

ORIGINAL ARTICLE

The role of human demographic history in determining the distribution and frequency of transferase-deficient galactosaemia mutations

JM Flanagan^{1,2}, G McMahon², SH Brendan Chia², P Fitzpatrick², O Tighe², C O'Neill¹, P Briones³, L Gort³, L Kozak⁴, A Magee⁵, E Naughten¹, B Radomyska⁶, M Schwartz⁷, JS Shin⁸, WM Strobl⁹, LA Tyfield¹⁰, HR Waterham¹¹, H Russell¹², G Bertorelle¹³, JKV Reichardt¹⁴, PD Mayne¹ and DT Croke²

¹Department of Pathology, The Children's University Hospital, Dublin, Ireland; ²Molecular & Cellular Therapeutics, The Royal College of Surgeons in Ireland, Dublin, Ireland; ³Institut de Bioquímica Clínica, Hospital Clínic and CIBERER, Barcelona, Spain; ⁴Center of Molecular Biology & Gene Therapy, University Hospital Brno, Brno, Czech Republic; ⁵Northern Ireland Regional Genetics Service, Belfast, UK; ⁶Department of Pediatrics, Institute of Mother and Child, Warsaw, Poland; ⁷Department of Clinical Genetics, University Hospital Rigshospitalet, Copenhagen, Denmark; ⁸Children's Hospital, University of Munich, Munich, Germany; ⁹Center for Physiology and Pathophysiology, Medical University of Vienna, Vienna, Austria; ¹⁰Southmead Hospital and University of Bristol, UK; ¹¹Academic Medical Centre, University of Amsterdam, The Netherlands; ¹²Centre for Cancer Research and Cell Biology, Queen's University, Belfast, UK; ¹³Dipartimento di Biologia, Università di Ferrara, Ferrara, Italy and ¹⁴University of Sydney, New South Wales, Australia

Classical or transferase-deficient galactosaemia is an inherited metabolic disorder caused by mutation in the human *Galactose-1-phosphate uridylyl transferase* (*GALT*) gene. Of some 170 causative mutations reported, fewer than 10% are observed in more than one geographic region or ethnic group. To better understand the population history of the common *GALT* mutations, we have established a haplotyping system for the *GALT* locus incorporating eight single nucleotide polymorphisms and three short tandem repeat markers. We analysed haplotypes associated with the three most frequent *GALT* gene mutations, Q188R, K285N and Duarte-2 (D2), and estimated their age. Haplotype diversity, in conjunction with measures of genetic diversity and of linkage disequilibrium,

indicated that Q188R and K285N are European mutations. The Q188R mutation arose in central Europe within the last 20 000 years, with its observed east–west cline of increasing relative allele frequency possibly being due to population expansion during the re-colonization of Europe by *Homo sapiens* in the Mesolithic age. K285N was found to be a younger mutation that originated in Eastern Europe and is probably more geographically restricted as it arose after all major European population expansions. The D2 variant was found to be an ancient mutation that originated before the expansion of *Homo sapiens* out of Africa.

Heredity (2010) **104**, 148–154; doi:10.1038/hdy.2009.84; published online 29 July 2009

Keywords: galactosaemia; Q188R; K285N; Duarte-2; haplotype; population genetics

Introduction

Galactosaemia is an autosomal recessive disorder of carbohydrate metabolism that can be caused by a defect in any one of the three enzymes involved in the metabolism of galactose through the Leloir pathway. The most common cause of galactosaemia, however, is deficiency of the galactose-1-phosphate uridylyl transferase (*GALT*) enzyme; this is sometimes termed as transferase-deficient or classical galactosaemia (On-line Mendelian Inheritance In Man no. 230400). The disease usually presents in the neonatal period with symptoms including vomiting, diarrhoea, failure to thrive, jaundice and *Escherichia coli* sepsis. The incidence of transferase-deficient galactosaemia varies worldwide, ranging from

1 in 23 000–40 000 in Europe (Murphy *et al.*, 1999) and 1 in 50 000 in the USA (Suzuki *et al.*, 2001) to 1 in 1 000 000 in Japan (Hirokawa *et al.*, 1999).

The human *GALT* gene maps to chromosome 9p13 and more than 170 causative mutations in transferase-deficient galactosaemia have been described (ARUP Laboratories *GALT* Mutation Database), fewer than 10% of which are found in more than one geographic region or ethnic group (Tyfield *et al.*, 1999). The most common mutations are Q188R (c. 563A>G), K285N (c. 855G>T), L195P (c. 584T>C), and S135L (c. 404C>T). Q188R is the predominant mutation in galactosemic patients of European descent and accounts for approximately 64% of all *GALT* mutations. The mutation reaches its highest frequency in Ireland (93.6%) (Murphy *et al.*, 1999) and shows a gradient of declining frequency through Europe in a south-easterly direction (Tyfield *et al.*, 1999). K285N is the second most common European mutation, especially in the countries of central and eastern Europe where it can account for up to 34% of galactosaemia alleles (Greber-Platzer *et al.*, 1997; Kozak *et al.*, 2000).

Correspondence: Dr JM Flanagan, Department of Hematology, St Jude Children's Research Hospital, 262 Danny Thomas Place, D-2057, Mail Stop 310, Memphis, TN 38103, USA.

