Original Paper

Human Heredity

Hum Hered 2004;57:142–155
DOI: 10.1159/000079245

Received: November 3, 2003
Accepted after revision: April 6, 2004

# Evidence of Admixture from Haplotyping in an Epidemiological Study of UK Caucasian Males: Implications for Association Analyses

Xiao-he Chen[a,1] Santiago Rodríguez[a,1] Emma Hawe[b] Philippa J. Talmud[b] George J. Miller[c] Peter Underhill[d] Stephen E. Humphries[b] Ian N.M. Day[a]

[a]Human Genetics Division, School of Medicine, University of Southampton, Southampton General Hospital, Southampton; [b]Centre for the Genetics of Cardiovascular Disease, British Heart Foundation Laboratories, Royal Free and University College London Medical School, and [c]Medical Research Council Cardiovascular Research Group, Wolfson Institute of Preventive Medicine, Barts and the London Queen Mary's School of Medicine and Dentistry, London, England; [d]Department of Genetics, Stanford, Calif., USA

**Key Words**

Genetic association studies · Population subdivision · Stratification · Y chromosome · Autosomal SNPs · Haplotypic analysis

**Abstract**

***Objective:*** Cohort and case-control genetic association studies offer the greatest power to detect small genotypic influences on disease phenotypes, relative to family-based designs. However, genetic subdivisions could confound studies involving unrelated individuals, but the topic has been little investigated. We examined geographical and interallelic association of SNP and microsatellite haplotypes of the Y chromosome, of regions of chromosome 11, and of autosomal SNP genotypes relevant to cardiovascular risk traits in a UK-wide epidemiological survey. ***Results:*** We show evidence (p = 0.00001) of the Danelaw history of the UK, marked by a two-fold excess of a Viking Y haplotype in central England. We also found evidence for a (different) single-centre geographical over-representation of one haplotype, both for *APOC3-A4-A5* and for *IGF2*. The basis of this remains obscure but neither reflect genotyping error nor correlate with the phenotypic associations by centre of these markers. A panel of SNPs relevant to cardiovascular risks traits showed neither association with geographical location nor with Y haplotypes. ***Conclusion:*** Combinations of Y haplotyping, autosomal haplotyping, and genome-wide SNP typing, taken together with phenotypic2 associations, should improve epidemiological recognition and interpretation of possible confounding by genetic subdivision.

## Introduction

Cohort and case-control association studies are powerful tools for the search for disease genes. These approaches are based on establishing a significantly increased or decreased occurrence of a marker allele in correlation with a disease trait [1]. A statistically significant association between the marker locus and the trait is

[1] Equal authorship.

KARGER

Fax + 41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

© 2004 S. Karger AG, Basel
0001–5652/04/0573–0142$21.00/0

Accessible online at:
www.karger.com/hhe

Dr. Santiago Rodríguez
Human Genetics Division, School of Medicine
University of Southampton, Duthie Building (MP808)
Southampton General Hospital, Tremona Road, Southampton SO16 6YD (UK)
Tel. +44 23 8079 4141, Fax +44 23 8079 4264, E-Mail santi@soton.ac.uk

usually considered to imply close physical linkage with the disease locus [2], although only family-based association methods, not cohort or case-control methods, prove this. The efficiency of methods based on unrelated subjects to map disease genes by allelic association is from 3/2 to 6-fold greater than in family-based procedures [3]. In addition, it has been demonstrated that the power of association studies can be much greater than that of linkage studies [4]. This higher power, together with other useful characteristics of population-based and case-control association studies [3, 5], has suggested these approaches as the methods of choice for mapping complex-trait loci [3, 4, 6]. However, many of the reports of positive association in the literature cannot be replicated [5], although broad literature metanalyses of genetic association studies indicate that many reports must have a biological basis rather than representing statistical artefacts [7, 8]. A common explanation proposed for spurious associations in population studies is the occurrence of confounding due to genetic subdivisions produced by admixture or assortative mating, although they have rarely been demonstrated as the cause [4, 5, 9]. In addition, it is believed that population stratification to an extent large enough to distort results is unlikely to occur in many realistic situations [10, 11]. Nevertheless, the number of studies directly evaluating for the presence of substructure in genetic epidemiological studies based on unrelated subjects is to date scarce [12].

Generically, the term *genetic subdivision*, also known as population 'substructure' or population 'stratification', refers to the presence of subgroups of individuals in the population who differ systematically in allele frequencies at many loci [9]. Admixed populations are the result of the *admixture* of different parental populations or different ethnic groups [13], whereas *assortative mating* occurs within the population as a result of the separation of individuals into cohabiting subgroups on the basis of cultural, genetic or other differences. One approach to dealing with the problem of confounding due to genetic subdivision or 'stratification' is to match the ethnic backgrounds at the design stage. However, 'cryptic stratification' may remain even when such matching is performed [14, 15]. 'Cryptic stratification' or 'cryptic relatedness' is defined as unobserved ancestral relationships between individual cases and controls who are naively treated as independent in the standard chi square test. Its possible effect on association analyses would be to inflating the effective sample size, thereby increasing the false positive rate, even in the absence of any confounding bias [16]. Other approaches recently suggested are based on the idea that if population substructure affects allele frequencies of the disease locus, then allele frequencies at other genes across the genome should be affected similarly. Thus, population substructure can be detected and it should even be possible to adjust for subdivision, by typing a panel of polymorphic markers that are not linked to the candidate gene under study but whose frequencies would be sensitive to substructure [2, 15, 17]. Analysis of the variation of Y chromosome haplotypes constituted by SNPs and microsatellites, has been reported to be more sensitive to population subdivision compared with analysis of autosomal markers, given the uniparental transmission through the male lineage which leads to a smaller effective population size, and the absence of recombination of the non-recombining region of the Y chromosome [18].

We describe here a preliminary search for the presence of genetic subdivision in a large cohort used for population-based association studies for the search for disease susceptibility genes. Population subdivision was analysed by examining the geographical variation of haplotypes from the Y chromosome and two autosomal regions as constituted by SNPs and microsatellites. We have also examined associations of these haplotypes with a range of autosomal SNPs relevant to cardiovascular disease traits and we have examined geographical stratification of SNP genotype frequencies. The approaches form a framework to test for possible confounding in genetic association studies and confirm the power of multicentre studies, sensitivity of haplotyping and use of Y chromosome for recognising admixture, which can then be controlled for.

## Methods

*DNA Samples*

The individuals analysed in this study were 2,654 unrelated healthy Caucasian men aged 51–62 years of the Northwick Park Heart Study II (NPHSII) cohort as described previously [19]. In brief, NPHSII is a prospective study of middle-aged Caucasian male subjects (mean ± SEM age 56.1 ± 3.5 years) free from symptomatic or electrocardiographic evidence of coronary artery disease at enrolment, and recruited from nine United Kingdom general practices. Among those individuals, 114 were from Aylesbury (Buckinghamshire), 351 from Carnoustie (Tayside Region), 184 from Chesterfield (Derbyshire), 437 from Halesworth (Suffolk), 309 from North Mimms (Hertfordshire), 143 from Harefield (Greater London), 392 from Parkstone (Dorset), 408 from Camberley (Surrey) and 316 from St. Andrews (Fife Region) (fig. 1). Local medical general practitioners collected clinical data and peripheral blood from each individual. Genomic DNAs were extracted from blood samples and equalised to 10 ng/µl as PCR working stock using protocols previously described (www.sgel.humgen.soton.ac.uk).

