European Journal of Human Genetics (2015) 23, 1158–1164

www.nature.com/ejhg

ARTICLE

# Confirmation of a founder effect in a Northern European population of a new β-globin variant: HBB:c.23_26dup (codons 8/9 (+AGAA))

Nina Marchi[1,2,3], Serge Pissard[4,5], Manuel Cliquennois[6], Christian Vasseur[4], Nathalie Le Metayer[4], Claude Mereau[7], Jean Pierre Jouet[8], Anne-France Georgel[6], Emmanuelle Genin[1,9] and Christian Rose[6]

**β-Thalassemia is a genetic disease caused by a defect in the production of the β-like globin chain. More than 200 known different variants can lead to the disease and are mainly found in populations that have been exposed to malaria parasites. We recently described a duplication of four nucleotides in the first exon of β-globin gene in several families of patients living in Nord-Pas-de-Calais (France). Using the genotypes at 12 microsatellite markers surrounding the β-globin gene of four unrelated variant carriers plus an additional one recently discovered, we found that they shared a common haplotype indicating a founder effect that was estimated to have taken place 225 years ago (nine generations). In order to determine whether this variant arose in this region of Northern Europe or was introduced by migrants from regions of the world where thalassemia is endemic, we genotyped the first 4 unrelated variant carriers and 32 controls from Nord-Pas-de-Calais for 97 European ancestry informative markers (EAIMs). Using these EAIMs and comparing with population reference panels, we demonstrated that the variant carriers were very similar to the controls and were closer to North European populations than to South European or Middle-East populations. Rare β-thalassemia variants have already been described in patients sampled in non-endemic regions, but it is the first proof of a founder effect in Northern Europe.**
*European Journal of Human Genetics* (2015) **23,** 1158–1164; doi:10.1038/ejhg.2014.263; published online 3 December 2014

## INTRODUCTION

Thalassemia is one of the most common genetic diseases worldwide (60 000 new-borns every year). It results from a defect in the production of either the alpha (α-thalassemia) or the beta (β-thalassemia) globin chains forming the tetramer of hemoglobin A1 (Alpha 2, Beta 2). β-Thalassemia is a heterogeneous group of autosomal recessive disorders characterized by the reduced ($\beta^+$) or absent ($\beta^0$) production of the β-globin chain.[1] More than 200 variants have been described and classified in HbVar database (http://globin.bx.psu.edu/hbvar).[2]

Sixty years ago, JBS Haldane first suggested that the uneven global distribution of β-globin disorders might be explained by malaria.[3–6] He hypothesized that β-globin variants could have been naturally selected, providing some kind of protection against malaria: the red blood cells of variant carriers could be more resistant to attacks by the sporozoa causing malaria. For discussion and review of this hypothesis, see Taylor *et al*,[3] Fowkes *et al*[5] and Weatherall.[6]

Indeed, β-thalassemia variants occurred in a wide range of ethnic groups, and they are prevalent in some human populations that have historically been exposed to malaria parasite, mainly *Plasmodium falciparum*.[3] There are several reports of β-thalassemia mutations in patients from Northern Europe with no known ancestry in regions where thalassemia is endemic,[7–9] but so far it cannot be ruled out that these mutations might have been introduced by recent migrations from regions where β-thalassemia mutations are frequent.

In 2010, we found a new variant in the β-globin gene (duplication of four nucleotides from the codons 8/9 (+AGAA) of the first exon (HGVS nomenclature: NM_000518.4(HBB):c.23_26dup, NP_000509 (HBB).1: p.(Ser10Glufs*15))). This variant is registered into HbVar database (http://globin.bx.psu.edu/hbvar/menu.html) with ID 2928.

The variant was observed in eight unrelated families (a total of 20 carriers).[10] Curiously, none of them had known ancestors from any area where β-thalassemia is endemic. The aim of this work was first to evidence a founder effect for this variant and estimate the age of the most recent common ancestor and second to infer by genetic analyses the native location (Northern Europe) of this new β-globin variant.

## METHODS

### Material

*β-Thalassemia carriers.* Five unrelated variant carriers were included in the study of the founder effect, and four of them were genotyped for ancestry informative markers to infer their geographic origin. All subjects gave informed consent for genetic analysis of β-globin gene during routine hematology consultation for biological abnormalities (non-explained microcytosis).

