European Journal of Human Genetics (2009) 17, 219–227

www.nature.com/ejhg

ARTICLE

Natural selection and the molecular basis of electrophoretic variation at the coagulation *F13B* locus

Anthony W Ryan*[,1,2], David A Hughes[1], Kun Tang[1], Dermot P Kelleher[2], Thomas Ryan[2], Ross McManus[2] and Mark Stoneking[1]

[1]*Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany;* [2]*Department of Clinical Medicine, Trinity College Dublin and Institute of Molecular Medicine, Trinity Centre for Health Sciences, St James's Hospital, Dublin, Ireland*

Electrophoretic analysis of protein variation at the coagulation *F13B* locus has previously revealed three alleles, with alleles 1, 2, and 3 each being at high frequency in European, African, and Asian populations, respectively. To determine if this unusual pattern of interpopulation differentiation reflects local natural selection or neutral genetic drift, we re-sequenced 4.6 kb of the gene, encompassing all exons, splice junctions, and 1.4 kb of the promoter, in African, European, and Asian samples. These analyses revealed three major lineages, which correspond to the common protein alleles and differ from each other at a non-synonymous substitution in exon 3 and a novel splice acceptor in intron K. There is previous evidence that these lineages are not functionally equivalent; we therefore carried out case–control analyses and confirmed that variability at *F13B* modulates susceptibility and/or survivorship in coronary artery disease ($P<0.05$) and type II diabetes within the coronary artery disease cohort ($P<0.01$). Tajima's D and Fu and Li's tests did not indicate significant departures from neutral expectations. However, publicly available data from SeattleSNPs and HapMap do indicate highly unusual levels of population differentiation ($P=0.003$) and an excess of allele-specific, extended haplotype homozygosity within the African population ($P=0.0125$). Possible causes of this putative signal of selection include hematophagous organisms, infection by pathogens that cause disseminated intravascular coagulation, and metabolic or dietary factors.
European Journal of Human Genetics (2009) **17,** 219–227; doi:10.1038/ejhg.2008.137; published online 20 August 2008

Keywords: coagulation factors; population genetics; classical marker genetics; natural selection; alternative splicing

Introduction

Factor XIII or fibrin-stabilizing factor circulates in the blood as the A_2B_2 tetramer, composed of two identical A subunits (which have catalytic activity) and two identical B subunits, which are encoded by separate genes *F13A* and *F13B* (OMIM 134570 and 134580, respectively). The A subunit, a transglutaminase, crosslinks fibrin and thus stabilizes the blood clot at the end of the coagulation cascade. The A subunit also functions as an intracellular enzyme, produced by megakaryocytes and packaged into platelets, where it likely plays a role in cytoskeletal remodeling. FXIIIA is also detected in monocytes and

*Correspondence: Dr AW Ryan, Department of Clinical Medicine, Trinity College Dublin, St James's Hospital, Dublin 8, Ireland.
Tel: +353 1 896 3273; Fax: +353 1 896 3503;
E-mail: aryan12@tcd.ie
Received 31 March 2008; revised 25 June 2008; accepted 25 June 2008; published online 20 August 2008

macrophages.[1] The B subunit, a protein with 10 sushi domains, has no known catalytic activity and is thought to act as a carrier protein for circulatory factor XIII.[2] Individuals with factor XIII deficiency exhibit defective coagulation, slow wound healing and a higher rate of spontaneous abortions.[3,4]

Genetic variation at the *F13B* locus was first demonstrated using protein electrophoresis through non-denaturing gels, which resolved three common alleles.[5] Later work cast doubt on the existence of three alleles at the locus,[6,7] but these were later resolved, with the difficulties due, in part, to the rarity of one of the common alleles (allele-2) in the Japanese population.[8] The protein polymorphism shows considerable geographic differentiation, with alleles 1, 2, and 3 being the most common alleles (>60%) in populations of European, African, and Asian descent, respectively.[9,10]

There is some evidence that the three common protein alleles at *F13B* are not functionally equivalent. Factor XIII activity in samples representing different geographic human populations demonstrated that at least some of the variance was due to the B subunit, despite the fact that it has no known catalytic activity.[11] A later molecular study[12] showed that the *95Arg* allele of the B subunit was associated with an increased rate of dissociation of the A_2B_2 tetramer, a key step in the activation of the enzyme. Case–control data from the same study also suggested a role for the *F13B 95Arg* variant in susceptibility to thrombosis. Furthermore, the *95Arg* variant appears to interact with the *34Leu* variant of the A subunit to reduce the risk of non-fatal myocardial infarction (MI) in women undergoing postmenopausal estrogen therapy.[13]

The high level of interpopulation differentiation[10] and the suggestion that at least some of the polymorphism at this locus may be functionally significant[11,12] point to the possibility that this locus may have been under the effect of local natural selection. However, a neutral explanation is also possible. To distinguish between these possibilities, we sequenced all of the exons, splice junctions, and a large portion of the region immediately upstream, in samples of Asians, Africans, and Europeans. The aims of this study were to: (1) determine the molecular basis of the tri-allelic *F13B* protein polymorphism; (2) confirm the relevance of genetic variation at this locus to genetic susceptibility to coronary artery disease (CAD); and (3) to investigate the locus for signatures of localized natural selection.