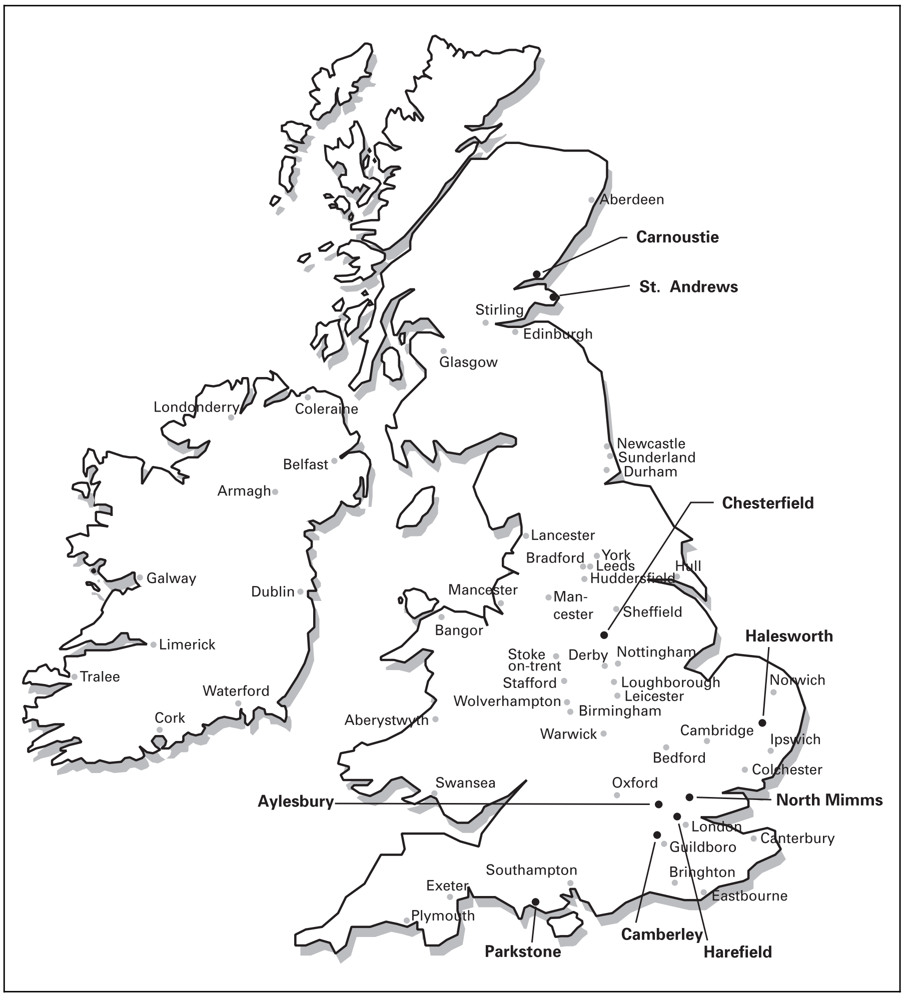

**Fig. 1.** Schematic UK map showing the geographic locations of nine populations in NPHS II cohort. Black markers and their names indicate the locations.

*DNA Markers*

Six Y-specific markers were analysed: two microsatellites (DYS390 and DYS392) and four SNPs (M9, M170, M173 and M223). We also analysed six *IGF2* SNPs located in human chromosome 11p15 (*IGF2* 2482, *IGF2* 2722, *IGF2* 266, *IGF2* 1926, *IGF2* 2207 and *IGF2* 3750) and nine SNPs spanning the *APOC3-A4-A5* gene cluster in 11q (S19W, –1131T→C, intergenic T→C, Q360H, T347S, –2845T→G, –482C→T, 1100C→T and 3238C→G). 55 SNPs located in different genes relevant to cardiovascular traits were also analysed.

*Genotyping*

The two Y microsatellites were amplified by PCR. Y-SNP M9 was amplified by PCR adopting the Amplification Refractory Mutation System (ARMS) Y-SNPs M170, M173 and M223 were amplified using allele-specific tetra-primer ARMS Sequences of the Y polymorphisms, sequences of the primers and PCR thermal cycling conditions for each marker are listed at www.sgel.humgen.soton.ac.uk. Amplification products were resolved by Microplate Array Diagonal Gel Electrophoresis (MADGE) as previously described [22–24].

The six *IGF2* SNPs were genotyped by ARMS-MADGE as previously described [25]. The nine *APOC3-A4-A5* SNPs were genotyped as previously described [26]. The name of each one of the 55 SNPs relevant to cardiovascular traits and the references indicating the details for their genotyping are available at www.sgel.humgen.soton.ac.uk.

*Statistical Analysis*

Y chromosome haplotype frequencies were directly established from the counts of the Y markers analysed. On account of potential inconsistencies using haplotype inference in unrelated subjects, two different haplotype algorithms were used to infer haplotype frequencies (absolutely or to some predefined level of confidence) from the *IGF2* gene region and the *APOC3-A4-A5* gene cluster: Partition-Ligation-Expectation-Maximization (PL-EM) [27] and PHASE version 2.0 [28]. PL-EM is a deterministic maximum likelihood algorithm. It is based on breaking down the marker loci into stretches of 'atomistic' units and then using the EM algorithm to construct haplotypes for each unit. Afterwards, a ligation process of adjacent partial haplotypes is performed by using the EM algorithm again until the complete phase is determined. It has been claimed that this approach performs better than both the conventional EM algorithm and an enhanced version of PHASE [27]. The PHASE program version 2 uses a Bayesian algorithm based on Gibbs sampling, a type of Markov-chain Monte Carlo algorithm [29]. It has been suggested that this approach outperforms the previously published Bayesian methods HAPLOTYPER [30] and a modified Stephens-Smith-Donnelly method [31]. The predefined level of confidence was set at ≥90% and the numbers of iterations and burns-in performed were 10,000, each iteration consisting of performing 100 steps through the Markov chain.

The pairwise Slatkin linearized $F_{ST}$ between the 9 locations analysed were calculated from the observed haplogroups for the 4 Y-SNPs. Slatkin linearized $F_{ST}$ can be used to estimate effective migration rates or times since population divergence and is expressed as $F_{ST} = (\bar{t} - \bar{t}_0)/\bar{t}$, were $\bar{t}_0$ is the average coalescence time of two copies of a gene drawn from the same population and $\bar{t}$ is the average coalescence time of two copies of a gene drawn from the collection of populations [32]. The null distribution of pairwise $F_{ST}$ values under the hypothesis of no difference between the locations was obtained by permuting haplogroups between locations. The p value of the test is the proportion of permutations leading to a $F_{ST}$ value larger or equal to the observed one. Both the pairwise $F_{ST}$ values and associated p values based on 10,000 permutations were calculated by the program Arlequin, version 2000 [33].

For examination of genetic subdivision between locations, the geographical variation of counts of the commonest haplotypes for three chromosomal regions (Y chromosome, *IGF2* gene region and *APOC3-A4-A5* gene cluster) was analysed by contingency-table chi square test (Pearson $\chi^2$). For this purpose, two programs were used: CONTING ver 2.61 [34] to compute asymptotic p values, and Monte Carlo R × C contingency table test ver 2.1 (http://engels.genetics.wisc.edu/pstat/) to compute p values based on 50,000 permutations. The same approaches were used for analysing the geographical variation of 55 autosomal SNPs relevant to cardiovascular traits, based on the differences in genotype distribution by location. Finally, we analysed the association of haplotypes (Y chromosome and two regions of chromosome 11) with autosomal SNPs relevant to cardiovascular traits in order to test whether the subdivisions found for both chromosomes correlate with variation of autosomal SNPs in each location. On account of the large number of comparisons involved in the analyses of the 55 autosomal SNPs, it is likely that at least one haplotype or genotype would be nominally significant even if there were no real association. The usual experimental error rate of 0.05 for multiple comparisons was controlled for using the Bonferroni method [35, 36]. Two levels of significance were established: one conservative ($\alpha_c$) and one under conservative ($\alpha_{uc}$). The conservative level of significance after Bonferroni correction was $\alpha_c = 0.00091$ (0.05 divided by 55, the total number of tests). The under conservative correction adopted was $\alpha_{uc} = 0.0016$ (0.05 divided by 31, 31 being the total number of different genes relevant to cardiovascular traits analysed) which took into account that the SNPs within a gene are non-independent because they are in partial or substantial linkage disequilibrium.

## Results

### *Differentiation between Pairs of Locations Based on Y Haplogroups*

Table 1 shows the pairwise Slatkin linearized $F_{ST}$ values and their probabilities for all pairs of locations in NPHSII, obtained from the Y haplogroups derived from the 4 Y SNPs analysed. The majority of pairwise comparisons (26 out of 36) are non-significant, with $F_{ST}$ values ranging from 0.000 to 0.015. Five of the pairwise comparisons involving Scottish locations (Carnoustie and Saint Andrews) are significant, with $F_{ST}$ values ranging from 0.011 to 0.062. Notably, six out of eight pairwise comparisons involving Chesterfield are significant, with a mean $F_{ST}$ of 0.041 ± 0.005. This suggests that this location in central England is the most differentiated location of the nine analysed in NPHSII, on the basis of Y haplogroups.