E-mail: Jonathan.Flanagan@stjude.org

Received 11 February 2009; revised 21 April 2009; accepted 21 May 2009; published online 29 July 2009

Both of these mutations are rare in galactosaemic patients of non-European ancestry (Ng *et al.*, 1994; Ashino *et al.*, 1995).

Another common *GALT* nucleotide substitution, N314D (c. 940A>G), is associated with two variants of galactosaemia that have altered transferase activity but have subclinical phenotypes. These variants are Duarte-1 (Los Angeles, D1) and Duarte-2 (D2). The D1 variant is associated with increased transferase activity (110–130% of normal) whereas the D2 variant is associated with decreased enzyme activity. For D1 alleles, N314D is in linkage disequilibrium with the L218L silent substitution (c. 652C>T) (Reichardt *et al.*, 1991; Elsas *et al.*, 1994). In the case of D2 alleles, the N314D polymorphism is in linkage disequilibrium with three intronic substitutions (IVS4nt-27g>c, IVS5nt-24g>a, IVS5nt+62g>a) and a 4 bp deletion in the *GALT* promoter (Kozak *et al.*, 1999). This 5'UTR-119delGTCA deletion has been identified as the cause of reduced transferase activity in D2, due to its effect on the *GALT* promoter (Elsas *et al.*, 2001; Trbusek *et al.*, 2001), and it has been shown more recently to reduce *GALT* mRNA abundance (Carney *et al.*, 2009). The overall frequency of N314D worldwide is approximately 10%, with some variation in the ratio of D2/D1 allele frequencies (Suzuki *et al.*, 2001; Tighe *et al.*, 2004). The N314D polymorphism has been found in all populations screened to-date and in association with the same linked polymorphisms for D1 and D2, suggesting that the Duarte variants are the result of ancient mutations.

Although some inferences can be made about the history of *GALT* mutations based upon frequency data, a more complete description of their population history depends on determining the place and time of origin of the mutations and an explanation of their presence, frequency and geographic distribution. To perform this task, we have identified flanking polymorphic markers that have allowed the reconstruction of haplotypes associated with particular *GALT* alleles. Using haplotype analysis, we have now defined the origin and age of the three most common galactosaemia mutations (Q188R, K285N and D2) in a large sample of galactosaemia patients from 10 European population groups.

Methods

DNA samples from patients and controls

DNA samples ($n = 320$) were obtained from galactosaemia patients and heterozygotes of known genotype from 10 European population groups with appropriate local ethical approval through the collaborating centers: Republic of Ireland ($n = 62$), Northern Ireland ($n = 14$), United Kingdom ($n = 38$), Denmark ($n = 23$), The Netherlands ($n = 27$), Germany ($n = 41$), Austria ($n = 58$), Poland ($n = 31$), Czech Republic ($n = 12$) and Spain ($n = 14$). These comprised Q188R alleles (65.9%; $n = 422$), K285N (9.2%; $n = 59$), Duarte-2 (D2), (9.4%; $n = 60$), low-frequency mutations (6.9%; $n = 44$) and wild-type alleles (8.6%; $n = 55$). Control DNA samples ($n = 60$) from individuals of Irish ethnicity were obtained with informed consent and local ethical approval; these were used to analyse the extent of polymorphism of candidate SNP and STR markers. In addition, a set of anonymized singleton DNA samples from 15 European populations ($n = 117$) was used to assess the level of wild-type *GALT*

haplotype diversity; these were obtained previously in the context of a study of *PAH* gene VNTR sequence diversity (Tighe *et al.*, 2003) and included: Republic of Ireland ($n = 72$), Northern Ireland ($n = 10$), Scotland ($n = 8$), Spain ($n = 5$), Germany ($n = 7$), Denmark ($n = 4$) and Poland ($n = 11$).

Identification of polymorphic markers

A largely bioinformatics-driven approach was used to identify polymorphic single nucleotide polymorphism (SNP) and short tandem repeat (STR) markers within an approximately 230 kbp interval flanking the galactose-1-phosphate uridylyl transferase locus. The SNP Consortium Database (<http://www.hapmap.org/>) was used to search for SNPs that lay within the *GALT* interval. Ten candidate SNPs were identified, of which six were found to be polymorphic (having a minor allele frequency of 5% or greater) by genotyping in a panel of 60 control Irish DNA samples: rs1799729 [GC/TT], rs1104748 [C/T], rs2252539 [A/T], rs2249561 [C/T], rs1411207 [C/T] and rs1566591 [C/T]. One STR marker, AFM326vd1, was identified within the *GALT* interval by searching the ENSEMBL database (<http://www.ensembl.org/>). To identify additional STRs, three genomic sequence contigs spanning the *GALT* interval (AL162231, AL450283 and AL160270) were downloaded from ENSEMBL and candidate dinucleotide STRs were identified by simple repeat text-string searches. This process yielded nine candidate STRs, which were assessed for polymorphism by genotyping in the same control DNA panel. Three of nine (STRs J5, J7 and J8) were found to be polymorphic and were fully characterized by sequencing (data not shown); J7 was found to be AFM326vd1, whereas J5 and J8 were novel STR markers. The oligonucleotide primers used for genotyping are given in the Supplementary Table S1.

Genotype acquisition and haplotype reconstruction

Patient and control DNA samples were genotyped for the eight SNP and three STR markers by standard laboratory methods, typing the SNPs in PCR-RFLP assays and the STRs by PCR amplification using fluorescently labelled primers and allele sizing on an ALFexpress automated DNA sequencer (Amersham Pharmacia Biotech, Little Chalfont, UK). SNP genotypes were called by visual inspection of electrophoresis gels, whereas STR genotypes were determined using the AlleleLinks software package (Amersham Pharmacia Biotech). Haplotypes were reconstructed from genotypic data by means of the PHASE 2.1 software (Stephens *et al.*, 2001; Stephens and Donnelly, 2003); optimal run length was determined empirically according to the recommendations of Stephens and Donnelly (2003).

Statistical methods

Measures of genetic diversity and linkage disequilibrium were estimated using the Arlequin (Excoffier *et al.*, 2005) and DnaSP (Rozas and Rozas, 1999) software packages. Median-Joining (MJ) networks were constructed using the Network v4.112 software (Bandelt *et al.*, 1999). The ages of the most common mutations were estimated using two different methods. The first technique used is based on a maximum-likelihood method (Slatkin and Rannala, 1997) that assumes the frequency of the

mutation and the genetic variation in the sample of affected chromosomes (related to the coalescence times of the intra-allelic gene genealogy) are expected to increase with time since the mutation occurred. This method jointly analyses these two pieces of information using a stochastic birth–death process. The second method is based on the $(\delta\mu)^2$ distance at microsatellite markers between all chromosomes carrying a specific mutation, with the expectation that $(\delta\mu)^2$ will increase in an approximately linear fashion with time (Goldstein *et al.*, 1995).

Results

Development of a *GALT* Haplotype

The purpose of this work was to develop a set of SNP and STR markers for the construction of haplotypes in transferase-deficient galactosaemia. We used *in silico* analyses to identify six polymorphic SNPs in the region flanking the *GALT* gene: rs1799729 [GC/TT], rs1104748 [C/T], rs2252539 [A/T], rs2249561 [C/T], rs1411207 [C/T] and rs1566591 [C/T]. We combined these SNPs with the two previously characterized intragenic *GALT* SNPs, IVS5nt-24g>a (*Sac* I) and c.652c>t (L218L), to form the basis of the haplotype framework. We further identified three dinucleotide STR markers, J5, AFM326vd1 (J7) and J8, through a combination of bioinformatics and genetic screening. We combined the eight SNP and three STR markers to produce the *GALT* compound haplotype (Figure 1). The compound haplotype flanks the 4.2 kbp *GALT* locus on the short arm of chromosome 9 (nt. 34 636 634–34 640 570), covering an approximately 232 kbp interval (nt. 34 561 450–34 793 688) (Figure 1). The eight SNP markers that constitute the backbone of the haplotype cover a span of approximately 166 kbp (nt. 34 627 990–34 793 688). We analysed this chromosomal

region using the HapMap data (www.hapmap.org; October 2005 release) and our analysis indicated the presence of a possible recombination ‘hot-spot’ between SNP markers rs2252539 and rs2249561 in the approximate 34 688 000–34 695 000 interval (Figure 1). However, when we examined analyses of pairwise linkage disequilibrium between SNPs across the interval, the ‘hot-spot’ had negligible effects on the haplotypes observed (data not shown). We genotyped patient and control DNA samples for all 11 polymorphic markers and we constructed haplotypes using the PHASE 2.1 software. A total of 179 compound haplotypes were found in 275 control chromosomes (all haplotype definitions are listed in the Supplementary Table S2).

Transferase-deficient galactosaemia haplotypes

Analysis of the Q188R alleles determined that there were 34 compound haplotypes associated with them, of which G1A [01011100 (AC)₁₄ (CA)₁₄ (GT)₁₀] was the most common, occurring at an overall frequency of 0.823 (347 out of 422 alleles). Within the Irish samples, 32 out of 80 Q188R alleles (40%) were of Irish traveller ethnicity. The travellers are a distinct ethnic group in Ireland and there is some debate as to their origins (Murphy *et al.*, 1999). The G1A compound haplotype was the prevalent haplotype in both populations (~100% frequency). The MJ network of Q188R compound haplotypes (Figure 2a) showed a relatively extended network centred on the majority, presumably ancestral, G1A haplotype. In contrast, we observed that the MJ network of wild-type *GALT* chromosomes (Figure 2d) was greatly extended and contained 179 haplotypes, the most common of which [G2D; 11010100 (AC)₁₅ (CA)₁₈ (GT)₁₀] occurred at a frequency of 0.053. We found that G1A, the putative ancestral Q188R haplotype, occurred at a frequency of 0.018 in all of the wild-type samples. Estimates of genetic diversity (*h*) for SNP-only haplotypes revealed

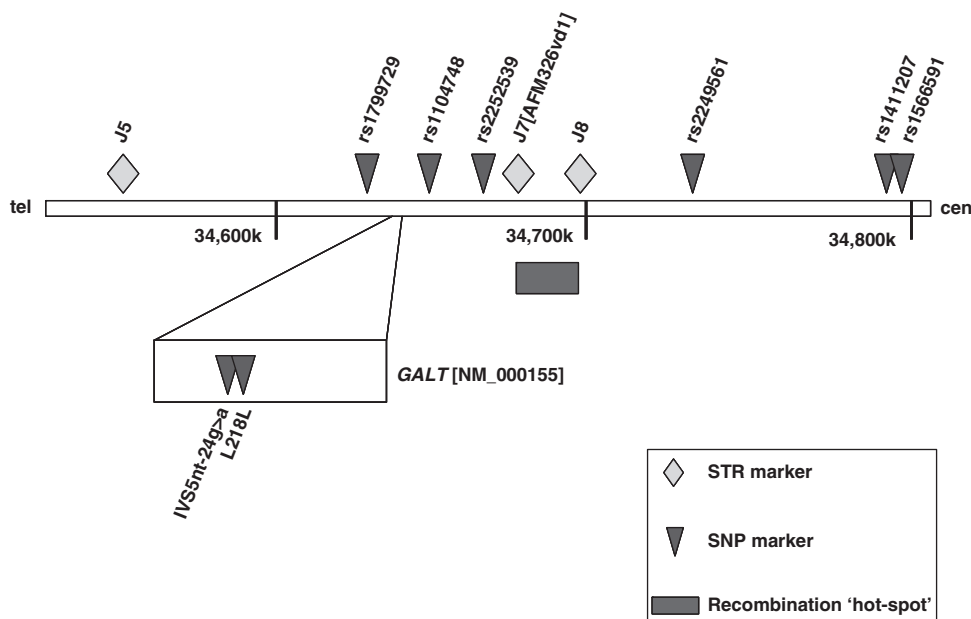


Figure 1 Map positions of the SNP and STR markers constituting the *GALT* haplotype on chromosome 9p13. Marker positions are given relative to the genomic sequence indicated as nucleotide numbers below the horizontal bar representing the chromosome; ‘cen’ and ‘tel’ indicate centromeric and telomeric orientation of the chromosome interval.

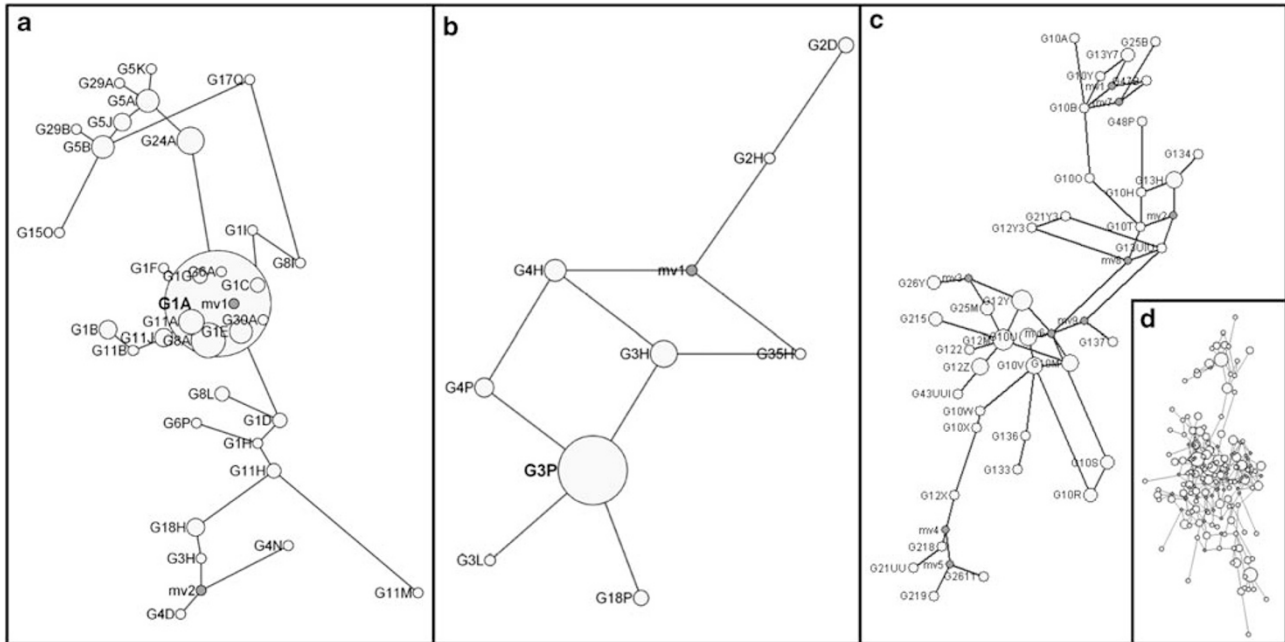


Figure 2 Median-joining networks of *GALT* haplotypes associated with Q188R (a), K285N (b), D2 (c) and wild-type (d) alleles.

Table 1 Geographic variation in SNP haplotype genetic diversity (*h*) for Q188R, K285N, D2 and wild-type alleles in Europe