Materials and methods

DNA samples from nine Nigerians (representing Yoruba, Ibo, and Hausa ethnicities), nine Han Chinese from Taiwan, and eight Norwegians from Trondheim were used for the re-sequencing study. In addition, samples from 26 individuals of known protein phenotype (including four allele-1 homozygotes, one allele-2 homozygote, and three allele-3 homozygotes) were used to confirm the cosegregation of the molecular sequence variants and protein alleles. All sampling followed the ethical guidelines of the Max Planck Institute for Evolutionary Anthropology, Leipzig, including informed consent. In addition, a single orangutan was sequenced for the same regions, for use as an outgroup in tests for departure from neutral expectation.

PCR fragments (Figure 1) were amplified using the primers and conditions shown in Supplementary Table 1, and direct sequencing of PCR fragments was performed on these amplicons using the same or nested primers and the BigDye Terminator Cycle Sequencing Kit (Applied Biosystems, Foster City, CA, USA). All exon fragments were sequenced on both strands, with the exceptions of exons 4 and 11, which were found to be flanked by polymorphic repeats; a 5′-poly-A with A_4 and A_5 alleles is present just upstream of exon 4, whereas a poly-T is present just downstream from exon 11. PCR fragments could not be directly sequenced on both strands in heterozygous individuals, so in these cases, the exons were sequenced twice on the same strand. Sequences were aligned using ClustalW[14] and converted to MEGA file format for sequence analysis (standard tests for departure from neutrality) in DNAsp.[15] Gametic phase was inferred using the SSD algorithm[16] for these data and for the publicly available SeattleSNP (http://pga.mbt.washington.edu/) *F13B* gene sequence data.

The publicly available HapMap Phase II (http://www.hapmap.org/) data were used to calculate pairwise and three-way weighted average F_{ST} for 1000 randomly selected, SNP density-controlled, 100 kb windows on chromosome 1 as well as for a 100 kb window centered on *F13B*.[17,18] Empirical significance levels were obtained from the position of observed values within the distribution of random values. Integrated extended haplotype

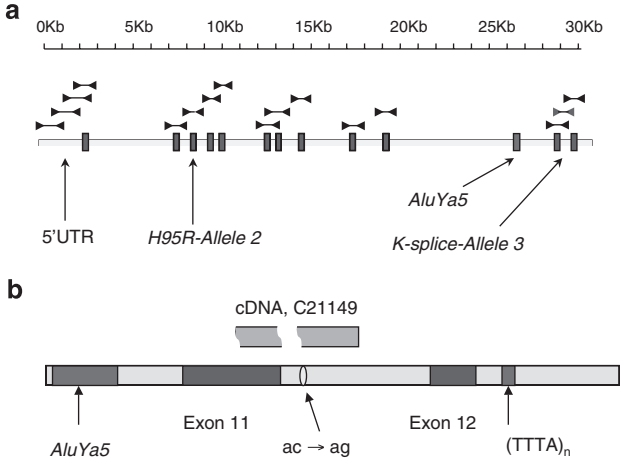

Figure 1 The *F13B* gene. (**a**) Schematic representation of the *F13B* gene, showing the exons (blue) and the areas sequenced. (**b**) Molecular basis of protein allele-3.

homozygosity (iEHH) values were calculated for each allele in HapMap phase I data as found in Tang *et al*.[19] Relative extended haplotype homozygosity within a population (Raw) and between populations (Rab) was determined as previously described.[19,20] The significance of Rab and Raw values was determined by calculating the proportion of sites within the observed window, which lies above the 95th and 99th percentiles of the HapMap distribution. First, each Rab or Raw value was binned according to the frequency of the allele in the numerator; all alleles below 5% frequency were excluded. Then the 95th and 99th percentiles of each bin were determined. Finally, the proportion of sites in each sampled window, which was above the 95th and 99th percentiles for the respective frequency bin, was calculated. For Raw calculations, the allele with the largest iEHH was placed in the numerator. As done with F_{ST} calculations, a thousand 100 kb windows were sampled from chromosome 1, controlling for SNP density by only including windows that contained ± 5% of the observed number of SNPs.

The case–control study included 144 individuals who had undergone coronary artery bypass graft (CABG) for CAD. Individuals who had undergone CABG for congenital defects were not included in the analyses. In all, 311 unmatched controls (250 healthy blood donors and 61 health-care workers) were also analyzed. All individuals were of Irish ethnicity and gave informed consent for the study. Genotypes at the *F13A Val34Leu* and *F13B His95Arg* SNPs were determined using Amplifluor™ technology, and variation at the *F13B* Intron K splice variant was assayed using a Taqman™ 5′ nuclease technology. Statistical analyses of the case–control data were performed using Genepop[21] and EpiInfo.[22] Haplotypes for the *F13B* loci were computationally inferred using HITAGENE (http://hitagene.tchpc.tcd.ie/hitagene/), and haplotype score statistics for disease association were calculated using Haplo Stats.[23]

Results

Identification of the molecular basis of the protein polymorphism

Exon 1 and 1358 bp of the sequence immediately upstream were sequenced, from position 1573 to 2930 of GeneBank accession M64554, inclusive. In addition, all coding exons from 2 to 12 were also sequenced. The total length of the sequences was 4638 bp; all polymorphic sites are shown in Supplementary Table 2.