### *Geographical Variation of Haplotypes of the Y Chromosome*

The counts of the three commonest haplogroups for the five Y SNPs, accounting for 90.2% of total haplogroups, were compared across locations. The results obtained show that there are significant differences among locations (p = 0.002). These differences are mainly attributable to a single location, Chesterfield, since its exclusion renders a non-significant difference across locations (p = 0.116). Indeed, comparison of haplogroup counts in Chesterfield with those of the remaining locations pooled all together indicates the significant presence of subdivision (p = 0.0002). The main contribution to $\chi^2$ is for haplogroup CCAC (markers in the order M9, M170, M173 and M223), named 0100 (0 indicating ancestral and 1 derived) or I*(xM223) [37] which is more than two-fold over-represented in Chesterfield (23.61%) compared with the remaining locations (10.91%).

A similar result was obtained when analysing the four Y SNPs in conjunction with the two Y microsatellites. Differences between the seven most frequent haplotypes (accounting for 67.4% of total haplotypes) across all locations were significant (p = 0.001) (table 2). By contrast, non-significant differences were seen when excluding Chesterfield from the analysis (p = 0.135). As in the case of the four SNPs, significant differences were observed when haplotype counts for Chesterfield were compared with the counts for the remaining locations (p = 0.00001). One of the main contributions to $\chi^2$ is for haplotype 22-11-CCAC (table 2), also over-represented more than two-fold in Chesterfield (10.53%) when compared with the remaining locations (5.24%).

**Table 1.** Magnitude (Slatkin linearized $F_{ST}$ values) and significance (in parentheses) of haplogroup differentiation between all possible pairs of locations in NPHSII when the 4 Y-SNPs are considered

| | CA | SA | CH | HA | NM | AY | HR | CM | PA |
|---|---|---|---|---|---|---|---|---|---|
| CA | – | | | | | | | | |
| SA | 0.028<br>**(0.004)** | – | | | | | | | |
| CH | 0.062<br>**(0.0002)** | 0.000<br>(0.327) | – | | | | | | |
| HA | 0.000<br>(0.335) | 0.011<br>**(0.040)** | 0.036<br>**(0.002)** | – | | | | | |
| NM | 0.004<br>(0.134) | 0.009<br>(0.082) | 0.027<br>**(0.010)** | 0.000<br>(0.324) | – | | | | |
| AY | 0.019<br>**(0.034)** | 0.000<br>(0.869) | 0.004<br>(0.222) | 0.003<br>(0.200) | 0.003<br>(0.211) | – | | | |
| HR | 0.000<br>(0.753) | 0.015<br>(0.093) | 0.043<br>**(0.016)** | 0.000<br>(0.674) | 0.000<br>(0.495) | 0.007<br>(0.207) | – | | |
| CM | 0.000<br>(0.557) | 0.016<br>**(0.025)** | 0.042<br>**(0.0008)** | 0.000<br>(0.668) | 0.000<br>(0.574) | 0.008<br>(0.111) | 0.000<br>(0.823) | – | |
| PA | 0.002<br>(0.184) | 0.009<br>(0.063) | 0.033<br>**(0.003)** | 0.000<br>(0.743) | 0.000<br>(0.365) | 0.002<br>(0.243) | 0.000<br>(0.726) | 0.000<br>(0.534) | – |

In bold are the significant p values. Centres are ordered in keeping with their geographic location, from the Northest (Carnoustie) to the Southest (Parkstone). CA is the abbreviation for Carnoustie, SA for St. Andrews, CH for Chesterfield, HA for Halesworth, NM for North Mimms, AY for Aylesbury, HR for Harefield, CM for Camberley and PA for Parkstone.

### *Geographical Variation of Haplotypes of Chromosome 11*

We have tested for the possibility of subdivision among locations by analysis of two independent sets of linked SNPs on human chromosome 11: one constituted by six SNPs spanning the *IGF2* gene region on the short arm of chromosome 11 and the other including nine SNPs at the *APOC3-A4-A5* gene cluster on the long arm of the same chromosome. The subdivision detected with Y chromosome haplotypes was not detected with autosomal haplotypes. Nevertheless, we found evidence of geographical subdivision for each autosomal region with over-representation of one single haplotype mainly in one single location, different for each region. Thus, highly significant differences ($p < 10^{-25}$) were detected among locations both with PL-EM and with PHASE ver 2 on the basis of the five commonest haplotypes of the six SNPs at the *IGF2* region (accounting for 75.2% of total haplotypes) (see the PL-EM results in table 3). These differences are, in most part, due to a single haplotype accounting for 5.7% of total haplotypes (111111, table 3), since the significant differences among locations reduce to p = 0.03 in the PL-EM analysis and p = 0.0001 in PHASE ver 2 when this haplotype is excluded from the contingency table. The residual significance appears to have a different basis depending on the algorithm used to estimate the haplotype frequencies. In the case of PL-EM it is due to the absence of the 121211 haplotype in Harefield, the expected number of haplotypes being 6 in this location. In the case of PHASE ver 2 it is due to an over-representation of haplotype 121111 in North Mimms, this over-representation being undetected by PL-EM. A more detailed analysis reveals that haplotype 111111 is not distributed uniformly across locations, its frequency being much greater in North Mimms than in the remaining locations. It represents 24.4% of the total haplotypes of North Mimms, the percentage of this haplotype in the remaining locations ranging from 1.5 to 11%. This over-representation is accompanied by an under-representation of the commonest haplotype for the six SNPs (2111111, table 3). The frequency of 211111 in North Mimms is 18.3%, ranging in the remaining locations from 25.9 to 41%. We considered the possibility that the subdivision seen in North Mimms, when compared with the remaining locations, could reflect to some extent the presence of genotyping or data handling error in North

**Table 2.** Y haplotype frequency distribution of the seven commonest haplotypes across geographical locations in the UK

| | Haplotype | Geographic Locations | | | | | | | | | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | CA | SA | CH | HA | NM | AY | HR | CM | PA | |
| 1 | 24-13-GACC | 98 (39.7) | 44 (31.0) | 31 (23.3) | 98 (33.7) | 69 (34.0) | 35 (35.4) | 13 (21.3) | 44 (26.0) | 90 (35.3) | 522 |
| 2 | 23-13-GACC | 23 (9.3) | 19 (13.4) | 28 (21.1) | 48 (16.5) | 22 (10.8) | 18 (18.2) | 5 (8.2) | 25 (14.8) | 44 (17.3) | 232 |
| 3 | 25-13-GACC | 21 (8.5) | 10 (7.0) | 3 (2.3) | 20 (6.9) | 5 (2.5) | 6 (6.1) | 4 (6.6) | 16 (9.5) | 15 (5.9) | 100 |
| 4 | 22-11-CCAC | 11 (4.5) | 11 (7.7) | 14 (10.5) | 11 (3.8) | 11 (5.4) | 8 (8.1) | 0 (0.0) | 8 (4.7) | 17 (6.7) | 91 |
| 5 | 23-11-GACC | 6 (2.4) | 8 (5.6) | 10 (7.5) | 9 (3.1) | 7 (3.4) | 4 (4.0) | 1 (1.6) | 6 (3.6) | 4 (1.6) | 55 |
| 6 | 25-14-GACC | 12 (4.9) | 5 (3.5) | 1 (0.8) | 6 (2.1) | 7 (3.4) | 0 (0.0) | 0 (0.0) | 5 (3.0) | 5 (2.0) | 41 |
| 7 | 24-14-CCAC | 6 (2.4) | 1 (0.7) | 8 (6.0) | 6 (2.1) | 7 (3.4) | 1 (1.0) | 2 (3.3) | 3 (1.8) | 4 (1.6) | 38 |
| | Others | 70 (28.3) | 44 (31.0) | 38 (28.6) | 93 (32.0) | 75 (36.9) | 27 (27.3) | 36 (59.0) | 62 (36.7) | 76 (29.8) | 521 |
| | TOTAL | 247 | 142 | 133 | 291 | 203 | 99 | 61 | 169 | 255 | 1,600 |

Percentages are shown in parentheses. Y markers are in the order DYS390, DYS392, M9, M170, M173 and M223. Centres are ordered in keeping with their geographic location, from the Northest (Carnoustie) to the Southest (Parkstone). CA is the abbreviation for Carnoustie, SA for St. Andrews, CH for Chesterfield, HA for Halesworth, NM for North Mimms, AY for Aylesbury, HR for Harefield, CM for Camberley and PA for Parkstone.

Pearson $\chi^2 = 83.90$; df = 48; p = 0.001. Pearson $\chi^2$ (without CH) = 52.20; df = 42; p = 0.135.