*Control population from Northern Europe.* A total of 32 individuals were enrolled in this study to serve as controls. They were volunteers recruited within the University or Hospital whose all four grandparents were born in

[1]Inserm UMR-946, Variabilité Génétique et Maladies Humaines, Institut Universitaire d'Hématologie, Université Paris Diderot, Centre d'Etude du Polymorphisme Humain, Paris, France; [2]Eco-Anthropologie et Ethnobiologie, UMR 7206 CNRS, MNHN, Université Paris Diderot, Sorbonne Paris Cite, France; [3]Ecole Normale Supérieure de Lyon, Master BioSciences, Lyon, France; [4]Département de Génétique, GHU H. Mondor-A. Chenevier, Créteil, France; [5]UPEC and INSERM-IMRB, Créteil, France; [6]Service d'Hématologie, Hôpital St Vincent de Paul, Université Catholique de Lille, GHICL, Lille, France; [7]Service de Biochimie, CHU, Lille Cedex, France; [8]Service des Maladies du Sang, Hôpital Claude Huriez, CHU Lille, Lille, France; [9]Inserm UMR-1078, CHU Morvan, Brest, France
*Correspondence: Professor C Rose, Service d'Onco-Hématologie, Université Catholique de Lille, Boulevard de Belfort, Lille 59000, France. Tel: +33 3 20 87 45 32; Fax: +33 3 20 87 45 85; E-mail: Rose.Christian@ghicl.net
Received 13 February 2014; revised 27 October 2014; accepted 30 October 2014; published online 3 December 2014

Nord-Pas-de-Calais (a region in France with four million people and located in the Northern Europe plain). Consent to a blood sample for genetic studies and with no direct benefit was signed by each of them after they received an information letter. An agreement to store DNA was obtained from the French agency of health products (AFSAPSS registered number 2010-A00653-36) and an agreement from the ethical committee of Lille 'Comité consultatif de protection des personnes' for the study was obtained in November 2010.

This control population was referred to as the 'Chtimi' sample where 'Chtimi' means 'people living in Nord-Pas-de-Calais' in a French dialect.

*Genetic data*. The geographic origin of the Chtimi sample and of four of the five variant carriers was determined by using a panel of 97 SNPs selected for being informative of North West–South East ancestry in Europe[11] Supplementary Data S1. SNP data of this study could be made available on request to authors.

Two worldwide reference control panels were used to determine the geographic origin:

- HapMap3 with 1151 individuals from 11 groups,[12] Supplementary Data S2A
- HGDP with 1056 individuals from 52 groups,[13] Supplementary Data S2B

To estimate the age of the variant, 12 microsatellites (D11S902, D11S905, D11S914, D11S915, D11S935, D11S1338, D11S4046, D11S4102, D11S4116, D11S4146, D11S4149, D11S4181), located on chromosome 11 around the β-globin gene cluster were genotyped in the 5 variant carriers and in 32 controls to obtain allele frequency estimates Supplementary Data S3A–C. All these STRs are registered and fully described in the UniSTS database: 'http://www.ncbi.nlm.nih.gov/unists'. A database containing all genetic data: microsatellites and SNPs for mutation carriers and controls from the Chtimi sample, has been deposited at the European Genome-phenome Archive (EGA, http://www.ebi.ac.uk/ega/), which is hosted by the EBI, under accession number EGAS00001000980.

## Methods

*Genotyping*. DNA was extracted from blood using a DNA extraction kit and according to the manufacturer's instructions (FlexiGene, QIAGEN (Venlo, Netherland), ref. 51206). The genotyping of the 97 European Ancestry Informative SNPs was done at the Département de génétique, CHU Hôpital H. Mondor_A. Chenevier, using allele-specific PCR assays (KASPar, KBioscience, Hoddesdon, UK) run in a real-time PCR device (Light Cycler 480 Instrument, Roche, Mannheim, Germany). The sizing of the 12 microsatellites was performed using the fluorescent PCR assays run in a Sanger sequencing device (Applied Biosystems 3130*xL* Genetic Analyzer, Life Technologies, New York, NY, USA).

*Estimation of the age of the variant*. The age of the variant was estimated by using the likelihood-based method implemented in ESTIAGE[14] on the microsatellite data of the five variant carriers after reconstructing variant carrier haplotypes by a parsimony method. This parsimony method consists of selecting among the different possible haplotype reconstructions the one that maximizes the length of the haplotype shared by all variant carriers. Briefly, starting from the location of the variant, each side was considered independently. For example, on the right side of the variant, we started from the first marker and we determined the most frequent allele seen in the genotypes of the variant carriers at this marker. This allele would be called $R_1$. Then, we restricted the sample to the individuals carrying this allele $R_1$ at the first marker and looked at the second marker on the right side and determined the most frequent allele, $R_2$. We then considered that the founder haplotype was Mut-$R_1$-$R_2$. We proceeded in the same way to determine the common haplotype and the possible phase at each marker in each individual up to the first marker where the individual carries alleles that were not shared with other affected individuals. We considered the location of this marker as the position where either a recombination has occurred (in which case it was informative) or a variant has occurred (in which case, it was not informative and we considered it as censored data). The probability of these two events, recombination or variant, depends on the alleles carried by the individual at the discordant marker. Indeed, a stepwise variant model was assumed (with a variant rate per marker per generation of $\mu$), and variants were given a higher probability if the individual carried an allele that was close to the ancestral allele in terms of numbers of repeats. Previous unpublished studies (for more details, you can contact E. Génin, emmanuelle.genin@inserm.fr) found that this reconstruction method provides more reliable age estimates than the one that considers all possible haplotype reconstructions that tended to provide underestimates of allele ages.