Given the high frequency of the protein alleles in the different geographic regions (ie, protein allele-1 has average frequency 0.73 in Europeans; protein allele-2 has average frequency 0.60 in African Americans; and protein allele-3 has average frequency 0.70 in Asians),[10] our expectation was that sequencing random individuals from these three geographic regions should reveal candidate non-synonymous substitutions for the molecular basis of the protein polymorphism at *F13B*. In the case of allele-2, this expectation was fulfilled, as a histidine to arginine substitution in exon 3 (M64554 position 8259) corresponded to this allele, the most common type in populations of African origin. Genotyping of individuals of known protein phenotype confirmed the cosegregation of the exon 3 substitution with protein allele-2 (data not shown). This variable site corresponds to *F13B His95Arg*.[12] The G (Arg) allele is at high frequency in the African samples and was not found among the samples from Asia, where protein allele-2 is known to be rare.[10]

A candidate for the molecular basis of protein allele-3 initially proved elusive, as no non-synonymous or 5′-UTR polymorphism appeared to correspond to this allele. However, a 3′-directed cDNA sequence (GenBank accession C21149) of an alternative splice variant of FXIIIB was found by BLAT search (http://genome.ucsc.edu/cgi-bin/hgBlat) to map onto the end of exon 11 and include part of intron K (Figure 1). As this cDNA clone (C21149) was isolated by a Japanese laboratory, in a geographical area where protein allele-3 is the predominant type, we investigated if this alternative splice variant could explain the protein allele-3. We sequenced this region in eight individuals of known protein phenotype and found that the three allele-3 homozygotes were all homozygous G (instead of the usual C) at position 29756. This changes an AC sequence in M64554 to a putative AG splice acceptor in individuals with this variant. The same two splice variants of *F13B* are described on the ACEVIEW database (http://www.ncbi.nlm.nih.gov/IEB/Research/Acembly/).[24] The splice variant (termed *F13B.bAug2005*) exhibits an alternative carboxy terminus and may be incomplete at the 5′ end. The *F13B C29756G* alleles will hereafter be referred to as *IntK-ve* (C) and *IntK + ve* (G), respectively.

In addition to the *His95Arg* substitution at M64554 position 8259, four additional SNPs in the region immediately upstream of exon 1, which is likely to contain regulatory elements, also show association with allele-2. Indeed, the protein allele-2 homozygote shows a TTAC motif, the predominant haplotype in Africa, at positions 1617, 1835, 2047, and 2416, respectively. The predominant motif for alleles 1 and 3 at these positions is CCGT. As protein allele-2, associated with the *95Arg* lineage, may have higher factor XIII activity,[11] we searched the sequences representing both major lineages (CCGT and TTAC) for common transcription factor consensus binding motifs using the program TFSearch (http://www.cbrc.jp/research/db/TFSEARCH.html). The *95Arg* lineage contains a T allele at position 1617, which creates a putative AP-4 (OMIM 600743) recognition site that is absent in the other lineage. Likewise, the *95Arg* lineage contains a T at position 1835, creating a putative GATA 1–3 (OMIM 305371, 137295, and 131320) recognition site that is absent in *95His*.

Table 1 Phase v2 haplotypes for F13B positions 1617, 1835, 2047, 2416, 8259 (His95Arg) and 29756 (intron K splice variant)

Haplotype	*Nigeria*	*Han*	*Norway*	*African American*	*European American*
TTACGC	0.722	0.000	0.187	0.454	0.043
CCATGC	0.055	0.000	0.000	0.022	0.000
TCATGC	0.001	0.000	0.000	0.045	0.000
TTACGG	0.000	0.000	0.000	0.000	0.000
CCATAG	0.000	0.000	0.000	0.001	0.000
CCGTAG	0.000	0.667	0.188	0.049	0.196
CCATAC	0.056	0.000	0.000	0.021	0.000
CCGTAC	0.167	0.333	0.625	0.407	0.761

African and European American samples are from the SeattleSNPs database.

Table 2 Genotype frequencies and Hardy–Weinberg Equilibrium (P_{HWE}) for *F13B H95R*, *F13B IntK,* and *F13A V34L* and haplotype frequencies for *F13B* in coronary artery bypass graft (CABG) cases and unmatched controls

(a) Genotype frequencies and frequency heterogeneity (P_{HET}) between CABG cases and controls

Locus	*Case/control*				P_{HWE}	P_{HET}
F13B H95R		**AA**	**AG**	**GG**		
	Control	269	36	6	0.0053	
	Case	115	25	1	1.0	0.4089
F13B IntK		**CC**	**CG**	**GG**		
	Control	216	61	8	0.1983	
	Case	85	49	2	0.1052	0.0330[a]
F13A V34L		**GG**	**GT**	**TT**		
	Control	153	116	21	1.0	
	Case	78	56	10	1.0	0.8124

(b) Comparison of haplotype frequencies and Haplo Stats analysis comparing CABG cases and controls[b]

Haplotype	*Control*	*Case*	*Haploscore*	P_{HS}	P_{SIM}
95His-IntK−ve	0.785	0.712	−2.53261	0.01132	0.018
95His-IntK+ve	0.137	0.192	2.25094	0.02439	0.026
95Arg-IntK−ve	0.076	0.095	0.91246	0.36153	0.351

(c) Haplo Stats analysis of type II diabetes (DM) susceptibility within the CABG cohort

Haplotype	*DM*	*No DM*	*Haploscore*	P_{HS}	P_{SIM}
95His-IntK−ve	0.639	0.754	−1.61393	0.10654	0.111
95His-IntK+ve	0.333	0.144	2.77181	0.00557	0.008
95Arg-IntK−ve	0.028	0.098	−1.23577	0.21654	0.214

[a]Odds ratio (CC *vs* CG/GG) OR = 1.88 (1.18–2.99), $\chi^2 = 7.98$, $P = 0.0047$.
[b]Statistical significance is given by P_{HS} and an empirical P_{SIM} from 1000 simulations.

Population variation

There are three major *F13B* haplotypes: *CCGT-95His-IntK−ve*, *TTAC-95Arg-IntK−ve,* and *CCGT-95His-IntK + ve* (positions 1617, 1835, 2047, 2416, 8259, and 29756). Two of these (*CCGT-95His-IntK−ve* and *TTAC-95Arg-IntK−ve*) differ by five SNPs spread over 6.6 Kb. The third lineage, *CCGT-95His-IntK + ve*, exhibits very low associated molecular diversity: individuals who are homozygous for *IntK + ve* are homozygous at all other SNPs. Despite this, *IntK + ve* is present at very high frequency (> 60%) in the Asian population, whereas it is rare or absent in Africa (Table 1).

Case–control study

Allele frequency heterogeneity test results are presented in Table 2. Only the *F13B-IntK* locus shows evidence of allele frequency heterogeneity between CABG cases and unmatched controls (Table 2a, $P = 0.0330$). This conclusion is supported by analysis of the data on the basis of a dominant model for carrier status of the *F13B-IntK* polymorphism (Odds ratio = 1.88 (1.18–2.99), $\chi^2 = 7.98$, $P = 0.0047$).

Haplotype analyses (Table 2b) showed that the *95His-IntK + ve* haplotype was significantly more frequent in the CABG group ($P = 0.026$), whereas the *95His-IntK−ve*

haplotype appeared to be protective ($P=0.018$) (Table 2b). The global haplotype score statistic (all haplotypes) was 6.76851, d.f. = 2, $P=0.0339$, indicative of significant haplotypic heterogeneity between samples. Within the CABG cohort, the *95His-IntK+ve* haplotype was significantly associated with risk of type II diabetes (global haplotype score = 8.48039, d.f. = 3, *P*-value = 0.03706, corrected for age and sex; for *95His-IntK+ve* haplotype, $P=0.008$, adjusted for age and sex, Table 2c).

Tests for departure from neutrality

Genetic variation at the *F13B* locus appears to be unusual, with three protein alleles found in three different parts of the world (Europe, Africa, Asia/Americas) each at a frequency of >60%. To determine how unusual such a pattern is, we used the HapMap data to calculate molecular F_{ST} for a 100 kb region centered on *F13B* and for 1000 randomly sampled 100 kb regions from the same chromosome (1) with similar SNP densities (Table 3a). It is apparent that allele frequencies across the *F13B* region in Yoruba (Africa) are greatly differentiated from those of CEPH (European) and Han Chinese (Asian), $F_{ST}=0.44$ ($P=0.005$) and 0.55 ($P=0.004$), respectively. This is consistent with theoretical and simulated expectations of directional selection[25–28] altering allele frequencies between populations beyond that which is expected from

Table 3 Statistical tests for departure from the expectations of the neutral theory

(a) Pairwise F_{ST} comparisons using HapMap data[a]

Comparison	Fst	P-*value*
HapMap F_{ST}		
AE	0.443833	0.005*
AC	0.553303	0.004*
CE	0.193754	0.141
3-way	0.46228	0.003*

(b) Standard tests for departure from neutrality[b]

		Tajima's D	*Fu and Li's D*	*Fu and Li's F*
This study:				
Nigeria	18	−0.31876	−0.09612	−0.17628
Han	18	0.76903	0.87120	0.98414
Norway	16	0.00188	0.15145	0.12877
SeattleSNPs:				
African American	46	−0.74294	−1.1542	−1.21127
European American	46	−0.23158	0.0541	−0.08325

(c) Extended haplotype homozygosity analysis[c]

	95th percentile			*99th percentile*			
Comparison	*No. of sig. sites*	*Proportion*	P-*value*	*No. of sig. sites*	*Proportion*	P-*value*	*Count*
HapMap Rab							
EA	12	0.179104	0.095	1	0.014925	0.218	67
CA	4	0.064516	0.313	0	0	1	62
JA	10	0.15873	0.145	4	0.063492	0.073	63
CE	8	0.126984	0.116	0	0	1	63
JE	7	0.109375	0.146	0	0	1	64
EC	0	0	1	0	0	1	63
EJ	0	0	1	0	0	1	64

	95th percentile			*99th percentile*			
Population	*No. of sig. sites*	*Proportion*	P-*value*	*No. of sig. sites*	*Proportion*	P-*value*	*Count*
HapMap Raw							
Africa	9	0.3	0.0125*	1	0.033333	0.033*	30
Europe	0	0	1	0	0	1	27
Han	1	0.066667	0.145	0	0	1	15
Japanese	0	0		0	0	1	15