**Table 3.** Haplotype frequency distribution across geographical locations in the UK of the five commonest haplotypes for the six SNPs spanning the *IGF2* gene region deduced by the PL-EM algorithm

| | Haplotype | Geographic Locations | | | | | | | | | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | CA | SA | CH | HA | NM | AY | HR | CM | PA | |
| 1 | 211111 | 108 (39.4) | 55 (40.4) | 41 (41.0) | 172 (36.9) | 45 (18.3) | 38 (38.0) | 44 (25.9) | 122 (38.1) | 180 (34.5) | 805 |
| 2 | 111211 | 3 (15.7) | 21 (15.4) | 14 (14.0) | 63 (13.5) | 20 (8.1) | 15 (15.0) | 23 (13.5) | 55 (17.2) | 86 (16.5) | 340 |
| 3 | 121111 | 33 (12.0) | 18 (13.2) | 7 (7.0) | 62 (13.3) | 34 (13.8) | 6 (6.0) | 20 (11.8) | 46 (14.4) | 81 (15.5) | 307 |
| 4 | 121211 | 19 (6.9) | 12 (8.8) | 6 (6.0) | 31 (6.7) | 6 (2.4) | 6 (6.0) | 0 (0.0) | 29 (9.1) | 44 (8.4) | 172 |
| 5 | 111111 | 4 (1.5) | 9 (6.6) | 2 (2.0) | 32 (6.9) | **60 (24.4)** | 4 (4.0) | 19 (11.2) | 9 (2.8) | 12 (2.3) | 132 |
| | Others | 67 (24.5) | 21 (15.4) | 30 (30.0) | 106 (22.7) | 81 (32.9) | 31 (31.0) | 64 (37.6) | 59 (18.4) | 119 (22.8) | 578 |
| | Total | 274 | 136 | 100 | 466 | 246 | 100 | 170 | 320 | 522 | 2,334 |

Percentages are shown in parentheses. *IGF2*-markers are in the order *IGF2* 2482, *IGF2* 2722, *IGF2* 266, *IGF2* 1926, *IGF2* 2207 and *IGF2* 3750. CA is the abbreviation for Carnoustie, SA for St. Andrews, CH for Chesterfield, HA for Halesworth, NM for North Mimms, AY for Aylesbury, HR for Harefield, CM for Camberley and PA for Parkstone. 1 represents the common allele and 2 the rare allele.

Pearson $\chi^2 = 248.68$; df = 28; $p < 10^{-35}$.

Mimms leading to the miscalling of allele 2 as allele 1 for SNP *IGF2* 2482. However, the genotypic counts for this SNP in North Mimms are in Hardy-Weinberg equilibrium ($\chi^2 = 0.144$; 1 d.f.; p = 0.704) and retrospective data review confirmed the veracity of relevant genotype calls. Removal of North Mimms from the analysis reduced the significance but not completely (p = 0.000003 in PL-EM and 0.00055 in PHASE ver 2), suggesting that the overall haplotype differences among locations are not all due to the haplotype frequencies observed in North Mimms. The main contributors to these still significant differences among locations are over representations of haplotype 111111 in the other two locations of the South East of England (Harefield, 11.2% and Halesworth, 6.9%). When North Mimms, Harefield and Halesworth are all excluded from the analysis, no significant differences exist between the locations analysed (p = 0.352 in PL-EM and 0.435 in PHASE ver 2).

**Table 4.** Haplotype frequency distribution across geographical locations in the UK of the five commonest haplotypes for the nine SNPs spanning the *APOC3-A4-A5* gene cluster deduced by the PL-EM algorithm

| | Haplotype | Geographic Locations | | | | | | | | | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | CA | SA | CH | HA | NM | AY | HR | CM | PA | |
| 1 | 111111111 | 110 (30.6) | 102 (42.1) | 78 (34.8) | 156 (34.1) | 93 (33.5) | 57 (31.0) | 102 (35.7) | 128 (35.0) | 111 (36.3) | 937 |
| 2 | 112122211 | 9 (8.1) | 25 (10.3) | 27 (12.1) | 33 (7.2) | 26 (9.4) | 14 (7.6) | 20 (7.0) | 33 (9.0) | 29 (9.5) | 236 |
| 3 | 112111111 | 31 (8.6) | 12 (5.0) | 20 (8.9) | 31 (6.8) | 20 (7.2) | **30 (16.3)** | 27 (9.4) | 21 (5.7) | 15 (4.9) | 207 |
| 4 | 111211111 | 32 (8.9) | 10 (4.1) | 12 (5.4) | 29 (6.3) | 11 (4.0) | 11 (6.0) | 32 (11.2) | 28 (7.7) | 15 (4.9) | 180 |
| 5 | 112122111 | 27 (7.5) | 18 (7.4) | 14 (6.3) | 18 (3.9) | 15 (5.4) | 4 (2.2) | 17 (5.9) | 32 (8.7) | 19 (6.2) | 164 |
| | Others | 131 (36.4) | 75 (31.0) | 73 (32.6) | 191 (41.7) | 113 (40.6) | 68 (37.0) | 88 (30.8) | 124 (33.9) | 117 (38.2) | 980 |
| | Total | 360 | 242 | 224 | 458 | 278 | 184 | 286 | 366 | 306 | 2,704 |

Percentages are shown in parentheses. Markers are in the order S19W, –1131T→C, intergenic T→C, Q360H, T347S, –2845T→G, –482C→T, 1100C→T and 3238C→G. CA is the abbreviation for Carnoustie, SA for St. Andrews, CH for Chesterfield, HA for Halesworth, NM for North Mimms, AY for Aylesbury, HR for Harefield, CM for Camberley and PA for Parkstone. 1 represents the common allele and 2 the rare allele.

Pearson $\chi^2 = 68.01$; df = 32; p = 0.0002.

**Table 5.** Autosomal SNPs relevant to cardiovascular traits showing nominal significance across locations before the Bonferroni correction

| Genotype name | Locus name | p |
|---|---|---|
| –574G/C | *GCH1* | 0.05 |
| 493 G→T | *MTP* | 0.04 |
| *IGF2 Alul* | *IGF2* | 0.03 |
| –174 G→C | *IL6* | 0.02 |
| R3611E G→A | *APOB* | 0.02 |
| Q95H | *MTP* | 0.01 |
| *INS Hphl* | *INS* | 0.005 |
| *ε2/3/4* | *APOE* | 0.004 |
| R353Q | *F7* | 0.004 |
| A222V | *MTHFR* | 0.0006 |

The conservative level of significance after Bonferroni correction is $\alpha_c = 0.00091$. The under conservative is $\alpha_{uc} = 0.0016$.

We also tested for Hardy-Weinberg equilibrium for each of the six markers spanning the *IGF2* gene region in each location. The results obtained indicate that only three out of 48 tests show significant differences from Hardy-Weinberg proportions. This low percentage (6%) can be explained by a type-I error given the large number of tests done. This result supports that the low percentage of missing data for each location and each marker is random and argues against the possibility of biased sampling within the locations.

With regard to the *APOC3-A4-A5* gene cluster at 11q, p values of 0.0002 (p L-EM) and 0.0001 (PHASE ver 2) were obtained when comparing the five commonest haplotypes from this cluster, accounting for 66.4% of total haplotypes, across locations. The significant differences are due to a single haplotype accounting for 7.7% of total haplotypes (112111111, table 4). This haplotype is significantly over-represented in Aylesbury (16.3%) when compared with the remaining locations (4.9–9.4%) (table 4). Indeed, the significant differences among locations reduced considerably both when excluding haplotype 112111111 from the analysis (p = 0.05 in PL-EM and p = 0.266 in PHASE ver 2) and when excluding Aylesbury from the analysis (p = 0.04 in PL-EM and p = 0.413 in PHASE ver 2). As in the case of the *IGF2* haplotype variation in North Mimms, the subdivision found in Aylesbury seems not to be an artifactual result caused by genotyping error in the intergenic T→C SNP, since genotypic counts for this SNP in Aylesbury are in Hardy-Weinberg equilibrium ($\chi^2 = 1.985$; 1 d.f.; p = 0.160).

The permutation-based p values for the chi-squared association tests obtained by the program Monte Carlo R × C contingency table test ver 2.1 were in complete agreement in all analyses (Y chromosome, *IGF2* gene and *APOC3-A4-A5* gene cluster) with those asymptotic p values obtained by the program CONTING ver 2.61, with only subtle quantitative differences (table 6).