*Ancestry determination*. To test the genetic homogeneity between the Chtimi sample and the variant carriers, Principal Component Analysis (PCA) was performed using SmartPCA[15] on the genotypes of 4 patients and 32 controls. The first two principal components were plotted, and carriers and controls were then visually compared. We tested by Mann–Whitney's tests if differences between cases and controls were observed, regarding the first two principal components.

Then, to determine the geographic origin of variant carriers and controls from the Chtimi sample, they were compared with worldwide reference panels using different methods.

First, PCA was performed using SmartPCA[15] on the genotypes of carriers and controls from the Chtimi sample and on the genotypes of the controls from the reference panels at the ancestry-informative SNPs. The first two principal components (PC1 and PC2) were plotted, and individuals from our study were visually compared with individuals from the reference panel. In order to compare the genetic distances between our samples and samples from particular geographic regions within the reference panels, we derived the average Euclidian distances for PC1 from each of our samples to individuals from the regions of interest. Mann–Whitney's tests were used to compare these average distances.

Second, the probability for an individual originating in the different populations represented in the reference control panels was estimated from SNPs' genotype data using the Bayesian method implemented in ADMIXTURE.[16]

*Simulations to assess the power to evidence ancestry differences using the ancestry-informative SNP panel*. Some simulations were performed to determine whether the tests used to compare the four variant carriers against the 32 controls could have been powerful at evidencing a difference if the variant carriers had been sampled in populations from different ancestries. Four individuals were randomly sampled in the different HGDP reference populations from Europe and the Middle-East and compared with the 32 controls. For each reference population, a total of 1000 random draws of four individuals were performed, and the values of the PC1 (respectively, PC2) obtained by the PCA performed on the genotypes at the 45 AIMs that were genotyped on the HGDP populations were compared against the values obtained for the 32 controls using a Mann–Whitney's test with bilateral option. A unilateral Mann–Whitney test was also performed for PC1. The power to reject the null hypothesis that the two samples were similar was then evaluated as the proportion of these draws where the *P*-value of the Mann–Whitney test was $<0.01$ Supplementary Data S4.

## RESULTS

### Estimation of the age of the founder effect of the variant

We were first interested in determining whether this variant could have been introduced in the population by a single ancestor and could thus exhibit a founder effect. If this was the case, we would expect carriers of the variant to share a common haplotype, the length of which would depend on the number of generations elapsed between this ancestor and the variant carriers we sampled. The more the generations, the smaller the expected length of the shared haplotype. Haplotypes of the five patients were reconstructed by parsimony for the 12 microsatellites, and the ancestral haplotype was defined (Table 1).

The time of introduction of this variant into this population was estimated at 9 generations, with a 95% confidence interval ranging

**Table 1 Haplotypes carrying the variant of the five variant carriers, reconstructed by parsimony**

| *Patients* | *D11S4046* | *D11S4146* | *D11S4181* | *Gene β-globin* | *D11S 1338* | *D11S 4149* | *D11S 4116* | *D11S 902* | *D11S 915* | *D11S 914* | *D11S 935* | *D11S 4102* | *D11S 905* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Chtimi33 | 189[a] | **205** | **215** | **Variant** | **261** | **224** | **208** | **154** | 260[a] | 287 | 213 | 150 | 277 |
| | 203 | 197 | 209 | Wild type | 269 | 226 | 210 | 156 | 266 | 287 | 213 | 166 | 293 |
| Chtimi34 | **187** | **205** | **215** | **Variant** | **261** | **224** | 210[a] | **154** | 268 | 283 | 213 | 168 | 285 |
| | 191 | 197 | 217 | Wild type | 263 | 226 | 216 | 160 | 274 | 287 | 215 | 162 | 289 |
| Chtimi35 | **187** | **205** | **215** | **Variant** | **261** | **224** | **208** | **154** | 270[a] | 283 | 213 | 168 | 277 |
| | 197 | 197 | **215** | Wild type | 267 | 230 | 214 | 160 | 268 | 277 | 213 | 150 | 271 |
| Chtimi36 | **187** | **205** | **215** | **Variant** | 267[a] | **224** | 214 | **154** | 270 | 277 | 213 | 162 | 277 |
| | 189 | 197 | 217 | Wild type | 269 | 220 | 216 | 158 | 268 | 277 | 211 | 164 | 275 |
| Chtimi37 | **187** | **205** | 209 | **Variant** | 267[a] | **224** | **208** | **154** | 270 | 283 | 203 | 168 | 279 |
| | 191 | 209 | **215** | Wild type | 269 | 222 | 210 | 162 | 270 | 283 | 211 | 168 | 279 |
| Ancestral haplotype | **187** | **205** | **215** | **Variant** | **261** | **224** | **208** | **154** | – | – | – | – | – |
| Mb[b] | 1.96 | 3.74 | 4.77 | 5.25 | 5.99 | 9.13 | 12.95 | 17.49 | 23.60 | 31.36 | 36.02 | 36.78 | 40.97 |
| Theta[c] | 0.0413 | 0.0189 | 0.0060 | | 0.0093 | 0.0488 | 0.0959 | 0.1496 | 0.2161 | 0.2886 | 0.3251 | 0.3305 | 0.3583 |
| Freq[d] | 0.16 | 0.42 | 0.39 | | 0.31 | 0.44 | 0.16 | 0.31 | – | – | – | – | – |