[a]A = Yoruba, C = Han Chinese, E = European in pairwise comparisons.
[b]All tests were performed using DNAsp (Rozas and Rozas, 1999). Data from this study used orangutan sequence as an outgroup, whereas *Pan troglodytes* was used as an outgroup for SeattleSNPs data.
[c]HapMap Rab and Raw values for the *F13B* region. *Comparison* is the two populations being compared. In the case of Rab the first population is the one being the tested for large iEHH. E = CEPH European, A = Yoruba, C = Han Chinese, J = Japanese, *no. of sig. sites* is the number of sites which were found above said percentile at the *F13B* region, *proportion* is the proportion of sites found above said percentile, *count* is the number of data points in each analysis. Significance is denoted with an asterisk (*).

drift alone. As we are able to compare the observed values with other regions of the same size and SNP density, it is arguable that the effects of population-specific demographic events may be excluded as being causative of the observations. Thus, it would appear that directional selection has acted in one or each of these human populations. To determine which population might have experienced a selective event, one may identify the population(s) that exhibits more homozygosity. At the *F13B* locus, there is ~2.15 times more haplotype homozygosity found in non-African populations than in Yoruba, but not significantly so, as detailed below.

Next, we utilized our sequence data and that of the SeattleSNPs sequence data of the *F13B* region to investigate departures from neutrality employing allele frequency tests, specifically Tajima's D, Fu and Li's D, and Fu and Li's F (Table 3b). Given the unusual F_{ST} values in the region summarized above, we may expect to find an excess of young alleles and thus large negative values in this region. However, as seen in Table 3b, the *F13B* region exhibits no departures from neutral expectations in the distribution of allele frequencies.

Finally, we applied recently developed EHH tests, Rab and Raw, to determine if this region has an excess of alleles with larger than expected haplotype homozygosity.[19] Raw[19] is analogous to the classic EHH test,[29] in which the EHH associated with two alternative alleles within a population are compared to each other. Raw differs from previous formulations[29] in that it is based on single SNPs rather than haplotypes. Under neutrality, one would expect high-frequency alleles to be old and to have experienced many recombination events and carry many mutations relative to the alternative allele, thereby exhibiting low EHH. However, a high-frequency allele that has experienced directional selection would have increased in frequency quickly and contain little variation thereby exhibiting high EHH. The Raw statistic compares two alternative alleles within a single population, whereas the Rab statistic compares the same allele between populations and has more power than Raw to detect sweeps that are at or near fixation.[19]

The Yoruba contain a significant excess of *F13B* alleles in the top 5% ($P = 0.0125$) and the top 1% ($P = 0.033$) of the Raw distribution as compared to other alleles of similar frequency (Table 3c). The nine significant sites in the 100 kb window around *F13B* have a median allele frequency of 0.69 (min = 0.56, max = 0.70) and a median Raw value of 2.197 (min = 2.045, max = 4.57), indicating that high-frequency alleles are driving the deviation from neutral expectations. Each of these sites is in LD ($D' = 1$) with non-synonymous substitution rs6003, which defines allele-2 (*His95Arg*) and has an allele frequency of 0.725 in HapMap Yoruba individuals. The other HapMap populations do not exhibit an excess of alleles with large EHH (Table 3c).

Discussion

Historically, the use of protein electrophoresis has made significant impact on population and anthropological genetics. It has generally been assumed that most of the changes detected reflect amino-acid substitutions that alter the charge, size, and shape of the resulting protein allele. Alternative splicing, which could potentially affect all three detectable protein properties, has not hitherto been recognized as a potential source of protein electrophoretic variation.

The classical protein assay for genetic variation at the *F13B* locus has been in use since the early 1980s. However, the molecular basis of this protein polymorphism has not been resolved to date. Our sequence analysis suggests that the *Arg95* allele corresponds to protein allele-2, with the expected frequencies in worldwide populations and in individuals of know protein type. Protein allele-3, the predominant Asian type, appears to be due to an SNP in intron K (allele *IntK + ve*), which creates a new consensus splice acceptor site. This is consistent with the observed sequence in three known protein allele-3 homozygotes and the observed frequencies in worldwide populations.

Genetic variation at splice sites, leading to population splice variation, has been observed as mutations in family studies[30,31] and as population polymorphisms.[32] However, to our knowledge, this represents the first instance in which a splice variant accounts for the molecular basis of a classical protein polymorphism.

The *His95Arg* site, which differs widely in frequency among worldwide populations and is particularly frequent in populations of African origin, does show signs of functional significance, both in terms of its affinity for the A subunit[12] and genetic susceptibility to thrombosis and MI.[12,13] In addition, each of these lineages is associated with a distinct haplotype (*TTAC-Arg* and *CCGT-His*) in the region immediately upstream of exon 1, a region likely to contain regulatory sequences for the *F13B* gene. Two SNPs (1617 and 1835, Supplementary Table 2) in the upstream region create putative transcription factor-binding sites, approximately 1 kb upstream of the first exon. In this respect, it has been shown that haplotype polymorphism in the 5′ promoter region can significantly influence gene expression, even when variation at individual SNP loci does not appear to do so.[33] In addition, some of the population variation in FXIII activity is due to genetic variation at the B subunit, despite the fact that it has no known catalytic activity. Protein allele-2 is associated with the highest activity, followed by allele-1 and then allele-3.[11] Biochemical data show that *Arg95* is associated with increased subunit dissociation,[12] consistent with our identification of this variant with protein allele-2, the most active form. Protein allele-3, which we identify as the *IntK* splice variant, exhibits the lowest activity.