**Table 6.** Comparison of asymptotic (CONTING ver 2.61) and permutation-based (Monte Carlo R × C contingency table test ver 2.1)

| | p | |
|---|---|---|
| | asymptotic | Monte Carlo |
| Y all locations | 0.0010 | 0.0007 |
| Y 8 locations (excluding CH) | 0.135 | 0.0846 |
| *IGF2* all loc. PL-EM | $2.3 \times 10^{-35}$ | 0 |
| *IGF2* all loc. PHASE 2 | $2.6 \times 10^{-27}$ | 0 |
| *IGF2* all loc. 4 haplot. PL-EM | 0.029 | 0.0038 |
| *IGF2* all loc. 4 haplot. PHASE 2 | 0.0001 | <0.00001 |
| *IGF2* 8 loc. (excluding NM) PL-EM | 0.000003 | <0.00001 |
| *IGF2* 8 loc. (excluding NM) PHASE 2 | 0.00055 | 0.00009 |
| *IGF2* 6 loc. (excluding NM, HR, HA) PL-EM | 0.352 | 0.414 |
| *IGF2* 6 loc. (excluding NM, HR, HA) PHASE 2 | 0.435 | 0.451 |
| *APO* all loc. PL-EM | 0.0002 | 0.00058 |
| *APO* all loc. PHASE 2 | 0.0001 | 0.00076 |
| *APO* all loc. 4 haplot. PL-EM | 0.045 | 0.0418 |
| *APO* all loc. 4 haplot. PHASE 2 | 0.266 | 0.2669 |
| *APO* 8 loc. (excluding AY) PL-EM | 0.039 | 0.0388 |
| *APO* 8 loc. (excluding AY) PHASE 2 | 0.413 | 0.4318 |

p values observed in the analyses performed for each chromosomal region. CH = Chesterfield, NM = North Mimms, HR = Harefield, HA = Halesworth and AY = Aylesbury.

### *Geographical Variation of Autosomal SNPs Relevant to Cardiovascular Traits*

Table 5 shows that 10 out of 55 SNPs have a p value <0.05 whereas 45 do not show significant differences in genotype distribution by location. After Bonferroni correction for multiple tests using $\alpha_{uc}$, one of the results was significant at $p < 0.05$ *(MTHFR)*. With $\alpha_c$, this result is of borderline significance. A detailed analysis of this marker revealed (unpublished data) that the significant difference is caused by differences in allele frequencies in two locations: a higher than average rare allele frequency in Saint Andrews (0.38 (0.34–0.41) vs. 0.31 (0.29–0.32) elsewhere, p = 0.001) and a lower than average frequency in North Mimms (0.24 (0.21–0.29), p = 0.005 vs. others).

### *Association of Haplotypes (Y Chromosome and Two Regions of Chromosome 11) with Autosomal SNPs Relevant to Cardiovascular Traits*

We analysed the association between the Y haplotype over-represented in Chesterfield and each autosomal SNP across each location. The lowest p value observed (p = 0.002) corresponded with the association of this haplotype with *APOB XbaI* genotype in Chesterfield, a value which is not significant after Bonferroni correction neither using $\alpha_c$ nor $\alpha_{uc}$. Similarly, no significant associations after Bonferroni correction were observed either when the association was analysed between the *IGF2* haplotype over-represented in North Mimms and the autosomal SNPs, or when the association between the *APOC3-A4-A5* haplotype over-represented in Aylesbury and the autosomal SNPs was analysed (unpublished data).

## Discussion

In this work we have tested for the occurrence of genetic substructure in NPHSII, a large UK-wide cardiovascular and genetic epidemiological study. Three instances of subdivision were detected on the basis of the variation of haplotypes across the nine locations analysed across the UK. These instances are likely the result of minor population genetic substructure in the UK, given that no location clearly differs from the others for all haplotype distributions analysed. The results presented here suggest an explanation for the historical basis of one of the subdivisions found and have implications for the design of population-based association studies, such that subdivision can be recognised and either avoided or controlled for during analysis.

### *Y Haplotype Subdivision*

Of the three observed subdivisions, one, the I*(xM223) Y-haplotype 22-11-CCAC was found to be over-represented more than two-fold in Chesterfield (central En-

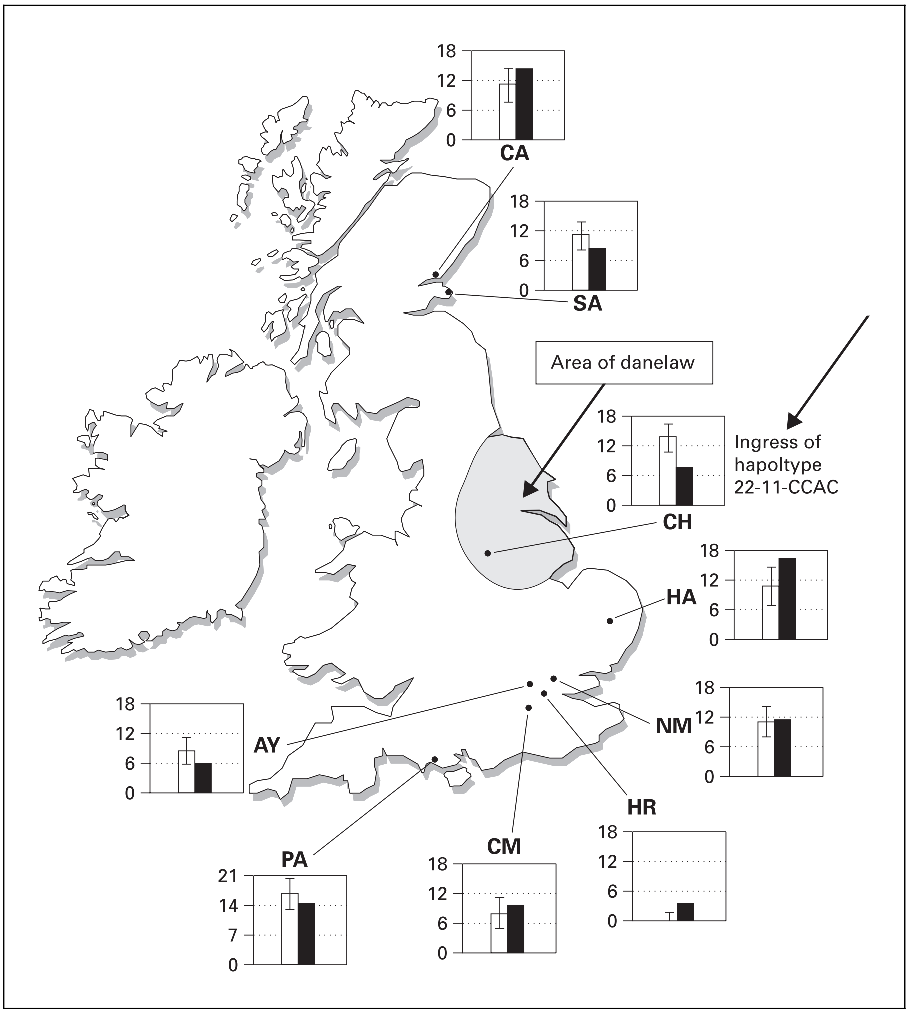

**Fig. 2.** Subdivision found by Y haplotypes in a location of central England, marked by a significant ingress of 22-11-CCAC haplotype (of likely Viking origin) in the area of Danelaw. Black and white bars represent, respectively, the expected and the observed numbers of 22-11-CCAC haplotypes for each location. Error bars represent differences of 1 s.d. either side. CA is the abbreviation for Carnoustie, SA for St. Andrews, CH for Chesterfield, HA for Halesworth, NM for North Mimms, AY for Aylesbury, HR for Harefield, CM for Camberley and PA for Parkstone.