In bold: the ancestral haplotype was evidenced.
[a]Corresponds to the position on each side of the variant where the haplotype is found divergent from the ancestral haplotype in each individual.
[b]Physical distance in Mb obtained from http://genome.ucsc.edu on Human Feb. 2009 (GRCh37/hg19) Assembly.
[c]Theta is the recombination rate between the variant locus and the markers shown and was estimated using the correspondence 0.7935 cM for 1 Mb over the whole region and Kosambi mapping function (for more details, see Supplementary Data 3A).
[d]Freq is the frequency of the marker allele shared by the patients, which was estimated from controls from the Chtimi panel.

from 4 to 22 generations. Taking 25 years for a generation, the studied variant would have been introduced in the population 225 years ago, with a confidence interval ranging from 100 to 550 years. This result was obtained assuming a stepwise mutation model at each marker with a mutation rate per marker per generation of $10^{-4}$. Increasing the mutation rate to $10^{-3}$ or decreasing it to $10^{-6}$ does not change the result.

### Genetic homogeneity between the Chtimi sample and the variant carriers

The single ancestor origin of the variant being established, we were interested in determining the geographic origin of this ancestor. The first point was to confirm that the patients and Chtimi controls formed together a genetically homogeneous group regarding the 97 SNPs they were genotyped for. We found that the cases did not differ from the controls on the first two PCs of the PCA (Figure 1). The Mann–Whitney's test comparing the top two PC values in cases and controls gave *P*-values of, respectively, 0.75 and 0.39 for the first and second PC.

### Northern European native character of the variant

*Ancestry study for the Chtimi sample and variant carriers at worldwide level.* The Chtimi sample and variant carriers were compared with the different populations of the HapMap3 panel based on the genotypes for the 54 SNPs that were available in both panels Supplementary Data S1. The PCA plot (Figure 2a) showed that the Chtimi sample and variant carriers clustered with the Europeans, and this was confirmed by the ancestry estimation using ADMIXTURE as the Chtimi sample and variant carriers have mainly only a European origin (Figure 2b). Two of the four variant carriers have an estimated posterior probability of European origin of 100%. The remaining two have low posterior probabilities of non-European ancestry, but these levels are not higher than the ones estimated for the Chtimi controls who are known to have their grandparents born in the Nord-

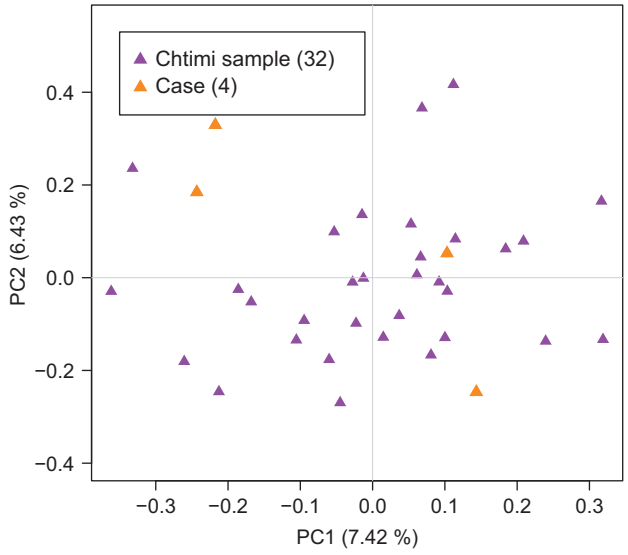


**Figure 1** Top two PCs of the PCA obtained from the genotypes at the 97 ancestry informative SNPs for the 32 Chtimi controls (in purple) and 4 cases (in orange). The percentage of variance explained by each PC is indicated between brackets in the axe legends. The PCA was performed using SmartPCA[11] with the default options.

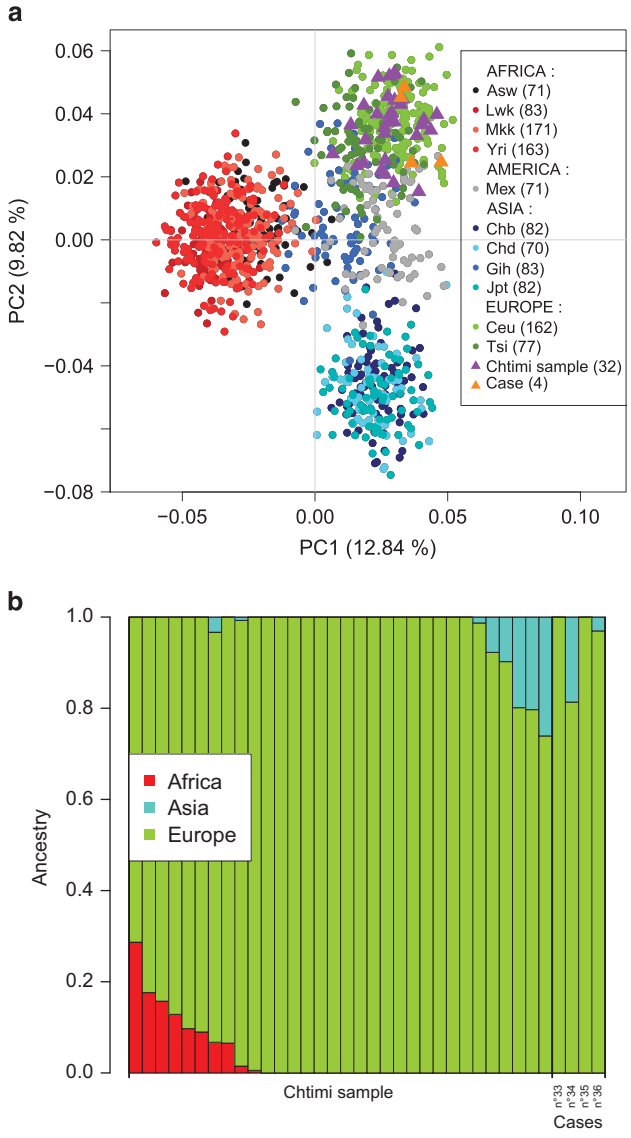


**Figure 2** Determination of the geographic origin of the Chtimi sample and variant carriers based on 54 of the ancestry informative SNPs available in the HapMap3 panel. (**a**) The top two Principal Components (PC1 and PC2) of the PCA represent 12.84 and 9.82% of the genetic variance. Each plot represents one individual: case, control from the Chtimi sample, or from the HapMap3 panel. (**b**) Ancestry estimation for the Chtimi sample and variant carriers using ADMIXTURE with three possible ancestry clusters: Africa, Asia, and Europe. Each column represents one individual.

Pas-de-Calais region. It could indicate either that the two patients are admixed and have some of parts of their genome of Asian ancestry or it could be due to uncertainties in the measures of posterior probability. The second explanation is more likely than the first one as we have run ADMIXTURE on a limited number of markers. If more markers spanning the whole genome were available, the estimates might have been more precise, and interestingly, it might have been possible to determine whether the studied variant falls in the inferred European or non-European part of the genome of these patients. However, given the fact that a founder effect was evidenced for the variant, with all the carriers having inherited the variant from a common ancestor, it is difficult to envisage a scenario where the variant could be in the non-European part of the genome of two out the four variant carriers.

A similar comparison was done with the HGDP Panel as this panel included more diverse populations than the HapMap3 panel. Genotypes were available in the HGDP panel for a total of 45 of the ancestry-informative SNP panel. The PCA plot Supplementary Data S5A was also consistent with an European ancestry of the four variant carriers, and the Bayesian analysis Supplementary Data S5B provided posterior probability estimates of European ancestry of 100% for two of them and >80% for the remaining two. Interestingly, the patients with some levels on non-European ancestry were not the same patients as the ones who showed some levels of non-European ancestry in the Bayesian analysis using HapMap3 reference panel. These discrepancies were probably due to uncertainties in the posterior probability estimates rather than indications of admixture.

These different analyses at worldwide level were consistent with the hypothesis that the variant carriers have a European origin and that the variant appeared in Europe and not in Africa or Asia where several variants involved in thalassemia are frequent.