The 29756-G splice variant in intron K (*IntK + ve*), located through a cDNA clone isolated in a Japanese

laboratory, is also potentially of functional significance, as it gives rise to an alternative carboxy terminus and may be truncated at the 5′ end. However, if the protein is not truncated then the effect is limited to the carboxy terminus and presumably does not influence the sushi structure of the protein.

Whether individuals who are homozygous for this polymorphism also produce 'normal' FXIIIB is not known. However, an indication may be obtained from the numerous electrophoretic studies of the locus.[5,8,34,35] Different separation methods (agarose electrophoresis, isoelectrophoretic focussing and combinations thereof) have been used, but all show allele 1–3 heterozygotes, which are recognizably distinct from allele-3 homozygotes. This would tentatively suggest that allele-3 homozygotes probably do not produce much, if any, of the allele-1 protein product. However, this interpretation is not definitive and is dependent on the sensitivity of the staining method used and the complex interactions of heterogeneous gene products on native protein gels.

Previous studies have shown association of *F13B H95R* variant with MI[13] and thrombosis.[12] Our case–control analysis suggests that the *IntK* polymorphism is associated with modified susceptibility or survivorship in a sample of CABG patients with severe atherosclerotic disease and/or MI. Although the sample size is relatively small, a power calculation using the Genetic Power Calculator (http://pngu.mgh.harvard.edu/~purcell/gpc/)[36] suggests that the minor allele frequency and effect magnitude (Odds ratio = 1.9) give >80% power with the sample sizes analyzed. Neither *F13A Val34Leu* nor *F13B His95Arg* showed association with CAD in our data. However, Haplo Stats analyses suggested a role for the *His/IntK + ve* haplotype.

Previous work has shown that concentration of FXIIIA subunit may be associated with infarct size and poor survival from ischemic stroke,[37] another atherosclerotic condition. This is consistent with our result that the *F13B IntK* splice variant is overrepresented in our CABG cohort, although it is difficult to explain in the context of low susceptibility to infarct yet high frequency of *F13B IntK* in Asian populations. Other genetic factors may account for this effect.

Regarding natural selection, a number of scenarios could explain the results obtained from tests for departure from neutral expectations. First, it may be that this region has not experienced a selective event and the observed high F_{ST} values are the result of complex demographic histories and a low local recombination rate. However, the excess of alleles with large EHH in the Yorubans argues against demographic scenarios. Alternatively, this region might have experienced a selective event, but in the sufficiently distant past and without fixation occurring, thus allowing enough time for recombination and mutation accumulation to reduce the power of allele frequency tests, such as Tajima's D and Fu and Li's tests, and LD-based tests such as those used in this study. Otherwise, a changing selection regime acting on standing molecular variation might have given rise to the observed patterns. Computer simulations[38] of the effect of natural selection on standing genetic variation, in which an existing allele becomes favorable due to environmental change, demonstrated that a broad range of results are possible using standard tests for departure from neutrality, depending on the frequency of the allele under positive selection. In summary, although allele frequencies are significantly differentiated between Yoruba and non-African populations, as well as in three-way comparisons, there is currently no conclusive evidence of selection acting in non-African populations. However, there are a significantly large number of alleles with extensive EHH in Yoruba, indicative of a partial selective sweep linked to the *95Arg* allele.

As genetic variation at the genes encoding both subunits contributes to the activity of plasma factor XIII,[11] natural selection on fibrin crosslinking activity could potentially influence either locus. A wide variety of selection pressures may impinge on the coagulation cascade. For example, hematophagous (blood-drinking) organisms typically produce anticoagulant proteins, which maintain blood in a liquid state during feeding. However, only one species, the Amazonian leech *Haementeria ghilianii*, is known to produce an anticoagulant that inhibits factor XIIIa.[39,40] This species is of limited geographic distribution and is therefore unlikely to have contributed to the worldwide geographic patterns observed at the *F13B* locus. However, the possibility remains that modification of FXIII activity may compensate for the attenuation of some other coagulation factors (eg, thrombin, factor Xa), which are more typically targeted by the anticoagulants of hematophagic organisms.[40]

A number of pathogenic agents (eg, hemorrhagic fevers, Gram-negative bacteria) may cause disseminated intravascular coagulation (DIC) in severely affected individuals.[41,42] In addition, *in vitro* infection of monocytes by *Yersinia pestis*, the etiologic agent of plague, causes a downregulation of thrombomodulin, which may play a role in DIC.[43] Interestingly, the experimental depletion of factor XIII can prevent organ damage during DIC in a rabbit model,[44] suggesting that the fibrinolytic system is capable of dissipating clots in the absence of fibrin crosslinking. A genetically determined modification in fibrin crosslinking activity may therefore influence an individual's chances of surviving an epidemic.

FXIII B-subunit antigen levels also correlate strongly with the biochemical and physical characteristics of type II diabetes[45] and cardiovascular disease.[46] The functional significance of the former association is not clear. However, it suggests that the association of *F13B* variation with CAD may involve metabolic or coagulation pathways. In addition, healthy relatives of patients with cardiovascular

disease exhibit altered fibrin clot structure,[47] which is potentially linked with FXIII activity. Although modified levels of a serum protein may merely be an effect, rather than a cause, of a condition, the association of *F13B* genetic variation with a cardiovascular disease phenotype (present study) suggests a more causative effect. In terms of natural selection, the relationship between FXIII B-subunit levels and type II diabetes raises the possibility of a role for the locus in metabolism and natural selection thereon.