gland) when compared with the remaining locations (fig. 2). Interestingly, the microsatellite features and SNP M9 of this haplotype are identical with the Viking haplogroup 2.47 [38] consisting of allele 22 at DYS390, allele 11 at DYS392 and the ancestral allele at M9 SNP. An excess of Viking Y chromosome haplotype seems then to occur in central England. An over-representation of haplotype 24-14-CCAC, carrying allele C at M9, was also found in Chesterfield. A trend of haplogroup 1 (including allele G at M9 [39]) has been observed [40] in Europe from the lowest frequency in the Near East (1.8% in Turkey) to the highest frequency in the West (98% in Connaught, Ireland). The frequency of allele C at M9 follows the opposite trend, allele C being common in the Near East and rare in the West. The increased presence of M9 allele C in Chesterfield (including both the Viking haplotype and haplotype 24-14-CCAC) may reflect this trend. Chesterfield is the only centre within our study falling within the historical Danelaw region of central England. The most likely origin of the Viking haplotype is Denmark [41, 42], although it has recently been suggested that any Danish Viking influence on the English gene pool may prove difficult to distinguish from Anglo-Saxon influence [43]. We do note that Weale's haplogroup 3 [43], now called R1a1 [37] which is prevalent in Norway and contains allele 25 at DYS390 and allele 11 at DYS392, does *not* form the basis of our observations. The more than two-fold over-representation of a Viking Y haplogroup in Chesterfield (10.53% vs. 4.81% elsewhere), is of a magnitude relevant to complex trait genetics. If this historical admixture marked arrival of autosomal genotypes or haplotypes with functional effects, then studies could be confounded if independent segregation had not since taken place. However, this risk seems unlikely. Firstly, autosomal haplotypes are unlikely to be as differentiated between Denmark and UK as are Y haplotypes. Secondly, the continued segregated existence of Vikings in Chesterfield seems unlikely. Thirdly, we have (both for Chesterfield and for other centres) examined for association of a range of autosomal haplotypes and SNP genotypes with the presumed admixed Y haplotype in other centres and found no evidence to support this. Nevertheless our present observations suggest that the power of Y haplotypes (or for maternal lineages, mitochondrial haplotypes) to detect admixture is greater than either the personal ancestral history that subjects can provide, or those general clinical observations available to the epidemiologist. Y and mitochondrial data should thus be useful adjuncts in controlling population-based genetic epidemiological studies.

### *Autosomal Haplotypes*

We detected a considerable over-representation of the *IGF2* haplotype 111111 in North Mimms when compared with the remaining locations. High frequencies of this haplotype were also found in the other two locations in the South East of England, but were considerably lower than in North Mimms. We found no evidence indicating that this subdivision could be caused by genotyping error but we did not find any particular characteristic of these locations that could account for such subdivision. We also detected over-representation of a single *APOC3-A4-A5* haplotype in Aylesbury when compared with the remaining locations. As in the previous case, the cause remains obscure. The subdivisions found by means of haplotypes in the *IGF2* gene and in the *APOC3-A4-A5* gene cluster, do not seem to be an artefactual result of cells with small numbers in chi-squared contingency tables, given the high counts of the corresponding haplotypes in these locations. They could be caused by unrecognised events in the population history of these locations including genetic drift, migration, stratification or admixture. Given sufficient data concerning autosomal haplotypes, it should in the future be possible to achieve much higher resolution information both for validation of laboratory data, and for recognising possible minor genetic substructure.

Some minor inconsistencies between the two inference programs used to estimate haplotype frequencies were found, namely an over-representation of the *IGF2* haplotype 121111 in North Mimms (detected by PHASE ver 2 and undetected by PL-EM) and minor differences in the numbers of *APOC3-A4-A5* haplotypes deduced by both programs leading to differences in the p values of the Pearson chi-squared contingency tests when comparing the haplotype counts observed across locations. Possible causes of these discrepancies include the fact that the simulations of PHASE are based on coalescent models, which may not be good approximates of human population evolutionary history [44], and the fact that PHASE incorporates the prior knowledge that unresolved haplotypes will tend to be the same, or similar to, known haplotypes [45]. Our results suggest that, in some instances, population subdivisions based on haplotype frequencies estimated from genotypic data could be the result of limitations or errors of a given algorithm, and support the comparison of different algorithms as a way to reduce the likelihood of reporting false subdivisions.

### *Subdivision and Haplotype Frequencies*

Interestingly, the three cases of subdivisions detected involve haplotypes with frequencies slightly higher than 5% in the UK. When considering only more common haplotypes, no significant evidence of subdivision across locations was found for any of the chromosomal regions analysed. These results have implications for association studies. The finding that more common haplotypes do not show subdivision across locations in the UK is in accordance with previous evidence indicating that stratification is a rare cause for confounding in association studies [10, 12, 46]. However, the subdivisions found, which represent less frequent haplotypes, could lead to confounding under particular circumstances. If one of the haplotypes showing subdivision across locations in the UK is associated with a particular trait or phenotype, then this association should be interpreted with care. It has recently been suggested that a positive association between the *CYP3A4-V* variant and prostate cancer in the African American population could be a confounded association attributable to population stratification [47]. Similarly, false positive results could arise from population admixture when analysing associations with the angiotensinogen (*AGT)* M235T polymorphism, given the strong variation of *AGT* M235T allele frequency across ethnic groups [48]. Ethnic heterogeneity in allele variation was also described in the *DRD4* gene in schizophrenia [49], and in the *HLA-A*, *B* and *C* alleles and haplotypes in the five major ethnic groups of the United States [50]. However, these examples are extreme and involve the admixtures of completely different ethnic groups, which should be obvious to the epidemiologist. A more subtle example is the almost two-fold greater prevalence of *APOE* ε4 allele, in north (~20%) than south (~10%) Europe [51]. This allele predisposes to both hypercholesterolaemia and late onset Alzheimer's disease. In Finland, blonde hair is prevalent, in the Mediterranean, black hair is prevalent. The mixing of Finns and Italians could thus generate a confounded study concerning hair colour and *APOE* ε4 effects. Again, personal ancestral history should alert the epidemiologist. Such situations may be much more prevalent in the USA and other recently admixed regions, than in provincial regions of Europe, where the population history is known and has been more stable over longer periods.

Low frequency mutations show much more regional variation. The same will be true for rare haplotypes. For example, regional over-representations of specific mutations have been described for the low-density lipoprotein receptor *(LDLR)* in the UK and for the hereditary hemochromatosis gene *(HFE)* in Portugal. The E80K mutation in *LDLR* was found to occur in 14.2% of probands studied in Manchester, the prevalence of this mutation ranging from 1 to 2% in the UK [52, 53]. Likewise, *LDLR* mutation R329X has been shown to occur in 11.5% probands in the south of England [54], though the prevalence elsewhere is low. The allele frequency of the C282Y mutation in the *HFE* was found to be higher in the North of Portugal (0.058) than in the South (0.009). This difference is in accordance with the observation of higher frequencies of this mutation in Northern European countries (an average of 10%) [55, 56], with a lower frequency observed in Southern European countries (1%) [57]. It has been suggested that the Nordic/Suevian occupation and settlement that occurred only in the North of Portugal, or a later Viking occupation, could be the origin of this difference [58, and references therein]. Under neutral selection, allelic frequency is generally proportional to age of allele it takes a long time to drift to high frequency. Frequency of rare alleles, by contrast, is less stable and this will vary more between locations. Familial hypercholesterolaemia (historically probably subject to little selection) has an incidence of 1/500 (1/1000 allele frequency). Peak prevalence of *LDLR* R329X or E80K is 0.01%, varying in the UK by 10-fold. *HFE* C282Y varies 6-fold in Portugal, peak prevalence 5.8%. The haplotypes of Y chromosome and 11 regions showing significant geographical frequency variations are also in the 5–10% frequency range. By contrast, the most prevalent haplotypes (e.g. >20%) show more stable frequencies. Geographical 'reference ranges' for haplotype, genotype or mutation frequencies will give some measure of the bounds of possible variation in frequency. Then, for example, frequency in cases outside of these bounds would strongly suggest causal association rather than artefacts of population genetic subdivision. Subdivision effects may turn out to be much more of an issue for rarer mutations and rarer haplotypes than for commoner ones.