*Ancestry study for the Chtimi sample and variant carriers at the European continental scale.* To gain further insights into the origin of the variant and particularly to determine whether it most likely originated in the North or South of Europe, we performed some additional analyses with the reference panels restricted to their European or Middle-East populations. In the HGDP panel, we selected populations from the Middle-East and Europe in order to estimate the part of Mediterranean origin in the Chtimi sample and variant carriers and to put the Chtimi sample and variant carriers on a North/South gradient. In this panel, the same 45 ancestry informative markers as the ones used in the worldwide analysis were available. We performed a PCA (Figure 3a), and it appeared that the Chtimi sample and variant carriers were closer to the European populations than to the Middle-Eastern ones (*P*-value of Mann–Whitney's test $= 5.70 \times 10^{-12}$ for PC1 distances). A distance on PC1 was computed between each variant carrier and each control from HGDP populations or from Chtimi sample (Figure 3b); these distances were smaller when considering the European and Chtimi samples than when considering Middle-Eastern populations, indicating a more likely European than Middle-Eastern origin of variant carriers. To infer a more precise European origin, we considered two populations from HapMap3 panel: CEU from Northern Europe and TSI from Tuscan, Southern Europe, and their genotypes at 62 of the ancestry-informative SNPs. We also realized a PCA: the PC1 only explained 3.84% of the variance but the Chtimi sample and variant carriers clustered with CEU individuals in a distinct group from TSI individuals. Indeed, for PC1 distances, Chtimi controls and variant carriers were closer to the HapMap3 CEU than to the HapMap3 TSI (*P*-value of Mann–Whitney's test $= 1.15 \times 10^{-5}$), indicating a more likely Northern than Southern origin in Europe, which was confirmed on the HGDP panel restricted to European population. Indeed, when classifying the European populations from HGDP into three groups: North (represented by the Orcadian population), Central (French Basque, French, North Italian and Tuscan populations) and South Europe (Sardinian population), we found that the Chtimi sample clustered with Northern and Central Europeans and the cases mainly with the Northern Europeans Supplementary Data S6. These different results argued in favor of the hypothesis that the studied variant was native to Northern Europe and was not introduced by recent migrations from other regions of the world.

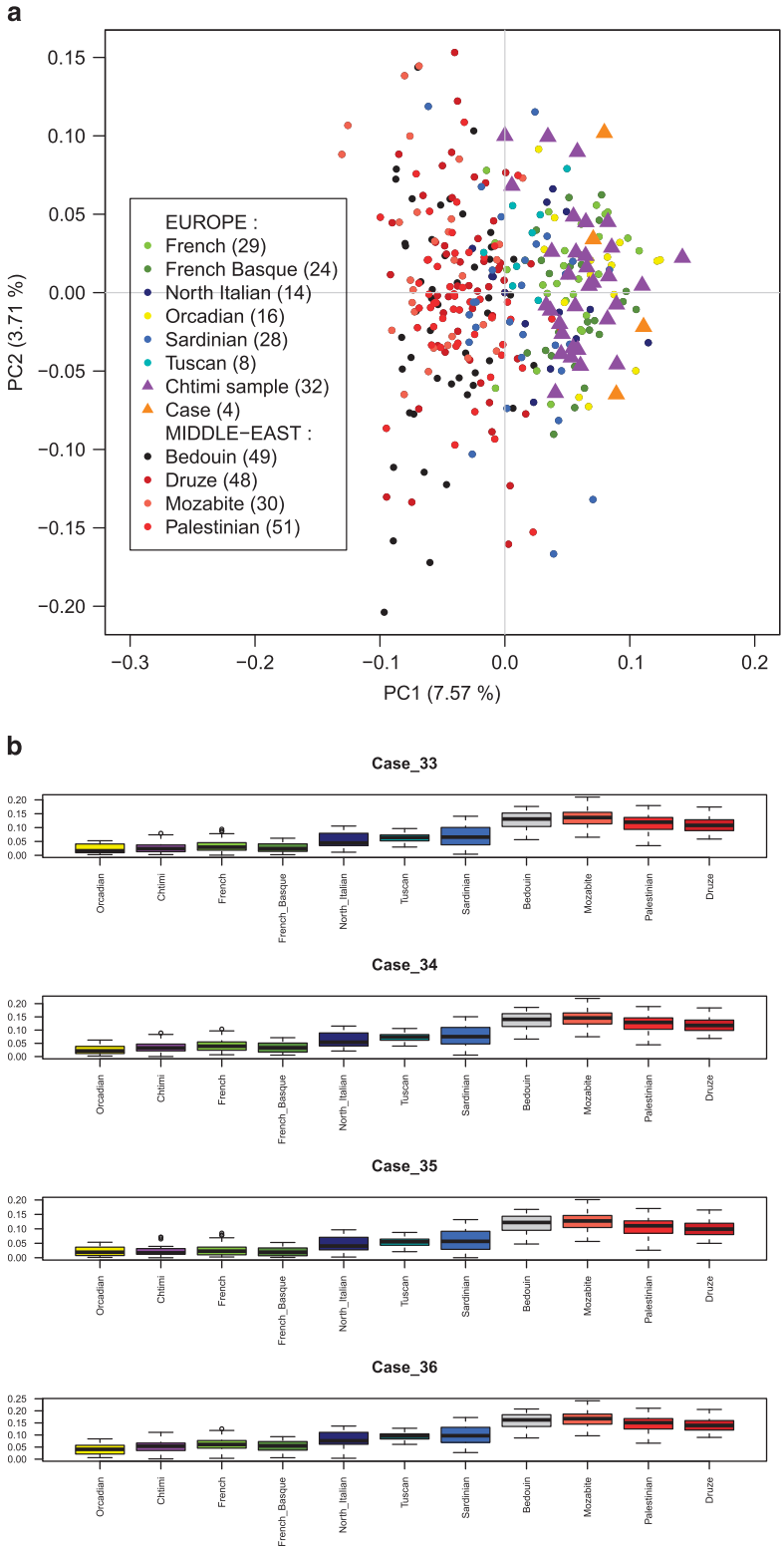


**Figure 3** Determination of the geographic origin of the Chtimi sample and variant carriers based on 45 of the ancestry informative SNPs available in the HGDP Panel. Only European and Middle-Eastern populations from HGDP Panel are considered. (**a**) The top two Principal Components (PC1 and PC2) of the PCA represent 7.57 and 3.71% of the genetic variance. Each plot represents one individual: case, or control from Chtimi sample or from the HGDP European and Middle-Eastern populations. (**b**) PC1 distances calculated between each variant carrier and each control from HGDP European and Middle-Eastern populations or Chtimi sample.

**Simulations to assess the power to evidence ancestry differences using the ancestry-informative SNP panel**

As described in the Methods section, some simulations were performed to determine whether the tests used to compare the 4 variant carriers against the 32 controls could have been powerful at evidencing a difference in the PCs values if the variant carriers had been sampled in populations from different ancestries. Based on Figure 3a, it seemed that variant carriers and Chtimi sample were on the extremity of the North–South PC1 axis, so we also performed for PC1 a Mann–Whitney's test against an unilateral null hypothesis testing whether the 4 sampled individuals have a lower PC1 value than the 32 Chtimi controls. Based on the PC1 values, the power (unilateral and bilateral) was 100% to discriminate between the 32 controls and 4 individuals from any of the Middle-East populations Supplementary Data S4. Power was more reduced when considering populations within Europe but clearly decreases from South to North. PC2, however, was not informative on the North–South ancestry. These simulation results showed that we could be confident that the four variant carriers were not recent migrants from Middle-East populations. We could not reject the possibility that they had a South European origin as the power to reject a Sardinian origin is only 53.2%. However, they show genetically closer relationships to the populations from Northern Europe and the Chtimi controls than to the Southern European populations (Figure 3a and b).

## DISCUSSION

In this study, we were interested in studying the origin of the duplication of four nucleotides in the first exon of β-globin gene that was previously described in several families of patients originating from Nord-Pas-de-Calais (France). Using several microsatellite markers surrounding the β-globin gene, we were able to show that all the five studied variant carriers shared a common haplotype. This is the signature of a founder effect and a piece of evidence that all the variant carriers have inherited the variant from a common ancestor who introduced it in the population about 225 years ago.

In order to determine whether the common ancestor could have been a recent migrant from a region where thalassemia is endemic, we used a panel of 97 European Ancestry Informative markers (EURO-AIMS) described by Price *et al*[11] to compare the origins of the variant carriers with those of controls whose four grandparents were born in the Nord-Pas-de-Calais region in the North of France and with publicly available data on controls of known origin from two worldwide reference control panels. We found that the genotypes at these EURO-AIMs of the variant carriers were genetically closer to individuals from Northern Europe than to individuals from Southern Europe or the Middle-East (Figure 3b) and that they were not different from the 32 controls whose grandparents were born in Nord-Pas-de-Calais. Although there have been several reports of variants of the β-globin genes found in patients living in the North of Europe and with no known ancestry in regions where thalassemia is endemic,[7–9] this is the first time a genetic study using ancestry-informative markers has been conducted to confirm the origin of patients.

To more finely trace the origin of the variants within Europe, it would be interesting to compare the patients with other reference panels with more European populations (Northern European such as Swedish, German, Belgian, British and Southern European such as Spanish or Portuguese), but it is more difficult to have access to these populations as most of the panels that were used to study the genetic diversity in Europe, such as the Europa panel[17] are not publicly available.

The variant was transmitted to all the five studied variant carriers by a same common ancestor who lived about 225 years ago, that is, in the second half of the eighteenth century, with a large range due to the small sample size. It is possible that the variant appeared earlier in the population but remains at very low frequency. Indeed, the method we used to estimate variant age provides an estimate of the time since the most recent common ancestor of variant carriers and not the time since the variant first appeared. It is possible that the mutation leading to the studied variant took place earlier and the estimate of nine generations ago is in fact the age estimate of a population bottleneck that occurred in the population. This bottleneck would then have created a genetic pattern that makes it look as though the mutation was first introduced into the population nine generations ago. Interestingly, there is evidence that for many centuries and up to 1900, malaria secondary to different species of *Plasmodium* (*vivax*, *falciparum*) was present all over Europe and Africa from Scandinavia to South Africa.[18] There are several reports of β-thalassemia variants in patients from Northern Europe with no known ancestry in regions where thalassemia is endemic,[7–9] but so far it cannot be ruled out that these variants might have been introduced by recent migrations from regions where β-thalassemia variants are frequent. There are some reports showing that, in the Nord-Pas-de-Calais region, swamps were present until the end of the nineteenth century, and malaria was prevalent.[19] As heterozygosity for β-thalassaemia is thought to provide some protection against death from malaria, it is possible that about 200 years ago a particularly severe malaria epidemic took place and gave a selective advantage to heterozygous variant carriers. However, given the short timescale (nine generations) and the fact that the speculated selective pressure is no longer present in the population, it is very difficult to draw any conclusion regarding the possible role of selection in the spread of this variant.

Moreover, the Northern extension of the area where malaria was present in the past is not well known, and thus the question of the relevance of determining the ethnic origin of the mutation carriers to study the relationship between malaria and β-globin could be raised. Although there have been several reports of variants of the β-globin genes found in patients living in the North of Europe and with no known ancestry in regions where thalassemia is endemic, to our knowledge, ancestry-informative markers spanning different genomic regions have not been used so far to confirm the origin of patients. This is, to our knowledge, the first work where the ethnic origin of mutation carriers is studied using ancestry-informative markers, and we hope it could open up the way to other similar studies that may collectively help resolve the question of the relationship between malaria and β-globin. In particular, studies with ancestry-informative markers on erythrocyte genetic disorders in geographical areas where malaria exerts a more significant selective pressure could really be helpful in the debate.

## CONFLICT OF INTEREST

The authors declare no conflict of interest.

## ACKNOWLEDGEMENTS

We thank F Ariey for his reviewing the manuscript; C Rose's mentor Professor F Bauters, for his large vision of the field of hematology and his concern for its development in his native area (Nord-Pas-de-Calais); and all volunteer controls.

1 Higgs DR, Engel JD, Stamatoyannopoulos G: Thalassaemia. *Lancet* 2012; **379**: 373–383.
2 Giardine B, van Baal S, Kaimakis P *et al*: HbVar database of human hemoglobin variants and thalassemia mutations: 2007 update. *Hum Mutat* 2007; **28**: 206.
3 Taylor SM, Parobek CM, Fairhurst RM: Haemoglobinopathies and the clinical epidemiology of malaria: a systematic review and meta-analysis. *Lancet Infect Dis* 2012; **12**: 457–468.
4 Haldane JB: Disease and evolution. *Ric Sci Suppl* 1949; **36**: 68–76.
5 Fowkes FJ, Allen SJ, Allen A, Alpers MP, Weatherall DJ, Day KP: Increased microerythrocyte count in homozygous alpha(+)-thalassaemia contributes to protection against severe malarial anaemia. *PLoS Med* 2008; **5**: e56.
6 Weatherall DJ: Genetic variation and susceptibility to infection: the red cell and malaria. *Br J Haematol* 2008; **141**: 276–286.
7 Cai SP, Eng B, Francombe WH *et al*: Two novel beta-thalassemia mutations in the 5' and 3' noncoding regions of the beta-globin gene. *Blood* 1992; **79**: 1342–1346.
8 Jamet D, Pissard S, Blouch MT, Berthou C, De Braekeleer M, Abgrall JF: Beta-thalassemia in the indigenous population of Brittany: identification of three rare mutations. *Haematologica* 2006; **91**: 1418–1419.
9 Vetter B, Schwarz C, Kohne E, Kulozik AE: Beta-thalassaemia in the immigrant and non-immigrant German populations. *Br J Haematol* 1997; **97**: 266–272.
10 Georgel AF, Mereau C, Willekens C *et al*: Identification of a new mutation on the beta-globin gene: codons 8/9 (+AGAA); GAG.AAG.TCT(Glu-Lys-Ser)>GAG. AAAGAAG, in a patient from the north of France with a phenotype of beta-thalassemia minor. *Hemoglobin* 2010; **34**: 389–393.
11 Price AL, Butler J, Patterson N *et al*: Discerning the ancestry of European Americans in genetic association studies. *PLoS Genet* 2008; **4**: e236.
12 International HapMap C, Altshuler DM, Gibbs RA *et al*: Integrating common and rare genetic variation in diverse human populations. *Nature* 2010; **467**: 52–58.
13 Rosenberg NA, Pritchard JK, Weber JL *et al*: Genetic structure of human populations. *Science* 2002; **298**: 2381–2385.
14 Genin E, Tullio-Pelet A, Begeot F, Lyonnet S, Abel L: Estimating the age of rare disease mutations: the example of Triple-A syndrome. *J Med Genet* 2004; **41**: 445–449.
15 Patterson N, Price AL, Reich D: Population structure and eigenanalysis. *PLoS Genet* 2006; **2**: e190.
16 Alexander DH, Novembre J, Lange K: Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* 2009; **19**: 1655–1664.
17 Heath SC, Gut IG, Brennan P *et al*: Investigation of the fine structure of European populations with applications to disease association studies. *Eur J Hum Genet* 2008; **16**: 1413–1429.
18 Hay SI, Guerra CA, Tatem AJ, Noor AM, Snow RW: The global distribution and population at risk of malaria: past, present, and future. *Lancet Infect Dis* 2004; **4**: 327–336.
19 Pierrard P: Histoire du Nord. *Hachette* 1978; pp 81–82.

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