was in familial ALS+FTD kindreds (48/67, 72%) but it was also prevalent in pure ALS kindreds (89/228, 39%), with the total familial frequency therefore being 46% (137/296, *P*-value 6.13×10^{-89}; OR = 244, 95% CI = 74.4–974.3) In sporadic ALS+/−FTD, HREM frequencies across Europe were higher than for any other known gene at 87/1048 (8%) (*P*-value 1.1×10^{-19}; OR 25.7, 95% CI = 7.8–102). Given that sporadic ALS accounts for 95% of all cases, then sporadic ALS+/−FTD cases with the HREM outnumber familial by a ratio of 4:1. Frequencies of the HREM in familial ALS+/−FTD were high but showed considerable variation by country: 19/22 (86%) in Belgium, 30/41 (73%) in Sweden, 10/27 (37%) in the Netherlands, 73/185 (39%) in England and 4/20 (20%) in Italy.

Genotype and phenotype

Phenotypic data were available on 189 ALS cases with the HREM and 870 cases without HREM (Table 2). The male:female ratio in HREM-positive cases was 1.1:1 compared with non-HREM cases 1.8:1 ($P = 0.009$), which is similar to population-based studies of familial and sporadic disease (Table 2). Patients with the expansion were more likely to present with cognitive/behavioural and bulbar symptoms than those without ($P = 0.02$). Kaplan–Meier estimates showed no difference between the two groups in the age at onset ($P = 0.27$) or disease duration ($P = 0.34$) (Supplementary Figure 1).

Characterising haplotypes across the locus

Sequencing of DNA captured across the 3.6-MB locus in 53 individuals generated 1.2 billion reads with an average of 487-fold depth across the region. We identified 10 604 SNPs that passed QC but no variants segregating with disease were identified in *C9ORF11*, *MOB3B*, *IFNK*, *C9ORF72* or *LINGO2* or other predicted genes within the locus. The largest number of GGGGCC repeats detected within intron 1 of *C9ORF72* was 8, occurring at the end of a single read in one affected individual. The pathological HREM was not identified because it is so GC-rich and fails to amplify by PCR without a repeat-specific primer. Thus, it is not surprising that none of the variant-calling algorithms we used detected this polymorphism.

We were able to phase 82 informative SNPs within the linked kindreds that defined a shared haplotype across the locus. These were further genotyped in 433 cases and 856 controls and correlated with the presence of the HREM. Detailed inspection of the SNP haplotype in the 137 HREM-positive cases revealed that a full 82-SNP haplotype existed in the vast majority of cases (111/137, 81%, $P = 8.33 \times 10^{-17}$) and 3 controls who were positive for the HREM. Despite significant recombination, alleles from the linked haplotype were always preserved in regions flanking one or other side of the HREM (80 SNPs telomeric and 2 SNPs centromeric), providing clear evidence of a single common founder in these European populations (Figure 1). The two SNPs centromeric to the HREM were rs1 17 89 520 and rs73440960, possessing allele frequencies in HREM-positive cases of 0.97% ($P = 0.00023$) and 0.93 ($P = 0.00006$), respectively, which demonstrate that the conserved haplotype straddles the expansion region.

Estimating the age of the founder event

Estimates of the age of the HREM using DMLE+ depend primarily on the growth rate of the population and generational interval, which are known to vary greatly over time (Table 3). Growth rates ranging from 2.5–8.5% and intergenerational intervals have been proposed between 20 and 30 years for most founder studies. If we take 5% as a conservative estimate of growth and 25 years as an intergenerational average, we estimate that the mutation arose around 251 generations ago, which equates roughly to 6300 years (see Table 3 for estimates based on a range of growth rates and intergenerational intervals). For comparison, we applied an alternative linkage disequilibrium method[19] and estimated that the mutation arose 131 generations ago (3300 years ago assuming a 25-year intergeneration time). We

Table 2 Gender, genotype and phenotype

Sex	*WT (n = 870)*	*HREM (n = 189)*
Male	564 (64.8%)	101 (53.4%)
Female	306 (35.2%)	88 (46.6%)
Site of symptom onset	*WT (n = 818)*	*HREM (n = 203)*
Bulbar	219 (26.8%)	77 (37.9%)
Spinal	573 (70%)	110 (54.1%)
FTD	26 (3.2%)	16 (8%)

Male to female ratios and frequencies of limb, bulbar and FTD onset in WT and HREM-positive cases.