In conclusion, two main lines of evidence suggest that genetic diversity at the *F13B* locus has been influenced by local natural selection. The first is the high degree of population subdivision exhibited by genetic variants that are functionally important in terms of observed protein chemistry (A–B subunit disassociation) and disease susceptibility to coronary heart disease and venous thrombosis (*His95Arg* and *IntK*). Second, analyses of HapMap data provide evidence for selection in African populations. In light of the multiple potential causes of selection and the geographic patterns of genetic variability observed at the *F13B* locus, it is unlikely that the genetic variation of the kind seen here is the result of a single cause and it may instead reflect the complex interaction of multiple selection, and possibly demographic, events over time and space.

Acknowledgements

We thank Ryk Ward (Oxford University), Manfred Kayser (MPI, Leipzig), Lutz Röwer (Humboldt-Universität, Berlin), Ilyas Kamboh (University of Pittsburgh) and numerous volunteers for the provision of samples for this study. RAS Ariëns (University of Leeds) kindly provided details of some of the primers and PCR conditions used in this study. The technical assistance of Karsten Schwarz, Birgit Nickel and Barbara Höffner (MPI, Leipzig) and sample collection by Edel Duggan (TCD, Ireland) are gratefully acknowledged. This project was supported by funds from the Max Planck Society, Germany, Hitachi Europe Ltd and the Higher Education Authority Program for Research in Third Level Institutions (PRTLI), Ireland.

Conflict of interest

The authors have no conflict of interest to declare.

References

1 Adany R, Bardos H: Factor XIII subunit A as an intracellular transglutaminase. *Cell Mol Life Sci* 2003; **60**: 1049–1060.
2 Ariens RA, Lai TS, Weisel JW, Greenberg CS, Grant PJ: Role of factor XIII in fibrin clot formation and effects of genetic polymorphisms. *Blood* 2002; **100**: 743–754.
3 Hashiguchi T, Saito M, Morishita E, Matsuda T, Ichinose A: Two genetic defects in a patient with complete deficiency of the b-subunit for coagulation factor XIII. *Blood* 1993; **82**: 145–150.
4 Saito M, Asakura H, Yoshida T, Ito K, Okafuji K, Matsuda T: A familial factor XIII subunit B deficiency. *Br J Haematol* 1990; **74**: 290–294.
5 Board PG: Genetic polymorphism of the B subunit of human coagulation factor XIII. *Am J Hum Genet* 1980; **32**: 348–353.
6 Nakamura S, Abe K: Genetic polymorphism of coagulation factor XIIIB subunit in Japanese. *Ann Hum Genet* 1982; **46**: 203–207.
7 Kera Y, Nishimukai H, Yamasawa K: Genetic polymorphism of the B subunit of human coagulation factor XIII: another classification. *Hum Genet* 1981; **59**: 360–364.
8 Board PG: Genetic heterogeneity of the B subunit of coagulation factor XIII: resolution of type 2. *Ann Hum Genet* 1984; **48**: 223–228.
9 Kamboh MI, Ferrell RE: Genetic studies of low abundance human plasma proteins. II. Population genetics of coagulation factor XIIIB. *Am J Hum Genet* 1986; **39**: 817–825.
10 Roychoudhury AK, Nei M: *Human Polymorphic Genes: World Distribution*. New York: Oxford University Press, 1988.
11 Saha N, Aston CE, Low PS, Kamboh MI: Racial and genetic determinants of plasma factor XIII activity. *Genet Epidemiol* 2000; **19**: 440–455.
12 Komanasin N, Catto AJ, Futers TS, van Hylckama Vlieg A, Rosendaal FR, Ariens RA: A novel polymorphism in the factor XIII B-subunit (His95Arg): relationship to subunit dissociation and venous thrombosis. *J Thromb Haemost* 2005; **3**: 2487–2496.
13 Reiner AP, Heckbert SR, Vos HL *et al*: Genetic variants of coagulation factor XIII, postmenopausal estrogen therapy, and risk of nonfatal myocardial infarction. *Blood* 2003; **102**: 25–30.
14 Thompson JD, Higgins DG, Gibson TJ: CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 1994; **22**: 4673–4680.
15 Rozas J, Rozas R: DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. *Bioinformatics* 1999; **15**: 174–175.
16 Stephens M, Smith NJ, Donnelly P: A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet* 2001; **68**: 978–989.
17 Weir BS: *Genetic Data Analysis II*. Sinauer Associates: Sunderland MA, 1996.
18 Weir BS, Cockerham CC: Estimating F-statistics for the analysis of population-structure. *Evolution* 1984; **38**: 1358–1370.
19 Tang K, Thornton KR, Stoneking M: A new approach for using genome scans to detect recent positive selection in the human genome. *PLoS Biol* 2007; **5**: e171.
20 Hughes DA, Tang K, Strotmann R *et al*: Parallel selection on TRPV6 in human populations. *PLoS ONE* 2008; **3**: e1686.
21 Raymond M, Rousset F: Genepop (version-1.2) – population-genetics software for exact tests and ecumenicism. *J Hered* 1995; **86**: 248–249.
22 Dean AG, Dean JA, Burton AH, Dicker RC: Epi Info™: a general purpose microcomputer program for health information systems. *Am J Prev Med* 1991; **7**: 178–182.
23 Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA: Score tests for association between traits and haplotypes when linkage phase is ambiguous. *Am J Hum Genet* 2002; **70**: 425–434.
24 Thierry-Mieg D, Thierry-Mieg J: AceView: a comprehensive cDNA-supported gene and transcripts annotation. *Genome Biol* 2006; **7** (Suppl 1: S12): 11–14.
25 Andolfatto P: Adaptive hitchhiking effects on genome variability. *Curr Opin Genet Dev* 2001; **11**: 635–641.
26 Bowcock AM, Kidd JR, Mountain JL *et al*: Drift, admixture, and selection in human evolution: a study with DNA polymorphisms. *Proc Natl Acad Sci USA* 1991; **88**: 839–843.
27 Beaumont MA, Balding DJ: Identifying adaptive genetic divergence among populations from genome scans. *Mol Ecol* 2004; **13**: 969–980.
28 Cavalli-Sforza LL: Population structure and human evolution. *Proc R Soc Lond B Biol Sci* 1966; **164**: 362–379.
29 Sabeti PC, Reich DE, Higgins JM *et al*: Detecting recent positive selection in the human genome from haplotype structure. *Nature* 2002; **419**: 832–837.
30 Murray A, Donger C, Fenske C *et al*: Splicing mutations in KCNQ1: a mutation hot spot at codon 344 that produces in frame transcripts. *Circulation* 1999; **100**: 1077–1084.