Sample	Q188R	K285N	Duarte D2	Wild-type
Ireland (Republic)	0.205 ± 0.057 (<i>n</i> = 80)	ND (<i>n</i> = 1)	1.000 ± 0.177 (<i>n</i> = 14)	0.854 ± 0.012 (<i>n</i> = 166)
Northern Ireland	0.100 ± 0.088 (<i>n</i> = 20)	ND (<i>n</i> = 1)	ND	0.881 ± 0.044 (<i>n</i> = 20)
United Kingdom	0.035 ± 0.034 (<i>n</i> = 57)	0.000 ± 0.000 (<i>n</i> = 9)	ND	ND
Netherlands	0.046 ± 0.043 (<i>n</i> = 44)	ND	0.700 ± 0.218 (<i>n</i> = 5)	ND (<i>n</i> = 3)
Denmark	0.139 ± 0.069 (<i>n</i> = 41)	ND	0.000 ± 0.000 (<i>n</i> = 3)	ND (<i>n</i> = 9)
Spain	0.255 ± 0.116 (<i>n</i> = 22)	0.000 ± 0.000 (<i>n</i> = 2)	ND	0.891 ± 0.063 (<i>n</i> = 11)
Germany	0.625 ± 0.072 (<i>n</i> = 49)	ND	0.831 ± 0.056 (<i>n</i> = 20)	0.868 ± 0.076 (<i>n</i> = 29)
Austria	0.438 ± 0.082 (<i>n</i> = 50)	0.684 ± 0.076 (<i>n</i> = 22)	0.667 ± 0.113 (<i>n</i> = 16)	ND (<i>n</i> = 15)
Poland	0.337 ± 0.091 (<i>n</i> = 43)	0.125 ± 0.106 (<i>n</i> = 16)	0.000 ± 0.000 (<i>n</i> = 2)	0.844 ± 0.062 (<i>n</i> = 22)
Czech Republic	0.000 ± 0.000 (<i>n</i> = 16)	0.000 ± 0.000 (<i>n</i> = 8)	ND	ND
Overall	0.258 ± 0.028 (<i>n</i> = 422)	0.381 ± 0.076 (<i>n</i> = 59)	0.793 ± 0.030 (<i>n</i> = 60)	0.895 ± 0.011 (<i>n</i> = 275)

Values of *h* are given as mean ± s.d. The number of chromosomes (*n*) included in the analysis for each allele type and population sample is given in brackets. ND, not determined; either allele not detected or no regional samples available.

substantial differences in the level of diversity associated with Q188R between the regional populations sampled (Table 1). We observed that populations on the ‘Atlantic façade’ of Europe exhibited *h* values lower than those in central/eastern Europe; for example, *h* values of 0.205 ± 0.057 in Ireland as compared with 0.625 ± 0.072 in Germany. Haplotype genetic diversity data for wild-type chromosomes showed similar levels of diversity across all populations sampled, with an overall *h* = 0.895 ± 0.011 (Table 1).

The second most common *GALT* mutation in Europe, K285N, was associated with nine compound haplotypes, of which G3P [11010011 (AC)₁₅ (CA)₁₈ (GT)₁₁] was the majority haplotype occurring at a frequency of 0.678 (40 out of 59 alleles). The MJ network of K285N haplotypes was minimally extended compared with that of Q188R and, particularly, to that of wild-type chromosomes (Figure 2). SNP-only haplotypes exhibited an overall genetic diversity of 0.381 ± 0.076.

Examination of the D2 variant chromosomes showed that the D2 compound haplotypes produced an extended MJ network that lacked an obvious ancestral node

(Figure 2c); six of the 30 haplotypes in the network [G10A, G10B, G10C, G12B, G12C, G13D] had frequencies of 0.051 or greater with all others occurring at lower frequencies. An examination of SNP-only haplotypes associated with D2 (data not shown) identified haplotype G10 [10011100] as the putative ancestral SNP haplotype. We found that D2 SNP-only haplotypes had an *h* value of 0.793 ± 0.030 overall (Table 1), a value closer to that determined for wild-type chromosomes than *h* values for either Q188R or K285N chromosomes.

GALT mutation age estimation

The differences in haplotype network structure and SNP-only haplotype diversity between Q188R, K285N, D2 and wild-type chromosomes were suggestive of different ages for the mutant alleles. To investigate this further, we used the δ parameter of Risch *et al.* (1995) to estimate the degree of linkage disequilibrium (LD) across the interval spanned by the *GALT* compound haplotype for Q188R, K285N and D2 (Figure 3). For each mutation, the δ parameter calculates the proportion of mutant chromo-

some still bearing the ancestral associated allele. In the case of all three mutant alleles, we observed the highest values of δ in the interval containing the *GALT* locus with values declining to lesser extents outside. Q188R and K285N showed similar extents of LD within and flanking this interval whereas δ values for D2 fell markedly on the centromeric and telomeric sides of the interval. This suggests that D2 is considerably older than either Q188R or K285N.

For a more accurate mutation age estimation, the $(\delta\mu)^2$ distance method of Goldstein *et al.* (1995) was performed using three different STR mutation rates (1×10^{-4} , 5×10^{-4} , 1×10^{-3}). Of these, the 5×10^{-4} mutation rate is likely to approximate most closely to the average mutation rate for autosomal STR loci (Cooper *et al.*, 1999) with the other values serving as upper and lower bounds. The $(\delta\mu)^2$ values yielded age estimates for Q188R, K285N and D2 of 19 467, 4617 and 246 051 years, respectively (Table 2). As an independent measure of allele ages based upon STR diversity, allele ages were computed also using the intra-allelic coalescent method of Slatkin and Rannala (1997) for a range of values of ξ , the growth/selection coefficient. The ξ coefficient

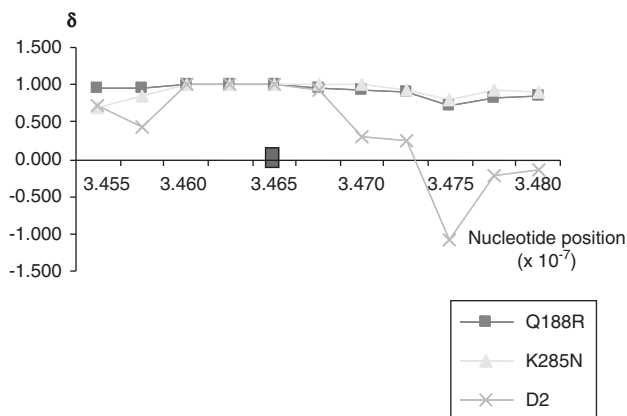


Figure 3 Decay of linkage disequilibrium across the *GALT* haplotype, estimated as the δ parameter for Q188R, K285N and D2 alleles. The red rectangle indicates the position of the *GALT* gene within the haplotype interval. δ is the linkage disequilibrium value between the specific mutation and haplotype marker.

varies in the range -1 to 1 , negative values indicating population decline and/or negative selection and positive values indicating population growth and/or positive selection (Slatkin and Rannala, 1997). Choosing a value of ξ is difficult, but estimates suggest that a value of between 0.005 and 0.010 may reflect most accurately population growth over the history of the European population (Slatkin and Bertorelle, 2001). Using a ξ value of 0.010, this method yielded age estimates for Q188R, K285N and D2 of 26 875, 20 575 and 33 100 years, respectively (Table 2).

Discussion

Considerable effort has been devoted to cataloguing *GALT* mutations in transferase-deficient galactosaemia worldwide, showing that most mutations are rare with only a few being observed in multiple geographic regions and/or ethnic groups (Tyfield *et al.*, 1999). Attention has focused on the small number of more widespread mutations (including Q188R, K285N and N314D) in an attempt to draw inferences about their origins and age from frequency and geographic distribution data. Such studies have suggested that Q188R and K285N arose in Europe within the last 50 000 years whereas N314D may have originated much earlier in an ancestral African population (Tyfield *et al.*, 1999). However, there is some debate as to where these mutations originated and their subsequent spread (Tyfield *et al.*, 1999). In an effort to shed more light on these basic conclusions, we have established a haplotyping system for the *GALT* locus and investigated the haplotype diversity associated with wild-type and mutant *GALT* alleles.

We have shown that the Q188R alleles gave rise to a moderately extended MJ network of haplotypes centred on the most common, putatively ancestral, G1A haplotype (Figure 2a). In contrast, wild-type alleles demonstrated a considerably more diverse network in which all haplotype nodes were represented at frequencies of 5% or less (Figure 2d). This difference in diversity was reflected at the level of SNP-only haplotypes and the Q188R haplotype genetic diversity demonstrated differences across Europe by region, with higher h values in central Europe relative to the 'Atlantic Façade' (Table 1).

Table 2 Allele ages estimates for the Q188R, K285N and D2 variant alleles of the Galactose-1-phosphate uridyl transferase (*GALT*) gene in the European population

Allele	$(\delta\mu)^2$ method			Intra-allelic coalescent method		
	Mutation rate	Age (years)	95% CI (years)	ξ value	TMRCAs (years)	95% CI (years)
Q188R	0.0001	97 335	25 110–169 560	0.01	26 875	23 225–33 475
	0.0005	19 467	5022–33 912	0.02	14 425	12 550–18 100
	0.0010	9747	2511–16 983	0.04	7700	6800–9550
K285N	0.0001	23 085	1 755–44 415	0.01	20 575	16 900–27 850
	0.0005	4617	351–8883	0.02	11 325	9450–14 975
	0.0010	2322	189–4455	0.04	6075	5175–8025
Duarte (D2)	0.0001	1 230 255	75 060–2 385 450	0.01	33 100	29 175–40 525
	0.0005	246 051	15 012–477 090	0.02	17 525	15 575–21 275
	0.0010	123 039	7506–238 572	0.04	9 225	8250–11 100

Allele ages were computed using the $(\delta\mu)^2$ distance and the intra-allelic coalescent methods. All allele ages are estimated based upon a generation time (g) of 27.

Given that the relative frequency of the Q188R allele is known to increase across Europe from south-east to north-west (Murphy *et al.*, 1999; Tyfield *et al.*, 1999), this implies that Q188R haplotype genetic diversity decreases as relative Q188R allele frequency increases from central Europe to the 'Atlantic Façade'. As more haplotype divergence is expected where the mutation has resided longest, the Q188R mutation appears to have originated in central Europe. Our haplotype analysis of the Q188R mutation described above defines an origin of the Q188R mutation in central Europe with age estimates of 19 467 (CI 5022–33 912) and 26 875 (CI 23 225–33 475) years using the $(\delta\mu)^2$ and the Slatkin methods, respectively. These results argue against Ireland having been the 'centre of diffusion' of Q188R in Europe (Murphy *et al.*, 1999), and this finding indicates that Q188R was carried across Europe from a central/eastern European source by human migration. There is extensive genetic and archaeological data that describes the population history of the European continent, with three large-scale population events having been inferred; the Palaeolithic, Mesolithic and Neolithic expansions (Barbujani and Bertorelle, 2001). The Mesolithic expansion occurred approximately 18 000 years ago, following the last glacial maximum, when human populations expanded from their glacial refugia to repopulate Europe (Hewitt, 2000). The estimated age of Q188R falls within this time period, indicating that the gradient of Q188R relative allele frequency across Europe may reflect repeated founder effects and genetic drift during this migration.