Table 1 Mutation frequencies by clinical diagnosis and country

		Clinical diagnosis				
Country	*Cases (n)*	*FALS*	*FALS/FTD*	*SALS*	*SALS/FTD*	*FTD*
London clinic	296	15/64 (23%)	14/16 (87.5%)	13/216 (6%)		
Other United Kingdom	870	35/93 (37%)	10/13 (77%)	58/737 (7.9%)	6/27 (22%)	0
Sweden	77	10/16 (63%)	20/25 (80%)	3/21 (14%)	1/12 (8.3%)	2/3 (67%)
The Netherlands	27	10/27 (37%)				
Belgium	22	16/19 (84%)	3/3 (100%)	0	0	0
Italy	55	3/10 (30%)	1/10 (10%)	1/5 (20%)	5/30 (17%)	0
Total	**1347**	**89/229 (39%)**	**48/67 (72%)**	**75/979 (7.6%)**	**12/69 (17%)**	**2/3 (66%)**

Mutation frequencies of the expansion repeat in cases according to clinical diagnosis and country. The total number of cases screened is indicated in bold on the bottom line including subtotals and corresponding mutation frequency according to clinical diagnosis.

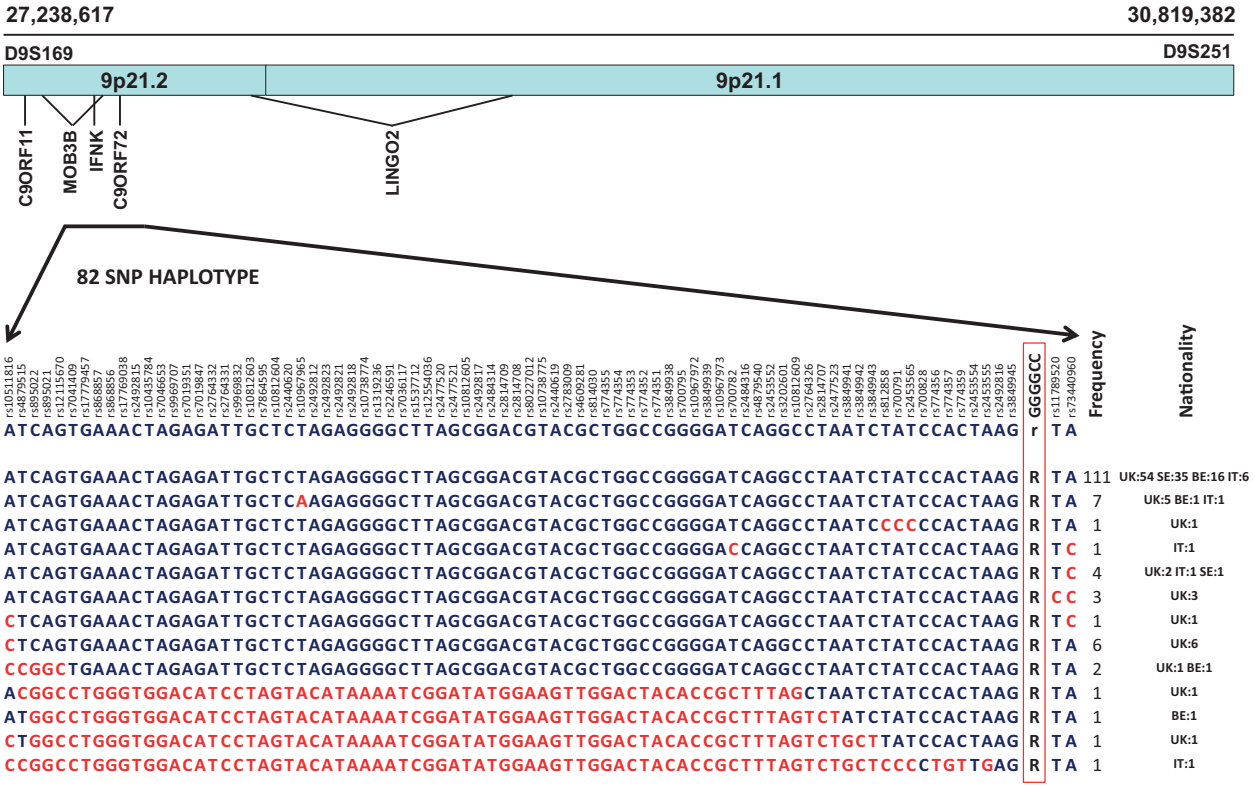

Figure 1 Details of the 82-SNP risk haplotype defined by rs10511816 (27468461 hg19) to rs7344960 (27578647 hg19), covering 110 kb region between *MOB3B* and *C9ORF72*. The top row represents the background haplotype on which the expansion arose (r), with the founder expansion directly below it (R). An additional 12 recombined HREM haplotypes are also shown along with their representation within the case cohort. The non-risk allele is highlighted in red.

Table 3 Age of the hexanucleotide repeat expansion mutation

Growth rate (%)	*Generations*	*Years*
8.5	157 (134–196)	3900 (2700–5900)
5	251 (220–287)	6300 (4400–8600)
2.5	479 (419–550)	12 000 (8400–16 500)

Estimates of the age of the HREM for a range of per-generation human population, growth estimates are given in 'generations' (with 2.5 and 97.5% quantiles) and 'years', estimated with a 25-generation interval (with 20- and 30-year intervals).

acknowledge the limitations of these analytical tools but it is encouraging that these figures are not greatly disparate.

Genomic instability of the GGGGCC repeat region

The human, chimpanzee and gorilla reference genomes contain three copies of the GGGGCC repeat (Figure 2a). Evidence from the NCBI Trace Archive database shows that chimpanzees can also have five or six repeats. No other species appear to contain the hexanucleotide motif, although orangutans possess a possible precursor sequence, 5′-GAGGCCGGGCCC-3′. Phylogenetic analysis of the human haplotypes reveals two main clades, one of which gave rise to the expansion mutation (Figure 2b).

The full background 82-SNP haplotype from which the HREM arose is present in almost all populations studied in the '1000 Genome' database of 1046 individuals (Figure 3). Its frequency in people of European ancestry averages at ~15.1%, which is nearly identical to the frequencies we derived from our analysis of controls of 262/1706 (14.9%) and our ALS+/−FTD cohort of 109/740 (14.8%). Repeat primer PCR genotyping does not give the number of repeats for the HREM allele, however, the longest number of repeats can be counted in cases without the expansion using fragment analysis. Sanger sequencing of 48 individuals showed a perfect correlation between the repeat number counted by fragment analysis and sequencing (Supplementary Figure 2). We measured the longest number of repeats in 1154 individuals and compared those with the background haplotype (r) to all other haplotypes (Figure 2c). The average number of repeats in those carrying haplotype (r) was 8 with a widespread of expanded alleles up to 26 (95% CI = 4–13), whereas the most prevalent number of repeats in all the other haplotypes was only 2 (95% CI = 1–7, $P<10^{-8}$). This indicates that the background haplotype on which the expansion arose is intrinsically unstable, tending to generate longer repeats.

We have also identified that rs2492816 independently tagged a repeat number >2 for the risk allele ($P<1.0E-13$, Fisher's exact test) and a repeat number of 2 for the non-risk allele ($P<1.0E-52$, Fisher's exact test), which accounts for the apparent bimodality of the non-risk haplotype distribution.

DISCUSSION

C9ORF72 mutations are common in familial and sporadic ALS

We have demonstrated that the hexanucletotide repeat expansion mutation in *C9ORF72* is the most common genetic cause of familial ALS+/−FTD across Europe, accounting for 20–86% of genetically undiagnosed familial cases, particularly where FTD and ALS co-segregate and in those presenting with bulbar or cognitive/behavioural symptoms. It is difficult to make robust conclusions about the origin of differences in frequency due to the small sample sizes for each country but they probably reflect the influence of a

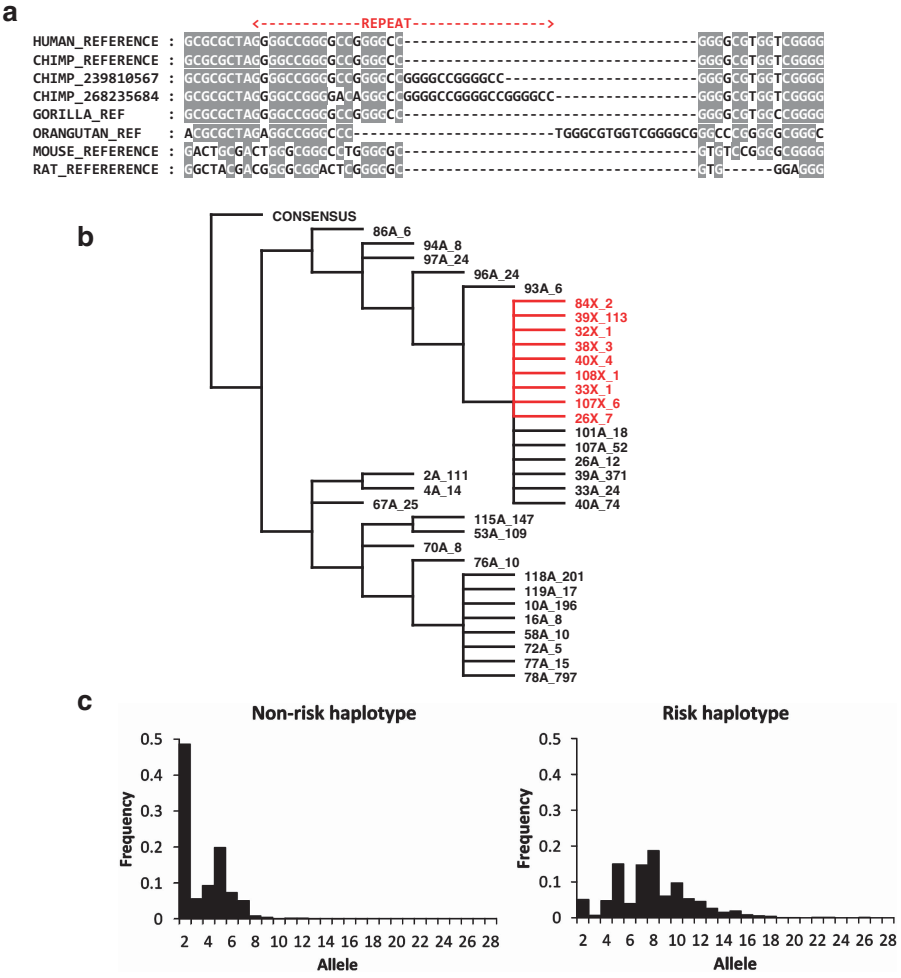

Figure 2 (a) Multiple alignment of the region surrounding the HREM from various mammals, showing the polymorphism in chimpanzees. Digits in identifiers refer to NCBI Trace Archive (ti) accession numbers, for example, ti 26 82 35 684. **(b)** Phylogeny of the unique haplotypes observed within our sample set, showing how the HREM occurs only within a single, distinct, clade of risk-associated haplotypes. Identifiers with an X contain the expanded HREM allele (highlighted in red). Digits after the underscore indicate the number of chromosomes in which the haplotype was observed. The phylogram was constructed from the consensus of 3807 best-scoring trees produced by the phylip 3.69 dnapars algorithm (Felsenstein, 1989), rooted using an unweighted consensus from all 139 unique haplotypes. To retain clarity, only those haplotypes with the HREM or those without but seen five or more times are shown. Additionally, nine haplotypes that had undergone major recombination events were also removed. Four of these contained the expansion (4 chromosomes) and five did not (33 chromosomes). **(c)** Figures showing repeat allele frequencies for risk and non-risk haplotypes. The repeat sizes are smaller for the non-risk haplotypes, consistent with the hypothesis that the risk haplotype predisposes to repeat instability. The difference in repeat allele frequency distribution for the two haplotype patterns is highly significant ($P \ll 10^{-8}$).

founder effect. In unselected patients from a London clinic with a family history of ALS (5–10% of all cases), a genetic diagnosis can now be confirmed in 55% of cases. *C9ORF72* HREM was the most common mutation (26%) followed by *SOD1* (24%), *FUS* (4%) and *TARDBP* (1%). The HREM is detected in 6% of sporadic ALS cases but is also present in the background population (0.3%). The penetrance of the HREM appears to be low, given that there is a common founder and the ratio of sporadic to familial HREM cases is 4:1. This is consistent with the incomplete penetrance reported in many linked kindreds. Further work is required to generate figures for age-related penetrance that can be used in genetic counselling and predictive gene testing.

The HREM arose from a common European founder around 6300 years ago

Following exhaustive sequencing, we have confidently identified a haplotype that proves that all HREM carriers arose from a single common founder. We phased a 82-SNP haplotype within linked kindreds that is conserved in its entirety in the majority of all the HREM carriers and flanks at least in part all of the remaining HREM cases. The most economical explanation is that the expansion mutation arose on just one occasion in the European population, however, we cannot exclude the possibility that it arose on multiple occasions on the same background haplotype (r). Estimates of founder age depend heavily on estimates of population growth rates, with smaller rates leading to older estimates, and to a lesser extent on intergenerational interval. Historical evidence is that growth rates have varied greatly being much slower in the distant past than in the last century.[24] Using averaged figures of a 5% growth rate and 25-year interval, we have estimated that the founder mutation arose around 6300 years ago (range 4400–8600 years).

A common founder was originally proposed for Finnish FALS cases based on a 42-SNP haplotype.[8] A subsequent meta-analysis of genome-wide association study data from five European

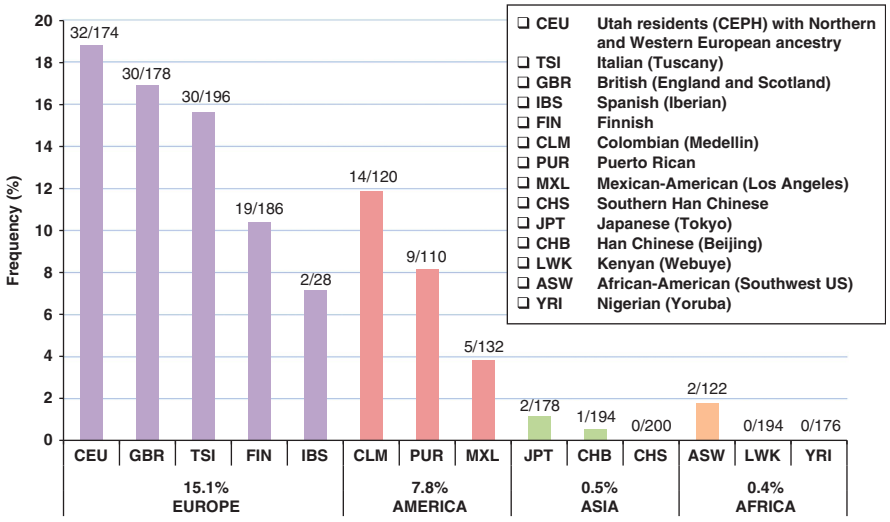

Figure 3 Bar chart showing how the frequency of the founder risk haplotype varies across continents and populations (data from the 1000 Genomes project, www.1000genomes.org), and how it is most prevalent in Europeans. Percentages indicate the number of chromosomes on which the haplotype was observed.

populations (Finnish, Irish, UK, US and Italian) reduced this to a common 20-SNP risk haplotype.[25] In the original report of the HREM by the same group, however, only two-thirds of Finnish cases were reported to have a common haplotype, implying that the other third of their HREM cases may have different founders.[13] By fine mapping across the locus in great detail, we have shown that all the HREM carriers (cases and controls) have conserved SNPs that flank one or other side of the HREM, confirming that all the carriers arose from a single founder haplotype (r). Given that the mean age at onset of our HREM cases is 60 years (see Supplementary Figure 1) and the penetrance is relatively low, we doubt that significant selection pressures would apply over past millennia where life expectancy was considerably lower. For these reasons, we would not expect the mutation to die out due to selective purification.

Hexanucleotide repeat instability is greater on the founder background haplotype (r)

We have uncovered evidence that the GGGGCC repeat arose during primate evolution and is highly polymorphic but the biological significance of this is unknown. Haplotype frequencies in different ethnic populations from the 1000 Genomes database strongly suggest a European origin for the background 82-SNP haplotype (r) on which the HREM arose. The maximum number of repeats on either allele is much greater in those with the (r) haplotype than all other haplotypes combined. Nearly 50% of the individuals with non-(r) haplotypes have a maximum of two copies (295/641) compared with 5% (29/513) of individuals with the (r) haplotype, where the average is eight repeats and some individuals have many more. This difference in repeat number confirms initial observations based on a single SNP rs38 49 942 marker for the HREM risk haplotype.[12] It is not clear why the (r) haplotype is prone to expansion but it is possible that 8–26 repeats, which are GC-rich, promote the formation of hairpin secondary loop structures that impair DNA replication. For instance, flap endonuclease 1 required for normal maturation of Okazaki fragments during replication fails to process flaps folded into aberrant hairpin structures and is thought to cause expansion at CAG repeats.[26,27] Alternatively, an independent *de novo* event may have occurred in a single person 6300 years ago, which affected the fidelity of DNA polymerase or a DNA mismatch repair enzyme, which in conjunction with the unstable repeat region resulted in the HREM.[28,29]

Role of the HREM in ALS and FTLD biology

The dominant pathology in 90% of ALS and tau-negative FTD inclusions contain the TAR DNA-binding protein (TDP-43) within the cytoplasm of neurons and glia.[30] TDP-43 inclusions are also prominent in cases linked to chromosome 9p[31] but HREM-specific pathology includes abundant cytoplasmic and intranuclear p62-positive inclusions in the hippocampus and cerebellum that are TDP-43-negative.[32,33] Precisely, how the HREM causes TDP-43 mislocalisation and neurodegeneration is not currently known. Evidence that the HREM reduces levels of *C9ORF72* transcripts implicates a loss of function, however, probes detecting the HREM transcript identified RNA foci within the nuclei of neurons in the frontal cortex and spinal cord.[12] In other dominant intronic repeat disorders, such as Myotonic Dystrophy (DM1), these foci have been shown to sequester RNA-binding proteins, which cause a range of deleterious changes in RNA processing.[34,35]

We propose that all *C9ORF72* HREM cases derive from a single common founder and are now the most common cause of familial and sporadic ALS in Western Europe. The GGGGCC repeat is highly polymorphic and particularly unstable in the context of a specific haplotype (r), but the massive pathogenic expansion may have arisen on just one occasion around 6300 years ago. Although gene testing will become widely available, further work is required to establish the disease risk for HREM carriers.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

This work was supported by the NIHR Biomedical Research Centre for Mental Health at the South London and Maudsley NHS Foundation Trust and Institute of Psychiatry, Kings College London, and the SLAGEN consortium. We would like to thank Antonia Ratti, Cinzia Tiloca, Barbara Castellotti, Viviana Pensato, Stefania Corti, Roberto del Bo, Gianni Sorarù, Carla

D'Ascenzo, Sandra D' Alfonso, Lucia Corrado, Cristina Cereda, Ceroni Mauro and Isabella Fogh for their help. This work was funded by the Medical Research Council, Motor Neuron disease Association (UK), American ALS Association and the Heaton-Ellis Trust. The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under the grant agreement number 259867.

1 Shaw CE, al-Chalabi A, Leigh N: Progress in the pathogenesis of amyotrophic lateral sclerosis. *Curr Neurol Neurosci Rep* 2001; **1**: 69–76.
2 Lomen-Hoerth C, Anderson T, Miller B: The overlap of amyotrophic lateral sclerosis and frontotemporal dementia. *Neurology* 2002; **59**: 1077–1079.
3 Vance C, Al-Chalabi A, Ruddy D *et al*: Familial amyotrophic lateral sclerosis with frontotemporal dementia is linked to a locus on chromosome 9p13.2-21.3. *Brain* 2006; **129**: 868–876.
4 Morita M, Al-Chalabi A, Andersen PM *et al*: A locus on chromosome 9p confers susceptibility to ALS and frontotemporal dementia. *Neurology* 2006; **66**: 839–844.
5 Luty AA, Kwok JB, Thompson EM *et al*: Pedigree with frontotemporal lobar degeneration – motor neuron disease and Tar DNA binding protein-43 positive neuropathology: genetic linkage to chromosome 9. *BMC Neurol* 2008; **8**: 32.
6 Boxer AL, Mackenzie IR, Boeve BF *et al*: Clinical, neuroimaging and neuropathological features of a new chromosome 9p-linked FTD-ALS family. *J Neurol Neurosurg Psychiatry* 2010; **82**: 196–203.
7 Shatunov A, Mok K, Newhouse S *et al*: Chromosome 9p21 in sporadic amyotrophic lateral sclerosis in the UK and seven other countries: a genome-wide association study. *Lancet Neurol* 2010; **9**: 986–994.
8 Laaksovirta H, Peuralinna T, Schymick JC *et al*: Chromosome 9p21 in amyotrophic lateral sclerosis in Finland: a genome-wide association study. *Lancet Neurol* 2010; **9**: 978–985.
9 van Es MA, Veldink JH, Saris CG *et al*: Genome-wide association study identifies 19p13.3 (UNC13A) and 9p21.2 as susceptibility loci for sporadic amyotrophic lateral sclerosis. *Nat Genet* 2009; **41**: 1083–1087.
10 Van Deerlin VM, Sleiman PM, Martinez-Lage M *et al*: Common variants at 7p21 are associated with frontotemporal lobar degeneration with TDP-43 inclusions. *Nat Genet* 2010; **42**: 234–239.
11 Rollinson S, Mead S, Snowden J *et al*: Frontotemporal lobar degeneration genome wide association study replication confirms a risk locus shared with amyotrophic lateral sclerosis. *Neurobiol Aging* 2011; **32**:758 e751–e757.
12 Dejesus-Hernandez M, Mackenzie IR, Boeve BF *et al*: Expanded GGGGCC hexanucleotide repeat in noncoding region of C9ORF72 causes chromosome 9p-linked FTD and ALS. *Neuron* 2011; **72**: 245–256.
13 Renton AE, Majounie E, Waite A *et al*: A hexanucleotide repeat expansion in C9ORF72 is the cause of chromosome 9p21-Linked ALS-FTD. *Neuron* 2011; **72**: 257–268.
14 Brooks BR, Miller RG, Swash M, Munsat TL: El escorial revisited: revised criteria for the diagnosis of amyotrophic lateral sclerosis. *Amyotroph Lateral Scler Other Motor Neuron Disord* 2000; **1**: 293–299.
15 McKenna A, Hanna M, Banks E *et al*: The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010; **20**: 1297–1303.
16 Danecek P, Auton A, Abecasis G *et al*: The variant call format and VCFtools. *Bioinformatics* 2011; **27**: 2156–2158.
17 Purcell S, Neale B, Todd-Brown K *et al*: PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007; **81**: 559–575.
18 Reeve JP, Rannala B: DMLE+: Bayesian linkage disequilibrium gene mapping. *Bioinformatics* 2002; **18**: 894–895.
19 Rannala B, Bertorelle G: Using linked markers to infer the age of a mutation. *Hum Mutat* 2001; **18**: 87–100.
20 Weale ME, Weiss DA, Jager RF, Bradman N, Thomas MG: Y chromosome evidence for Anglo-Saxon mass migration. *Mol Biol Evol* 2002; **19**: 1008–1021.
21 Meddison A: *Contours of the World Economy, 1-2030 AD: Essays in Macro-Economic History.* New York: Oxford University Press, 2007.
22 Hastbacka J, de la Chapelle A, Kaitila I, Sistonen P, Weaver A, Lander E: Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. *Nat Genet* 1992; **2**: 204–211.
23 Sham PC, Curtis D: Monte Carlo tests for associations between disease and alleles at highly polymorphic loci. *Ann Hum Genet* 1995; **59**: 97–105.
24 Risch N, de Leon D, Ozelius L *et al*: Genetic analysis of idiopathic torsion dystonia in Ashkenazi Jews and their recent descent from a small founder population. *Nat Genet* 1995; **9**: 152–159.
25 Mok K, Traynor BJ, Schymick J *et al*: The chromosome 9 ALS and FTD locus is probably derived from a single founder. *Neurobiol Aging* 2011; **33**: e203–e208.
26 Spiro C, McMurray CT: Nuclease-deficient FEN-1 blocks Rad51/BRCA1-mediated repair and causes trinucleotide repeat instability. *Mol Cell Biol* 2003; **23**: 6063–6074.
27 Henricksen LA, Tom S, Liu Y, Bambara RA: Inhibition of flap endonuclease 1 by flap secondary structure and relevance to repeat sequence expansion. *J Biol Chem* 2000; **275**: 16420–16427.
28 Daee DL, Mertz TM, Shcherbakova PV: A cancer-associated DNA polymerase delta variant modeled in yeast causes a catastrophic increase in genomic instability. *Proc Natl Acad Sci USA*, **107**: 157–162.
29 Foiry L, Dong L, Savouret C *et al*: Msh3 is a limiting factor in the formation of intergenerational CTG expansions in DM1 transgenic mice. *Hum Genet* 2006; **119**: 520–526.
30 Neumann M, Sampathu DM, Kwong LK *et al*: Ubiquitinated TDP-43 in frontotemporal lobar degeneration and amyotrophic lateral sclerosis. *Science* 2006; **314**: 130–133.
31 Pearson JP, Williams NM, Majounie E *et al*: Familial frontotemporal dementia with amyotrophic lateral sclerosis and a shared haplotype on chromosome 9p. *J Neurol* 2011; **258**: 647–655.
32 Al-Sarraj S, King A, Troakes C *et al*: p62 positive, TDP-43 negative, neuronal cytoplasmic and intranuclear inclusions in the cerebellum and hippocampus define the pathology of C9orf72-linked FTLD and MND/ALS. *Acta Neuropathol* 2011; **122**: 691–702.
33 Troakes C, Maekawa S, Wijesekera L *et al*: An MND/ALS phenotype associated with C9orf72 repeat expansion: Abundant p62-positive, TDP-43-negative inclusions in cerebral cortex, hippocampus and cerebellum but without associated cognitive decline. *Neuropathology*, e-pub ahead of print 19 December 2011.
34 Miller JW, Urbinati CR, Teng-Umnuay P *et al*: Recruitment of human muscleblind proteins to (CUG)(n) expansions associated with myotonic dystrophy. *EMBO J* 2000; **19**: 4439–4448.
35 Kanadia RN, Shin J, Yuan Y *et al*: Reversal of RNA missplicing and myotonia after muscleblind overexpression in a mouse poly(CUG) model for myotonic dystrophy. *Proc Natl Acad Sci USA* 2006; **103**: 11748–11753.

Supplementary Information accompanies the paper on European Journal of Human Genetics website (http://www.nature.com/ejhg)