31 Takada D, Ezura Y, Ono S *et al*: Apolipoprotein H variant modifies plasma triglyceride phenotype in familial hypercholesterolemia: a molecular study in an eight-generation hyperlipidemic family. *J Atheroscler Thromb* 2003; **10**: 79–84.
32 Iwao M, Morisaki H, Morisaki T: Single-nucleotide polymorphism g.1548G>A (E469K) in human ICAM-1 gene affects mRNA splicing pattern and TPA-induced apoptosis. *Biochem Biophys Res Commun* 2004; **317**: 729–735.
33 Drysdale CM, McGraw DW, Stack CB *et al*: Complex promoter and coding region beta 2-adrenergic receptor haplotypes alter receptor expression and predict *in vivo* responsiveness. *Proc Natl Acad Sci USA* 2000; **97**: 10483–10488.
34 Kamboh MI: Heterogeneity of factor XIIIB: A new method for the determination of factor XIIIB phenotypes by isoelectric focusing in 6 M Urea. *Electrophoresis* 1985; **6**: 185–186.
35 Leifheit HJ, Cleve H: Analysis of the genetic polymorphism of coagulation factor XIIIB (FXIIIB) by isoelectric focusing. *Electrophoresis* 1988; **9**: 426–429.
36 Purcell S, Cherny SS, Sham PC: Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics* 2003; **19**: 149–150.
37 Kohler HP, Ariens RA, Catto AJ *et al*: Factor XIII A-subunit concentration predicts outcome in stroke subjects and vascular outcome in healthy, middle-aged men. *Br J Haematol* 2002; **118**: 825–832.
38 Przeworski M, Coop G, Wall JD: The signature of positive selection on standing genetic variation. *Evolution* 2005; **59**: 2312–2323.
39 Finney S, Seale L, Sawyer RT, Wallis RB: Tridegin, a new peptidic inhibitor of factor XIIIa, from the blood-sucking leech Haementeria ghilianii. *Biochem J* 1997; **324** (Part 3): 797–805.
40 Ciprandi A, Horn F, Termignoni C: Saliva of hematophagous animals: source of new anticoagulants. *Revista Brasileira de Hematologia e Hematerapia* 2003; **24**: 250–262.
41 Levi M, van der Poll T: Coagulation in sepsis: all bugs bite equally. *Crit Care* 2004; **8**: 99–100.
42 Levi M: Current understanding of disseminated intravascular coagulation. *Br J Haematol* 2004; **124**: 567–576.
43 Das R, Dhokalia A, Huang XZ *et al*: Study of proinflammatory responses induced by *Yersinia pestis* in human monocytes using cDNA arrays. *Genes Immun* 2007; **8**: 308–319.
44 Lee SY, Chang SK, Lee IH, Kim YM, Chung SI: Depletion of plasma factor XIII prevents disseminated intravascular coagulation-induced organ damage. *Thromb Haemost* 2001; **85**: 464–469.
45 Mansfield MW, Kohler HP, Ariens RA, McCormack LJ, Grant PJ: Circulating levels of coagulation factor XIII in subjects with type 2 diabetes and in their first-degree relatives. *Diabetes Care* 2000; **23**: 703–705.
46 Warner D, Mansfield M, Grant PJ: Coagulation factor XIII levels in UK Asian subjects with type 2 diabetes mellitus and coronary artery disease. *Thromb Haemost* 2001; **86**: 1117–1118.
47 Mills JD, Ariens RA, Mansfield MW, Grant PJ: Altered fibrin clot structure in the healthy relatives of patients with premature coronary artery disease. *Circulation* 2002; **106**: 1938–1942.

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