*Subdivisions in Relation to Epidemiological Analyses*

Independently of the general validity of this hypothesis, the analysis of haplotypes in geographically distributed locations by means of chi-square contingency tables has proved to be useful for the examination of genetic subdivision in this work. We have previously found that one haplotype of *IGF2,* at a frequency of 10% in the NPHSII sample (haplotype 2122, markers in the order *IGF2* 6815, *IGF2* 1156, *IGF2 ApaI* and *IGF2* 1926), is significantly over-represented in lighter individuals, with mean body mass index (BMI) 1.19 kg/m$^2$ lower than for the remaining haplotypes (unpublished data). When comparing the distribution of this haplotype across locations with the distribution of the remaining haplotypes all together, by means of a chi-square contingency table, no significant evidence of subdivision was found ($p = 0.181$). In addition, the mean BMI of individuals with haplotype 2122 in each location (considering only locations where more than 10 individuals carrying this haplotype were found), ranged from –0.61 to –2.18 kg/m$^2$ when compared with the mean BMI of the remaining haplotypes. These results strengthen the conclusion that subdivision is not the cause of the association detected between 2122 haplotype and BMI and confirm previous evidence [25] suggesting that none of the associations detected between BMI and individual SNPs at the *IGF2* gene reflected any geographically evident subdivision. If the genotype causing this BMI effect were on another chromosome, then the allele conferring lower weight would have to be found in the individuals bearing *IGF2* 2122 haplotype, in every centre in the UK. The consistency of findings between geographically dispersed centres, known to represent admixtures or subdivisions within one ethnic group, should complement replication of observations between ethnic groups. However, for large population studies, the environmental exposures in the latter case may be very poorly matched.

The other approach we used to detect subdivision across the UK, i.e. the analysis of differences in genotype distribution by centre by analysis of a set of 55 autosomal SNPs relevant to cardiovascular traits, revealed that subdivision seem to be infrequent in NPHSII. Only one SNP (2% of total SNPs) showed significant or borderline significant difference in genotype distribution by centre after correction for multiple tests. This result suggests that haplotypes are more powerful than single markers to detect subdivision and confirms that none of the coronary heart disease (CHD) candidate SNPs, where an association with CHD risk has been reported in NPHSII (e.g. *APOE* [60], *IL-6* [60], *PPARα* [61], *APOA4* [62]) show any evidence for population substructure being a potential confounder (unpublished data). While we have not found any significant association between SNPs typed in relation to cardiovascular risk traits and Y or autosomal haplotypes, it is worth considering the Y haplotype findings in relation to *APOE* ε4 allele distribution. Viking Y haplotype and higher ε4 allele frequency of northern Europe [51] could have travelled together to the UK Danelaw region. The Finnish ε4 frequency (~20%) would be an overestimate of ε4 frequency in the invaders. The Viking Y haplogroup shows that at least a 7% addition to the male half of the Chesterfield gene pool was made by Viking influx. Since other common haplogroups indistinguishable from the indigenous population would also have arrived, the contribution may have been considerably greater. The effect on *APOE* ε4 allele frequency of 15% in the UK [63], could maximally have been to raise the 15% frequency by 2%, or 17.5% if only the male populus were replaced. The Viking Y haplotype represents 11% of all Y chromosomes in Chesterfield compared with 4% in other parts of the UK and with 38% in Norway and 16% in Friesland [38]. Assuming between 16 and 38% frequency in Denmark would imply approximately 17–44% replacement of the Chesterfield male population by Viking ingress. The local change in *APOE* ε4 frequency and consequent possible confounding was thus small, even through by the standards of complex trait genetics the ε4 allele confers substantial risk (×2) for Alzheimer's dementia and displays substantial frequency heterogeneity in Europe and the evidence of Viking influx is statistically striking. It seems impossible that any study could confoundedly 'discover' that Y haplogroup 2.47 predisposed to Alzheimer's disease, even if no segregation of Y and *APOE* markers had taken place. In fact we found no significant ε4 allele frequency difference in Chesterfield, compared with all the other centres [63].

## Conclusions

We conclude that Y chromosome haplotyping is a sensitive test of subdivision in a distributed epidemiological setting. Historically, in this particular case, the subdivision is due to Viking admixture. Indeed, given known human population history, it should be feasible to test for and control for admixture. On the other hand, examples of assortative mating within a population are few [64]. Geographical tests represent one powerful way for recognising possible subdivision effects, inviting the use of multicentre studies in genetic epidemiology. Autosomal haplotyping may also be useful in recognising subdivision and we propose that geographical mapping of autosomal haplotypes should form an adjunct to interpretation of haplotype-phenotype associations. Y haplotype association with autosomal haplotypes (or, less powerfully, single SNP genotypes) should provide a sensitive test for segregation. This will be applicable in instances in which historically known admixtures have taken place, of populations with known differential frequencies both of Y haplotypes and of autosomal haplotypes. However, the full prosecution of these strategies will require large genome-wide data sets as well as full geographical representation and knowledge of social and other substructures of the population.

## Acknowledgements

S.R. was an EC Marie Curie Research Fellow. I.N.M.D. was a Lister Institute Professor. The UK Medical Research Council and the British Heart Foundation are thanked for support.

## References

1 Zak NB, Shifman S, Shalom A, Darvasi A: Population-based gene discovery in the post-genomic era. Drug Discov Today 2001;6(21): 1111–1115.

2 Pritchard JK, Rosenberg NA: Use of unlinked genetic markers to detect population stratification in association studies. Am J Hum Genet 1999;65(1):220–228.

3 Morton NE, Collins A: Tests and estimates of allelic association in complex inheritance. Proc Natl Acad Sci USA 1998;95(19):11389–11393.

4 Risch NJ: Searching for genetic determinants in the new millennium. Nature 2000;405 (6788):847–856.

5 Cardon LR, Bell JI: Association study designs for complex diseases. Nat Rev Genet 2001; 2(2):91–99.

6 Risch N, Merikangas K: The future of genetic studies of complex human diseases. Science 1996;273(5281):1516–1517.

7 Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K: A comprehensive review of genetic association studies. Genet Med 2002; 4(2):45–61.

8 Lohmueller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN: Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. Nat Genet 2003;33(2):177–182.

9 Pritchard JK, Donnelly P: Case-control studies of association in structured or admixed populations. Theor Popul Biol 2001;60(3):227–237.

10 Wacholder S, Rothman N, Caporaso N: Population stratification in epidemiologic studies of common genetic variants and cancer: quantification of bias. J Natl Cancer Inst 2000;92(14): 1151–1158.

11 Cardon LR, Palmer LJ: Population stratification and spurious allelic association. Lancet 2003;361(9357):598–604.

12 Ardlie KG, Lunetta KL, Seielstad M: Testing for population subdivision and association in four case-control studies. Am J Hum Genet 2002;71(2):304–311.

13 Deng HW, Chen WM, Recker RR: Population admixture: detection by Hardy-Weinberg test and its quantitative effects on linkage-disequilibrium methods for localizing genes underlying complex traits. Genetics 2001;157(2):885–897.

14 Clayton D: Population association, in Balding DJ, Bishop M, Cannings C (eds): Handbook of Statistical Genetics. Chichester, John Wiley and Sons, 2001, pp 519–540.

15 Reich DE, Goldstein DB: Detecting association in a case-control study while correcting for population stratification. Genet Epidemiol 2001;20(1):4–16.

16 Thomas DC, Witte JS: Point: population stratification: a problem for case-control studies of candidate-gene associations? Cancer Epidemiol Biomarkers Prev 2002;11(6):505–512.

17 Devlin B, Roeder K: Genomic control for association studies. Biometrics 1999;55(4):997–1004.

18 Shen P, Wang F, Underhill PA, Franco C, Yang WH, Roxas A, Sung R, Lin AA, Hyman RW, Vollrath D, Davis RW, Cavalli-Sforza LL, Oefner PJ: Population genetic implications from sequence variation in four Y chromosome genes. Proc Natl Acad Sci USA 2000;97(13): 7354–7359.

19 Miller GJ, Bauer KA, Barzegar S, Cooper JA, Rosenberg RD: Increased activation of the haemostatic system in men at high risk of fatal coronary heart disease. Thromb Haemost 1996; 75(5):767–771.

20 Newton CR, Graham A, Heptinstall LE, Powell SJ, Summers C, Kalsheker N, Smith JC, Markham AF: Analysis of any point mutation in DNA. The amplification refractory mutation system (ARMS). Nucleic Acids Res 1989; 17(7):2503–2516.

21 Ye S, Dhillon S, Ke X, Collins AR, Day IN: An efficient procedure for genotyping single nucleotide polymorphisms. Nucleic Acids Res 2001;29(17):E88.

22 Day IN, Humphries SE: Electrophoresis for genotyping: microtiter array diagonal gel electrophoresis on horizontal polyacrylamide gels, hydrolink, or agarose. Anal Biochem 1994; 222(2):389–395.

23 O'Dell SD, Gaunt TR, Day IN: SNP genotyping by combination of 192-well MADGE, ARMS and computerized gel image analysis. Biotechniques 2000;29(3):500–506.

24 Chen XH, O'Dell SD, Day IN: Microplate array diagonal gel electrophoresis for cohort studies of microsatellite loci. Biotechniques 2002;32(5):1080–1082, 1084, 1086.

25 Gaunt TR, Cooper JA, Miller GJ, Day IN, O'Dell SD: Positive associations between single nucleotide polymorphisms in the IGF2 gene region and body mass index in adult males. Hum Mol Genet 2001;10(14):1491–1501.

26 Talmud PJ, Hawe E, Martin S, Olivier M, Miller GJ, Rubin EM, Pennacchio LA, Humphries SE: Relative contribution of variation within the APOC3/A4/A5 gene cluster in determining plasma triglycerides. Hum Mol Genet 2002; 11(24):3039–3046.

27 Qin ZS, Niu T, Liu JS: Partition-ligation-expectation-maximization algorithm for haplotype inference with single-nucleotide polymorphisms. Am J Hum Genet 2002;71(5):1242–1247.

28 Stephens M, Donnelly P: A comparison of bayesian methods for haplotype reconstruction from population genotype data. Am J Hum Genet 2003;73(6):1162–1169.

29 Gilks WR, Richardson S, Spiegelhalter DJ: Markov chain Monte Carlo in practice. London, Chapman & Hall, 1996.

30 Niu T, Qin ZS, Xu X, Liu JS: Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. Am J Hum Genet 2002;70(1):157–169.

31 Lin S, Cutler DJ, Zwick ME, Chakravarti A: Haplotype inference in random population samples. Am J Hum Genet 2002;71(5):1129–1137.

32 Slatkin M: A measure of population subdivision based on microsatellite allele frequencies. Genetics 1995;139(1):457–462.

33 Schneider S, Roessli D, Excoffier L: Arlequin, version 2000. Genetics and Biometry Laboratory, University of Geneva, Switzerland, 2000.

34 Ott J: CONTING ver 2.61. Utility programs for analysis of genetic linkage. 1988.

35 Holm S: A simple sequentially rejective multiple test procedure. Scand J Stat 1979;(6)65–70.

36 Zapata C, Alvarez G: On the detection of nonrandom associations between DNA polymorphisms in natural populations of Drosophila. Mol Biol Evol 1993;10:823–841.

37 Y Chromosome Consortium: A nomenclature system for the tree of human Y-chromosomal binary haplogroups. Genome Res 2002;12(2): 339–348.

38 Wilson JF, Weiss DA, Richards M, Thomas MG, Bradman N, Goldstein DB: Genetic evidence for different male and female roles during cultural transitions in the British Isles. Proc Natl Acad Sci USA 2001;98(9):5078–5083.

39 Hurles ME, Irven C, Nicholson J, Taylor PG, Santos FR, Loughlin J, Jobling MA, Sykes BC: European Y-chromosomal lineages in Polynesians: a contrast to the population structure revealed by mtDNA. Am J Hum Genet 1998; 63(6):1793–1806.

40 Hill EW, Jobling MA, Bradley DG: Y-chromosome variation and Irish origins. Nature 2000; 404(6776):351–352.

41 Richards JD: Viking age England. Stroud, Tempus, 2000.

42 Helgason A, Hickey E, Goodacre S, Bosnes V, Stefansson K, Ward R, Sykes B: mtDna and the islands of the North Atlantic: estimating the proportions of Norse and Gaelic ancestry. Am J Hum Genet 2001;68(3):723–737.

43 Weale ME, Weiss DA, Jager RF, Bradman N, Thomas MG: Y chromosome evidence for Anglo-Saxon mass migration. Mol Biol Evol 2002; 19(7):1008–1021.

44 Zhang S, Pakstis AJ, Kidd KK, Zhao H: Comparisons of two methods for haplotype reconstruction and haplotype frequency estimation from population data. Am J Hum Genet 2001; 69(4):906–914.

45 Stephens M, Smith NJ, Donnelly P: A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 2001; 68(4):978–989.

46 Pankow JS, Province MA, Hunt SC, Arnett DK: Regarding 'Testing for population subdivision and association in four case-control studies'. Am J Hum Genet 2002;71(6):1478–1480.

47 Kittles RA, Chen W, Panguluri RK, Ahaghotu C, Jackson A, Adebamowo CA, Griffin R, Williams T, Ukoli F, Adams-Campbell L, Kwagyan J, Isaacs W, Freeman V, Dunston GM: CYP3A4-V and prostate cancer in African Americans: causal or confounding association because of population stratification? Hum Genet 2002;110(6):553–560.

48 Jeunemaitre X, Gimenez-Roqueplo A, Célérier J, Soubrier F, Corvol P: Angiotensinogen and hypertension; in Dominiczak AF, Connell JMC, Soubrier F (eds): Molecular Genetics of Hypertension. Oxford, BIOS Scientific Publishers Ltd, 1999, pp 201–230.

49 Lung FW, Tzeng DS, Shu BC: Ethnic heterogeneity in allele variation in the DRD4 gene in schizophrenia. Schizophr Res 2002;57(2–3): 239–245.

50 Cao K, Hollenbach J, Shi X, Shi W, Chopek M, Fernandez-Vina MA: Analysis of the frequencies of HLA-A, B, and C alleles and haplotypes in the five major ethnic groups of the United States reveals high levels of diversity in these loci and contrasting distribution patterns in these populations. Hum Immunol 2001;62(9): 1009–1030.

51 Haddy N, Bacquer DD, Chemaly MM, Maurice M, Ehnholm C, Evans A, Sans S, Martins MC, Backer GD, Siest G, Visvikis S: The importance of plasma apolipoprotein E concentration in addition to its common polymorphism on inter-individual variation in lipid levels: results from Apo Europe. Eur J Hum Genet 2002;10(12):841–850.

52 Webb JC, Sun XM, Patel DD, McCarthy SN, Knight BL, Soutar AK: Characterization of two new point mutations in the low density lipoprotein receptor genes of an English patient with homozygous familial hypercholesterolemia. J Lipid Res 1992;33(5):689–698.

53 Wenham PR, Haddad L, Panarelli M, Ashby JP, Day IN, Giles PD, Humphries SE, Penney MD, Rae PW, Walker SW: Simplified detection of a mutation causing familial hypercholesterolaemia throughout Britain: Evidence for an origin in a common distant ancestor. Ann Clin Biochem 1998;35(Pt 2)226–235.

54 Day IN, Haddad L, O'Dell SD, Day LB, Whittall RA, Humphries SE: Identification of a common low density lipoprotein receptor mutation (R329X) in the south of England: complete linkage disequilibrium with an allele of microsatellite D19S394. J Med Genet 1997; 34(2):111–116.

55 Ryan E, O'Keane C, Crowe J: Hemochromatosis in Ireland and HFE. Blood Cells Mol Dis 1998;24(4):428–432.

56 Murphy S, Curran MD, McDougall N, Callender ME, O'Brien CJ, Middleton D: High incidence of the Cys 282 Tyr mutation in the HFE gene in the Irish population – implications for haemochromatosis. Tissue Antigens 1998;52 (5):484–488.

57 Carella M, D'Ambrosio L, Totaro A, Grifa A, Valentino MA, Piperno A, Girelli D, Roetto A, Franco B, Gasparini P, Camaschella C: Mutation analysis of the HLA-H gene in Italian hemochromatosis patients. Am J Hum Genet 1997;60(4):828–832.

58 Cardoso CS, Oliveira P, Porto G, Oberkanins C, Mascarenhas M, Rodrigues P, Kury F, de Sousa M: Comparative study of the two more frequent HFE mutations (C282Y and H63D): significant different allelic frequencies between the North and South of Portugal. Eur J Hum Genet 2001;9(11):843–848.

59 Kimura M, Ohta T: The age of a neutral mutant persisting in a finite population. Genetics 1974;75:199–212.

60 Humphries SE, Luong LA, Ogg MS, Hawe E, Miller GJ: The interleukin-6 -174 G/C promoter polymorphism is associated with risk of coronary heart disease and systolic blood pressure in healthy men. Eur Heart J 2001;22(24):2243–2252.

61 Flavell DM, Jamshidi Y, Hawe E, Pineda Torra I, Taskinen MR, Frick MH, Nieminen MS, Kesaniemi YA, Pasternack A, Staels B, Miller G, Humphries SE, Talmud PJ, Syvanne M: Peroxisome proliferator-activated receptor alpha gene variants influence progression of coronary atherosclerosis and risk of coronary artery disease. