Aller, E.; Larrieu, L.; Jaijo, T.; David Baux; Espinós-Armero, CÁ.; González-Candelas, F.; Nájera, C.... (2010). The USH2A c.2299delG mutation: dating its common origin in a Southern European population. European Journal of Human Genetics. 18(7):788-793. https://doi.org/10.1038/ejhg.2010.14

TITLE:

The *USH2A* c.2299delG mutation: dating its common origin in a Southern European population.

RUNNING TITTLE: Allelic age of the *USH2A* c.2299delG mutation**.**

AUTHORS:

Elena Aller[1,2], Lise Larrieu[3], Teresa Jaijo[1,2], David Baux[3], Carmen Espinós[2,4], Fernando González-Candelas[5], Carmen Nájera[6], Francesc Palau[2,4], Mireille Claustres[3,7,8], Anne-Françoise Roux[3,7], José M Millán[1,2].

AFFILIATIONS:

[1]Unidad de Genética, Hospital Universitario La Fe, Valencia, Spain.

[2]CIBER de Enfermedades Raras (CIBERER), Valencia, Spain.

[3] CHU Montpellier, Laboratoire de Génétique Moléculaire, Montpellier, France.

[4]Unidad de Genética y Medicina Molecular, Instituto de Biomedicina de Valencia (CSIC), Valencia, Spain.

[5]Instituto Cavanilles de Biodiversidad y Biología Evolutiva, Universitat de València: Genómica y Salud (CSISP, Valencia) and CIBERESP, Spain.

[6]Departamento de Genética, Universitat de València, Valencia, Spain.

[7] Inserm, U827, Montpellier, France.

[8] Univ, Montpellier I, Montpellier, France.

Corresponding author:

Jose M Millán, PhD

Unit of Genetics
Hospital Universitario La Fe
Avda. Campanar, 21
46009 Valencia. Spain
Tel. +34 96 197 3153
Email. millan_jos@gva.es
chema_millan@yahoo.es

ABSTRACT

Usher syndrome type II is the most common form of Usher syndrome. *USH2A* is the main responsible gene of the three known to be responsible for this pathology. It encodes two isoforms of the protein usherin. This protein is part of an interactome that plays an essential role in the development and function of inner ear hair cells and photoreceptors. The gene contains 72 exons spanning over a region of 800 kb. Although numerous mutations have been described, the c.2299delG mutation is the most prevalent in several populations. Its ancestral origin was previously suggested after the identification of a common core haplotype restricted to 250 kb in the 5´ region, which encodes the short usherin isoform. By extending the haplotype analysis over the 800 kb region of the *USH2A* gene with a total of 14 intragenic SNPs, we have been able to define 10 different c.2299delG haplotypes showing high variability but preserving the previously described core haplotype. An exhaustive c.2299delG/control haplotype study suggests that the major source of variability in the *USH2A* gene is recombination. Furthermore, we have evidenced twice the amount of recombination hotspots located in the 500 kb that covers the 3´ end of the gene, explaining the higher variability observed in this region when compared to the 250 kb of the 5' region. Our data confirm the common ancestral origin of the c.2299delG mutation and the estimation of its allelic age reveals that c.2299delG may have arisen about 5 500-6 000 years ago.

KEYWORDS:

***USH2A*, c.2299delG, haplotype, dating**

INTRODUCTION

Usher syndrome type II (USH2) belongs to a genetically and phenotypically heterogeneous group of recessively inherited disorders that combine hearing loss and retinitis pigmentosa (RP). More specifically, USH2 displays moderate to severe hearing loss, postpubertal onset of RP and normal vestibular reflexes. Although three genes are responsible for USH2, *USH2A* accounts for more than 75% of USH2 cases[1,2]. Usher syndrome type IIA (USH2A; MIM 276901) represents the most common form of inherited deaf-blindness and is estimated to affect 1 in 17,000 individuals[3].

The underlying *USH2A* gene was isolated by positional cloning[4]. It was initially described as including 21 exons, with the first exon being entirely non-coding, spanning a region of 250 kb and it was predicted to encode a 1546 amino acid protein of 171KDa. Today, this protein is recognized as the short isoform of usherin and is predicted to be a secreted extracellular protein[5].

Because mutation detection rates obtained in mutation screening studies were lower than those expected, the existence of additional uncharacterised exons of *USH2A* was postulated. van Wijk *et al.*[6] identified 51 novel exons at the 3´end of the gene increasing its size to 800 kb. These authors also provided some indications for alternative splicing. The predicted protein encoded by the longest open reading frame (5202 residues) is a member of the protein network known as the Usher interactome. This interactome plays an essential role in the development of the stereocilia of the hair cells in the organ of Corti. In photoreceptors, the Usher interactome localises in the periciliary region and could be involved in the cargo transport between the inner and outer segment[7,8,9].

Since the identification of the *USH2A* long isoform, a small number of mutation screenings have been reported, which indicates that screening of all of the encoding 72 exons is mandatory for efficient molecular diagnosis[1,2,10,11,12]. As a result of these studies, together

with the mutations of the short isoform reported before 2004, more than 210 mutations have been described. A great majority of these mutations are private or present in a few families[13]. However, a prevalent mutation located in exon 13, designated as c.2299delG, is frequently found in European and US patients but also in isolated cases from South America, South Africa and China. The allele frequency distribution of c.2299delG varies geographically in Europe. This mutation represents 30.6% of the USH2A cases in a Scandinavian population[2], while an allelic frequency of 31% was found in the Netherlands[14], 16 to 36 % in the UK[15,16] and 15 and 10% in Spain[10] and France (unpublished result), respectively. A common ancestral origin has been hypothesised for the c.2299delG mutation on the basis that alleles bearing the c.2299delG mutation share the same core haplotype, restricted to the first 21 exons of the *USH2A* gene[17]. In this study, we carry out an exhaustive analysis of the 52 additional exons of the long isoform which reveals high variability in numerous associated intragenic Single Nucleotide Polymorphisms (SNPs) giving rise to, at least, 10 different c.2299delG haplotypes. Using *USH2A* (c.2299delG/control) haplotyes data we have estimated the allelic age of this mutation, revealing that c.2299delG may have arisen about 5 500-6 000 years ago.

MATERIAL AND METHODS

Patients

Twenty-seven patients were included in this study. Seventeen were of Spanish origin and were recruited from the Federación de Asociaciones de Afectados de Retinosis Pigmentaria del Estado Español (FAARPEE) and from the Ophthalmology and ENT Services of several Spanish hospitals. Ten patients were French and were recruited from medical genetic and ophthalmology clinics distributed all over France. The patients were classified as Usher type II on the basis of ophthalmologic studies, including visual acuity, visual field and fundus

ophthalmoscopy, electroretinography, pure-tone and speech audiometry and vestibular evaluation. For each patient, samples from parents were considered as well as siblings, when possible. This study was approved by both the Hospital La Fe and CHU Montpellier Ethical Committees and consent to genetic testing was obtained from adult probands or parents in the cases of minors.

Controls

97 control chromosomes were used to establish the distribution of *USH2A* normal alleles. They were generated from fifty trios (subject and both parents). Twenty-five of each were of French and Spanish origins and randomly chosen as the healthy control group. These trios did not refer symptoms or a history of Usher syndrome or related disorders.

DNA analysis of *USH2A* gene

Patient and control genomic DNA was extracted from peripheral blood samples using standard protocols. The 14 SNPs used to construct *USH2A* haplotypes of control and c.2299delG alleles were PCR-amplified using the primers and PCR conditions previously described (5,6). PCR products were directly sequenced on an ABI PRISM 3130xl (Applied Biosystems, CA, U.S.A). The polymorphism IVS17-8T>G was not considered in this study. Because this variant was included in the core haplotypes defined by Dreyer *et al*.[17] we indicate it in brackets to avoid any confusion when referring to Dreyer's data.

Construction of haplotypes

Parents and available siblings of the c.2299delG patients were used to infer the haplotypes linked to the c.2299delG mutation (M haplotypes). Similarly, control trios were used to establish normal *USH2A* haplotypes in a healthy population (C haplotypes). In some cases, the data were not informative enough to establish the phase of the SNPs and some ambiguities

remained. When possible, ambiguous haplotypes were ascribed to an already existing haplotype.

Construction of phylogenetic trees.

Relationships between haplotypes were inferred using two approaches with three different data sets: the complete set of SNPs, the first five SNPs included in the first 21 exons of the gene, and the last 9 SNPs located in the 3' end of the gene (see table 1). In the first approach, we constructed phylogenetic trees using a variety of methods and evolutionary models. However, the high levels of homoplasy present in this dataset prevented the derivation neither of a single most reliable phylogenetic tree, neither with the whole set nor with any of the other subsets of SNPs. Among the different trees obtained, we present the results obtained with the neighbor-joining method[18], using the uncorrected number of differences between pairs of SNPs as a measure of their genetic divergence. Bootstrap support values were obtained using version 4.1 of the MEGA software.

Additionally, a median-joining network was obtained with the program Network 4.5.10 (Fluxus Technology, http://www.fluxus-technology.com). A network represents all the alternative possibilities linking every haplotype considered through a minimum number of mutation steps and is not restricted to represent relationships as a single pathway. This is a more appropriate methodology than dichotomous phylogenetic trees for establishing relationships among closely related allele variants[19].

Dating the *USH2A* c.2299delG mutation

To estimate the original date of the c.2299delG mutation in the *USH2A* gene, two mathematical approaches were applied: a Monte Carlo likelihood method implemented in the program BDMC21 v2.1[20] (http://www.rannala.org/labpages/software.html) and a Markov chain method by means of the DMLE+ v2.2 software[21,22,23] (http://www.dmle.org). For both

approaches, we used a generation time of 25 years and a carrier frequency of USH2 of 1/106 for Spain and France (data not shown). The program BDMC21 v2.1 relies on the assumption that genetic variation among a group of highly linked polymorphic markers, defining a haplotype in which a novel non-recurrent mutation arose, is a function of the mutation frequencies of those linked markers and the time since the first occurrence of this unique mutation. To achieve this approach, we considered information from the three variable SNPs closest to the c.2299delG mutation: c.4714C>T, c.6506T>C and c.6875G>A. Confidence interval was estimated following the standard theory of maximum likelihood estimation[24]. The second analysis performed was using the DMLE+ program version 2.2, which takes into account the marker information from the entire haplotype on the basis of:

5'c.373G>A_c.504A>G_c.1419C>T_IVS15+35G>A_c.4457G>A_c.4714C>T_c.6506T>C_c.6875G>A_c.10232A>C_c.11602A>G_c.11677C>A_c.12612A>G_c.12666A>G_c.13191G>A 3'.

This program allows Bayesian inference of the mutation age based on the observed linkage disequilibrium at multiple genetic markers.

RESULTS

c.2299delG haplotypes

The c.2299delG haplotypes were built for the 27 *USH2A* patients using the 14 SNPs represented in table 1. Seven of the patients were c.2299delG homozygotes (six were Spanish and one was French). A total of 10 different haplotypes were identified (M1-M10, see Table 2). The haplotypes were identical from exon 2 to 21, but the SNPs located along the 52 additional exons of the *USH2A* long isoform were variable (table 2). The variability rate of

the SNPs was uneven. For five of the SNPs (c.4714C>T, c.6875G>A, c.11602A>G, c.11677C>A and c.13191G>A) the same allele was present on at least 8 haplotypes.

Haplotype M1 was the most frequent in the Spanish population (8/23; frequency 0.35) followed by haplotype M2 (6/23; frequency of 0.26). Haplotype M1 was also the most prevalent in France, together with haplotype M8 (3/11; frequency 0.27). Haplotypes M4-M9 were restricted to either the Spanish or French populations. Haplotype M1 was the most common with a frequency of 0.32 (11/34) when both populations were pooled. **Control *USH2A* haplotypes**

Fifty-four different haplotypes could be defined from the 97 control chromosomes (Table 3). High variation was found along the entire gene, including the region encompassing exons 1 to 21. However, two SNPs remained invariable: c.4714C>T and c.11677C>A. In addition, the c.6875G>A SNP had the same G allele in 53 of 54 haplotypes. Interestingly, this variant corresponds to the only CpG dinucleotide identified among the 14 SNPs (table 1). Haplotype C1 was the most prevalent among the Spanish control population with a frequency of 0.1 (5/51) and haplotype C6 was the most frequent among the French controls with a frequency of 0.09 (4/46). Combining the data from both populations, haplotype C1 was the most prevalent with a frequency of 0.07 (7/97).

Relationship of haplotypes

The neighbor-joining tree for all the entire haplotypes was rooted with the corresponding *Pan troglodytes* haplotype. It did not present a well-defined structure, since none of the nodes were supported by bootstrap analysis. Nevertheless, a small cluster encompassing six haplotypes related to the disease (M1, M2, M5-M8) was observed. The remaining disease-associated haplotypes did not group with this clade, but were not too distant from it (Fig. 1A).

This pattern was very different from that obtained when only the 5 SNPs from the first 21 exons of the gene were analyzed. The common haplotype including disease-related alleles as well as many others from control chromosomes occupies an intermediate position between the oldest haplotypes, as inferred from their close relationship to the out-group, and the most recently derived, the group including C40-C46. Again, none of the nodes in this tree were supported by bootstrap analysis (Fig. 1B). This topology is markedly different from the one inferred from the remaining SNPs, those located at the 3'-end of the gene. Here, there was no longer a clear association between disease-related haplotypes, except for a small group including alleles M5-M8. Most of the other disease-related haplotypes were more closely related to control alleles than to any other disease allele but, again, these associations were not supported by bootstrap analysis (Fig. 1C).

The apparent lack of congruence between the phylogenetic histories of these alleles when considering SNPs from the 5'- and 3'-ends may be due to frequent recombination events. This was further checked by reconstruction of median-joining networks for the same three data sets described above. The three networks, but especially those derived from the complete and the 3'-end sets of SNPs, present a high level of connectedness with many alternative routes connecting every possible pair of haplotypes (see figures 2A, 2B and 2C). There are also many haplotypes connected to several others with a minimum number of intermediate steps and only a few haplotypes are connected to the rest through a single intermediate. This pattern is still present, although at a much reduced level, in the network derived from SNPs in the 5'-end of the gene (figure 2B), partly due to the reduced number of different haplotypes in this part of the gene. The ancestral (C19, which includes *Pan troglodytes*) and the most abundant haplotypes are connected through an intermediate haplotype (either C28 or C17) and two point changes in SNPs c.3157+15G>A and c.4457G>A. These observations easily explain the

difficulties encountered in reconstructing a phylogenetic tree with well supported relationships as previously commented. Although it is certainly possible to invoke homoplasic point mutations to explain these patterns, they are more likely due to high level of recombination, with an apparently higher rate in the second part of the gene.

Dating the c.2299delG mutation

The haplotype data were used to estimate the original date of the *USH2A* c.2299delG mutation using two mathematical approaches, the BDMC21 v2.1 program and the DMLE+ v2.2 software. Haplotype data were analysed for the Spanish and French populations separately and then in the pooled populations, although similar results were obtained (see table 4). These results showed that the c.2299delG mutation arose 219-245 generations ago. Assuming a generation time of 25 years, this would indicate that the *USH2A* c.2299delG mutation arose about 5 500-6 000 years ago.

DISCUSSION

c.2299delG/general population haplotypes

The data obtained from the entire haplotypes of the control population reveal a highly variable genetic background, since 54 haplotypes could be identified in the Spanish and French populations with no evidence of a prevalent common haplotype (table 3). In 2001, twelve core haplotypes were identified by Dreyer *et al*.[17] in a Scandinavian control population. These were based on partial information since only part of the *USH2A* gene was then recognized. These authors identified a major haplotype "A-G-C-A-(T)-A" with a frequency of 0.60. This core haplotype is also the most frequent one in our control group (C1 to C16), but overall represents less than 50%. The same core haplotype is found in all c.2299delG alleles within

the first 21 exons, confirming the existence of high linkage disequilibrium in this 250 kb region.

The C>T distribution of the c.4714 SNP is quite striking. The C allele is present in all control haplotypes, but it is carried by only two disease-associated haplotypes, M9 and M10, that represent less than 15 % of the c.2299delG alleles. Linkage disequilibrium between the c.4714T allele and the c.2299delG mutation had already been noted in a French study[1]. Dreyer *et al.*[2] identified this SNP in both the c.2299delG and control Scandinavian alleles. However, we do not know if the majority of the c.2299delG patients in North Europe also carry the T allele at this position. Extending the studies to Northern Europe and other populations should help to clarify this point.

The variability observed in the additional portion of the gene covering from exon 22 to 72 (i.e. about 500 kb) is quite puzzling. We analyzed the mutability rate of *USH2A* SNPs by looking at CpG dinucleotides (table 1). Only one CpG was found in exon 36 at position 6875 and, therefore, cannot explain the variability observed in the 3' region. The median-joining networks reconstructed in order to find the relationship between haplotypes showed a high level of connectedness, especially for the second part of the gene. These networks are more easily explained by the existence of high recombination rates than by point mutations. So, we can conclude that recombination events probably represent the predominant source of variability in this gene. Subsequently, we looked for a common sequence motif CCNCCNTNNCCNC associated with recombination hot spots in humans[25] along the entire *USH2A* DNA sequence. Twenty motif locations were found, 4 within the first 20 introns and 16 between introns 21-71. Therefore, twice amount of recombination hotspots are located in the most variable region.

Origin and spread of the c.2299delG mutation

Dating calculations estimates that c.2299delG could have arisen about 5 500-6 000 years ago. After the Last Glacial Maximum (15 000-10 000 years ago), the human populations that settled in the warmer areas of the Mediterranean coasts of the Iberian peninsula, Southern France and the Balkans began the colonization of Central and North-western Europe[26]. Whether this colonization was due to demic or cultural diffusion is controversial, although many authors argue for the latter[27,28].

The c.2299delG mutation shows a cline of decreasing frequency from Scandinavia to the Mediterranean countries[1,2,10,14,15,16,29]. This phenomenon has been observed for many other disease-causing mutations, such as delta F508 responsible for Cystic Fibrosis[30], p.R408W in the *PAH* gene responsible for Phenylketonuria[31], c.2588G>C in *ABCA4* responsible for Stargardt disease[32] and many others; some of which have been used to estimate the history of human migrations throughout Europe.

It is tempting to speculate that c.2299delG arose in a Neolithic population that lived on the Mediterranean coast and spread to Central Europe along the waterways as a leap-frog migration[33]. Thus, the high frequency of c.2299delG in Scandinavia could be due to one, or successive bottlenecks; whereas the migration fluxes that occurred between the Mediterranean populations (Phoenicians, Greeks, Romans, Arabs and North-Africans) and the mixing of the genetic pools of these peoples has contributed to the decrease of the allele frequency of c.2299delG in contemporary French and Spanish populations.

There are no data concerning the c.2299delG frequency among North-Africans. However, c.2299delG is not a prevalent mutation in the non-Ashkenazi Jewish populations from the South and Near East regions[11,16,34,35]. This supports the hypothesis of the more recent migration fluxes across the Mediterranean Sea as a cause of the reduced frequency of the

c.2299delG mutation within Northern Mediterranean populations. However, it is quite possible that, due to both the moderate and high variable USH2 phenotype, it remains under-diagnosed and therefore underestimated in certain populations. Another interesting point is the presence of c.2299delG in a Chinese population. This mutation has been found in isolated patients of Chinese origin[15]. The recent study carried out by Dai *et al.*[12] indicated that c.2299delG is not common among Chinese USH2 patients; although the authors only screened six patients and further studies are needed in order to investigate the frequency of c.2299delG in this and other non-European populations.

In relation to those territories with a history of European colonization, such as America and South Africa, it has already been pointed by Dreyer *et al.*[17] that the recent waves of European migration to the New World and other countries would definitely explain the presence of c.2299delG in these populations.

The exhaustive study of the 3´ region of the *USH2A* gene in our cohort of patients has revealed that haplotypes linked to the c.2299delG mutation show high variability, but preserve the previously described core haplotype "A-G-C-A-(T)-A". This common haplotype is restricted to 250 kb in the 5´ region of this gene, which corresponds to the USH2A protein short isoform. By extending this study to the control population we have evidenced the existence of linkage disequilibrium restricted to this 250 kb region. The analysis of the relationship between *USH2A* haplotypes suggests that the major source of variability in this gene is recombination. The higher variability observed in the 3´ region could be explained by the accumulation of recombination hotspots observed in specific intronic sequences of this portion of the gene. It is tempting to speculate that the structural and dynamic differences observed between the 5´and 3´region of the gene could have a functional significance.

Haplotypes data have also been used to estimate the allelic age of c.2299delG. Based upon historical migration fluxes of humans and the geographical distribution of this mutation, c.2299delG may have arisen arose 5 500-6 000 years ago in a Neolithic population that lived on the Mediterranean coast and, later on, spread to Central and Northern Europe.

ACKNOWLEDGEMENTS

Authors are grateful to the participating patients and their relatives and to the FAARPEE for their help and co-operation. This work was supported by grants from the Fondo de Investigaciones Sanitarias (FIS07/0558) and from Ministère de la Recherche "PHRC National 2004" (Protocole 7802).

We also acknowledge Fabiola Barraclough for English corrections.

REFERENCES

1. Baux D, Larrieu L, Blanchet C *et al*: Molecular and in silico analyses of the full length isoform of usherin identify new pathogenic alleles in Usher type II patients. *Hum Mutat* 2007, **28**(8):781-9.
2. Dreyer B, Brox V, Tranebjaerg L *et al*: Spectrum of USH2A mutations in Scandinavian patients with Usher syndrome type II. *Hum Mutat* 2008, **29**(3):451.
3. Kimberling WJ. Estimation of the frequency of occult mutations for an autosomal recessive disease in the presence of genetic heterogeneity: application to genetic hearing loss disorders**.** *Hum Mutat* 2005, **26**(5):462-70.

4. Eudy JD, Weston MD, Yao S *et al*: Mutation of a gene encoding a protein with extracellular matrix motifs in Usher syndrome type IIa. *Science* 1998, **280**(5370):1753-7.
5. Weston MD, Eudy JD, Fujita S *et al*: Genomic structure and identification of novel mutations in usherin, the gene responsible for Usher syndrome type IIa. *Am J Hum Genet* 2000, **66**(4):1199-210.
6. Van Wijk E, Pennings RJ, te Brinke H *et al*: Identification of 51 novel exons of the Usher syndrome type 2a (USH2A) gene that encode multiple conserved functional domains and that are mutated in patients with Usher syndrome type II. *Am J Hum Genet* 2004, **74**(4):738-44.
7. Adato A, Lefèvre G, Delprat B *et al*: Usherin, the defective protein in Usher syndrome type IIA, is likely to be a component of interstereocilia ankle links in the inner ear sensory cells. *Hum Mol Genet* 2005, **14**(24):3921-32.
8. Liu X, Bulgakov OV, Darrow KN *et al*: Usherin is required for maintenance of retinal photoreceptors and normal development of cochlear hair cells. *Proc Natl Acad Sci* U S A. 2007, **104**(11):4413-8.
9. Maerker T, van Wijk E, Overlack N *et al*: A novel Usher protein network at the periciliary reloading point between molecular transport machineries in vertebrate photoreceptor cells. *Hum Mol Genet* 2008, **17**(1):71-86.
10. Aller E, Jaijo T, Beneyto M *et al*: Identification of 14 novel mutations in the long isoform of USH2A in Spanish patients with Usher syndrome type II. *J Med Genet* 2006, **43**(11):e55.

11. Auslender N, Bandah D, Rizel L *et al*: Four USH2A Founder Mutations Underlie the Majority of Usher Syndrome Type 2 Cases among Non-Ashkenazi Jews. *Genet Test* 2008, (**2**):289-94.
12. Dai H, Zhang X, Zhao X *et al*: Identification of five novel mutations in the long isoform of the USH2A gene in Chinese families with Usher syndrome type II. *Mol Vis* 2008, **14**:2067-75.
13. Baux D, Faugère V, Larrieu L *et al*: UMD-USHbases: a comprehensive set of databases to record and analyse pathogenic mutations and unclassified variants in seven Usher syndrome causing genes. *Hum Mutat* 2008, **29**(8):E76-E87.
14. Pennings RJ, Te Brinke H, Weston MD *et al*: USH2A mutation analysis in 70 Dutch families with Usher syndrome type II**.** *Hum Mutat* 2004, **24**(2):185.
15. Liu XZ, Hope C, Liang CY *et al*: A mutation (2314delG) in the Usher syndrome type IIA gene: high prevalence and henotypic variation. *Am J Hum Genet* 1999, **64**(4):1221-5.
16. Leroy BP, Aragon-Martin JA, Weston MD *et al*: Spectrum of mutations in USH2A in British patients with Usher syndrome type II. *Exp Eye Res* 2001, **72**(5):503-9.
17. Dreyer B, Tranebjaerg L, Brox V *et al*: A common ancestral origin of the frequent and widespread 2299delG USH2A mutation. *Am J Hum Genet* 2001, **69**(1):228-34.
18. Saitou N and Nei M: The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* 1987, **4**:406-25.
19. Bandelt H-J, Forster P, Röhl A: Median-joining networks for inferring intraspecific phylogenies. *Mol Biol Evol* 1999, **16**:37-48).
20. Slatkin, M. and Rannala, B: Estimating the age of alleles by use of intraallelic variability. *Am J Hum Genet* 1997, **60**, 447-58.

21. Rannala, B. and Slatkin, M: Likelihood analysis of disequilibrium mapping, and related problems. *Am J Hum Genet* 1998, **62**, 459-73.
22. Reeve, J.P. and Rannala, B. DMLE+: Bayesian linkage disequilibrium gene mapping. *Bioinformatics* 2002, **18**, 894-5.
23. Rannala, B. and Reeve, J.P: Joint Bayesian estimation of mutation location and age using linkage disequilibrium. *Pac Symp Biocomput* 2003, 526-34.
24. Sorensen, D.A. and Gianola, D: Likelihood, Bayesian and MCMC methods in quantitative genetics. Springer-Verlag, New York. 2002.
25. Myers S, Freeman C, Auton A, Donnelly P, McVean G: A common sequence motif associated with recombination hot spots and genome instability in humans. *Nat Genet* 2008, **40**:1124-9
26. Clark JDG: Radicarbon dating and the expansion of farming culture from the Near East over Europe. *Proc Prehist Soc* 1965, **31**:57-73
27. Sokal RR, Oden NL and Wilson C: Genetic evidence for the spreading of agriculture in Europe by demic diffusion**.** *Nature* 1991, **315**:143-145.
28. Chiki L, Nichols RA, Barbujani G , Beaumont MA: Y genetic data support the Neolithic demic diffusion model. *Proc Natl Acad Sci* USA 2002, **99**:11008-11013.
29. Dreyer B, Tranebjaerg L, Rosenberg T, Weston MD, Kimberling WJ, Nilssen O: Identification of novel USH2A mutations: implications for the structure of USH2A protein. *Eur J Hum Genet* 2000, **8**(7):500-6.
30. Morral N, Bertranpetit J, Estivill X *et al*: The origin of the major cystic fibrosis mutation (delta F508) in European populations. *Nat Genet* 1994, **7**:169-75.
31. Tighe O, Dunican D, O'Neill C *et al*: Genetic diversity within the R408W phenylketonuria mutation lineages in Europe. *Hum Mutat* 2003, **21**(4):387-93.

32. Maugeri A, Flothmann K, Hemmrich N *et al*: The ABCA4 2588G>C Stargardt mutation: single origin and increasing frequency from South-West to North-East Europe. *Eur J Hum Genet* 2002, **10**(3):197-203.

33. Cruciani F, La Fratta R, Santolamazza P *et al*: Phylogeographic analysis of haplogroup E3b (E-M215) Y chromosomes reveals multiple migratory events within and out of Africa. *Am J Hum Gene* 2004, **74**:1014-1022

34. Adato A, Weston MD, Berry A, Kimberling WJ, Bonne-Tamir A: Three novel mutations and twelve polymorphisms identified in the USH2A gene in Israeli USH2 families. *Hum Mutat* 2000, **15**(4):388.

35. Kaiserman N, Obolensky A, Banin E, Sharon D: Novel USH2A mutations in Israeli patients with retinitis pigmentosa and Usher syndrome type 2. *Arch Ophthalmol* 2007, **125**(2):219-24.

TITLES AND LEGENDS TO FIGURES

Figure 1. Phylogenetic trees constructed using *USH2A* haplotypes data.

A. Neighbor-joining tree constructed using the complete set of SNPs (Hamming distance).
B. Neighbor-joining tree constructed using the first 5 SNPs (Hamming distance).
C. Neighbor-joining tree constructed using the last 9 SNPs (Hamming distance).

Pan: *Pan troglodytes*

Figure 2. Median-joining networks representing all the alternative possibilities linking every *USH2A* haplotype.

A. Median-joining network constructed using the complete set of SNPs (Hamming distance).

mv1-mv12 nodes represent haplotypes that have not been found in the sample of study. These haplotypes are automatically generated by programe Network 4.5.10 in order to connect the haplotyes found in our study.

PAN: *Pan troglodytes*.

B. Median-joining network constructed using the first 5 SNPs (Hamming distance).

M1 node includes haplotypes M1-M10 + C1-C15.

C19 node includes haplotypes C19-C21 + C24-C26 + Pan (*Pan troglodytes*).

C29 node includes haplotypes C29-C31.

C33 node includes haplotypes C33-C35.

C40 node includes haplotypes C40-C46.

C48 node includes haplotypes C48-C49.

C50 node includes haplotypes C50-C51.

C52 node includes haplotypes C52-C53.

C. Median-joining network constructed using the last 9 SNPs (Hamming distance).

mv1-mv12 nodes represent haplotypes that have not been found in the sample of study. These haplotypes are automatically generated by programe Network 4.5.10 in order to connect the haplotyes found.

PAN: *Pan troglodytes*.

M9 node includes haplotypes M9, C29 and C47.

M10 node includes haplotypes M10, C14 and C37.

C1 node includes haplotypes C1, C34, C40 and C48.

C2 node includes haplotypes C2 and C25.

C3 node includes haplotipes C3 and C49.

C5 node include haplotypes C5, C17, C33, C36, C48 and C41.

C6 node includes haplotypes C6, C26, C42 and C53.

C7 node include haplotypes C7 and C39.

C9 node includes haplotypes C9, C28, C35, C43 and C51.

C15 node includes haplotypes C15, C20, C31 and C46.

C19 node includes haplotypes C19 and C45.

C24 node includes haplotypes C24 and C54.

Table 1. Location and repartition of the 14 SNPs used to establish the USH2A haplotypes.

The encoded short and long transcripts are indicated. Total distance between c.373G>A and c.13191G>A is approximately 730 kb. The distance between two SNP can be less than 1 kb. Entrez accession number is indicated for each SNP except for c.4714C>T, which is in linkage disequilibrium with the c.229delG mutation and is not referenced in dbSNP.

Table 2. Representation of the ten different c.2299delG linked haplotypes.

Table 3. Representation of the fifty-four different control *USH2A* haplotypes.

Table 4. Summarized results of c.2299delG dating using BMC21 and DMLE+ programs. Results are given in number of generations with a confidence interval of 95%.