We found that the haplotypes of the K285N alleles produced an MJ Network centred on the ancestral G3P haplotype that was considerably less extended than that for Q188R and, particularly, that for wild-type alleles (Figure 2). Like Q188R, this was reflected in a reduced overall h value for SNP-only haplotypes associated with K285N compared with wild-type alleles (Table 1). The SNP-only haplotypes showed an overall genetic diversity of 0.381 ± 0.076 with some indication of differences by region, though sample numbers were too small to allow a pattern to be discerned (Table 1). Our age estimates of K285N yielded ages of 4617 (C.I 351–8883) and 20 575 (C.I 16 900–27 850) years using the $(\delta\mu)^2$ and the Slatkin methods, respectively. We found that the age estimates generated for K285N were consistently lower than Q188R for each method used. This confirms that K285N is a younger mutation than Q188R, and a more recent origin would explain why K285N is more geographically restricted compared with Q188R.

The D2 alleles produced a markedly different picture, with a substantially extended MJ network lacking an obvious ancestral node (Figure 2) and an overall h value for SNP-only haplotypes closer to that for wild-type alleles than to the values determined for Q188R and K285N (Table 1). As D2 alleles from all populations sampled thus far preserve the linkage disequilibrium relationship between N314D, the three intronic substitutions and the promoter deletion [IVS4nt-27g>c, IVS5nt-24g>a, IVS5nt+62g>a, 5'UTR-119delGTCA], it is likely that the D2 extended MJ network is not the result of recurrent mutation, but rather reflects a mutation event that occurred far earlier than those which gave rise to Q188R and K285N (Tyfield *et al.*, 1999). Given that

linkage disequilibrium decays progressively with allele age, we used the δ parameter of Risch *et al.* (1995) to investigate LD across the haplotype span for the three mutations (Figure 3). We demonstrated that there was strong LD in the region flanking the *GALT* locus and that the LD patterns supported the hypothesis that D2 is considerably older than either Q188R or K285N. It is interesting to note that δ values for D2 fell off more sharply on the centromeric side of *GALT*, perhaps consistent with the presence of a recombination 'hot-spot' between SNP markers rs2252539 and rs2249561. Estimations of D2 allele age yielded average values of 246 051 (C.I 15012–477 090) and 33 100 (C.I 29 175–40 525) years using the $(\delta\mu)^2$ and the Slatkin methods, respectively. Of all three mutations analysed, calculation of D2 allele age showed the greatest degree of variability between the age estimation methods. This is most likely due to the difficulty in accurately calculating the age of an allele associated with a large degree of diversity. In particular, the ages of origin generated by the intra-allelic coalescent method were considerably younger than those produced by the $(\delta\mu)^2$ method. This probably reflects the fact that the analysis was performed with the full STR data set, such that the diversity captured by the STR markers is actually distributed across a range of nodes within the D2 haplotype network rather than being associated with a majority (ancestral) node as is the case with Q188R and K285N. Overall, haplotype analysis of D2 was consistent with the mutation being an ancient event that probably arose early in human evolution despite having a wide range (15 012–477 090 years) in its estimated age of origin. This is in accordance with similar conclusions based upon cross-species analysis of *GALT* sequence variants (Carney *et al.*, 2009).

In this study, we have demonstrated the increased power afforded by the analysis of haplotypes in delineating the origins of disease-causing mutations. Similar to haplotype analysis of the phenylalanine hydroxylase (*PAH*) gene (Tighe *et al.*, 2003), haplotype analysis of the *GALT* gene has ascertained the genetic history of three common *GALT* mutation, while also giving some insight into human migration.

Acknowledgements

We acknowledge Professor Guido Barbujani [Dipartimento di Biologia, Università di Ferrara, Italy] and Professor Judith Fridovitch-Keil [Department of Human Genetics, Emory University School of Medicine, Atlanta, GA, USA] for helpful discussions and expert advice in the area of population genetics. We acknowledge with gratitude financial support received in the form of project grants from the Health Research Board of Ireland, the Research Committee of the Royal College of Surgeons in Ireland and an EMBO Short-Term Fellowship award to DTC [EMBO ASTF 110.00–03]. JKVR is a Medical Foundation Fellow at the University of Sydney.

Web Resources:

ARUP Laboratories *GALT* Mutation Database: http://www.arup.utah.edu/database/galactosemia/GALT_welcome.php

Arlequin Software Download: <http://lgb.unige.ch/arlequin/software/>

References

- Ashino J, Okano Y, Suyama I, Yamazaki T, Yoshino M, Furuyama J *et al.* (1995). Molecular characterization of galactosemia (type 1) mutations in Japanese. *Hum Mutat* **6**: 36–43.
- Bandelt HJ, Forster P, Rohl A (1999). Median-joining networks for inferring intraspecific phylogenies. *Mol Biol Evol* **16**: 37–48.
- Barbujani G, Bertorelle G (2001). Genetics and the population history of Europe. *Proc Natl Acad Sci USA* **98**: 22–25.
- Carney AE, Sanders RD, Garza KR, McGaha LA, Bean LJ, Coffee BW *et al.* (2009). Origins, distribution and expression of the Duarte-2 (D2) allele of galactose-1-phosphate uridylyltransferase. *Hum Mol Genet* **18**: 1624–1632.
- Cooper G, Amos W, Bellamy R, Siddiqui MR, Frodsham A, Hill AV *et al.* (1999). An empirical exploration of the ($\Delta\mu$)² genetic distance for 213 human microsatellite markers. *Am J Hum Genet* **65**: 1125–1133.
- Elsas LJ, Dembure PP, Langley S, Paulk EM, Hjelm LN, Fridovich-Keil J (1994). A common mutation associated with the Duarte galactosemia allele. *Am J Hum Genet* **54**: 1030–1036.
- Elsas LJ, Lai K, Saunders CJ, Langley SD (2001). Functional analysis of the human galactose-1-phosphate uridylyltransferase promoter in Duarte and LA variant galactosemia. *Mol Genet Metab* **72**: 297–305.
- Excoffier L, Laval G, Schneider S (2005). Arlequin ver. 3.0: An integrated software package for population genetics data analysis. *Evolutionary Bioinformatics Online* **1**: 47–50.
- Goldstein DB, Ruiz LA, Cavalli-Sforza LL, Feldman MW (1995). Genetic absolute dating based on microsatellites and the origin of modern humans. *Proc Natl Acad Sci USA* **92**: 6723–6727.
- Greber-Platzer S, Guldberg P, Scheibenreiter S, Item C, Schuller E, Patel N *et al.* (1997). Molecular heterogeneity of classical and Duarte galactosemia: Mutation analysis by denaturing gradient gel electrophoresis. *Hum Mutat* **10**: 49–57.
- Hewitt G (2000). The genetic legacy of the Quaternary ice ages. *Nature* **405**: 907–913.
- Hirokawa H, Okano Y, Asada M, Fujimoto A, Suyama I, Isshiki G (1999). Molecular basis for phenotypic heterogeneity in galactosaemia: Prediction of clinical phenotype from genotype in Japanese patients. *Eur J Hum Genet* **7**: 757–764.
- Kozak L, Francova H, Fajkusova L, Pijackova A, Macku J, Stastna S *et al.* (2000). Mutation analysis of the GALT gene in Czech and Slovak galactosemia populations: Identification of six novel mutations, including a stop codon mutation (X380R). *Hum Mutat* **15**: 206.
- Kozak L, Francova H, Pijackova A, Macku J, Stastna S, Peskovicova K *et al.* (1999). Presence of a deletion in the 5' upstream region of the GALT gene in Duarte (D2) alleles. *J Med Genet* **36**: 576–578.
- Murphy M, Mchugh B, Tighe O, Mayne P, O'Neill C, Naughten E *et al.* (1999). Genetic basis of transferase-deficient galactosaemia in Ireland and the population history of the Irish Travellers. *Eur J Hum Genet* **7**: 549–554.
- Ng WG, Xu YK, Kaufman FR, Donnell GN, Wolff J, Allen RJ *et al.* (1994). Biochemical and molecular studies of 132 patients with galactosemia. *Hum Genet* **94**: 359–363.
- Reichardt JK, Packman S, Woo SL (1991). Molecular characterization of two galactosemia mutations: Correlation of mutations with highly conserved domains in galactose-1-phosphate uridylyl transferase. *Am J Hum Genet* **49**: 860–867.
- Risch N, De LD, Ozelius L, Kramer P, Almasy L, Singer B *et al.* (1995). Genetic analysis of idiopathic torsion dystonia in Ashkenazi Jews and their recent descent from a small founder population. *Nat Genet* **9**: 152–159.
- Rozas J, Rozas R (1999). DnaSP version 3: An integrated program for molecular population genetics and molecular evolution analysis. *Bioinformatics* **15**: 174–175.
- Slatkin M, Bertorelle G (2001). The use of intraallelic variability for testing neutrality and estimating population growth rate. *Genetics* **158**: 865–874.
- Slatkin M, Rannala B (1997). Estimating the age of alleles by use of intraallelic variability. *Am J Hum Genet* **60**: 447–458.
- Stephens M, Donnelly P (2003). A comparison of bayesian methods for haplotype reconstruction from population genotype data. *Am J Hum Genet* **73**: 1162–1169.
- Stephens M, Smith NJ, Donnelly P (2001). A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet* **68**: 978–989.
- Suzuki M, West C, Beutler E (2001). Large-scale molecular screening for galactosemia alleles in a pan-ethnic population. *Hum Genet* **109**: 210–215.
- Tighe O, Dunican D, O'Neill C, Bertorelle G, Beattie D, Graham C *et al.* (2003). Genetic diversity within the R408W phenylketonuria mutation lineages in Europe. *Hum Mutat* **21**: 387–393.
- Tighe O, Flanagan JM, Kernan F, O'Neill C, Naughten E, Mayne PD *et al.* (2004). Frequency distribution of the Los Angeles and Duarte galactose-1-phosphate uridylyltransferase variant alleles in the Irish population. *Mol Genet Metab* **82**: 345–347.
- Trbusek M, Francova H, Kozak L (2001). Galactosemia: Deletion in the 5' upstream region of the GALT gene reduces promoter efficiency. *Hum Genet* **109**: 117–120.
- Tyfield L, Reichardt J, Fridovich-Keil J, Croke DT, Elsas LJ, Strobl WM *et al.* (1999). Classical galactosemia and mutations at the galactose-1-phosphate uridylyl transferase (GALT) gene. *Hum Mutat* **13**: 417–430.

Supplementary Information accompanies the paper on Heredity website (<http://www.nature.com/hdy>)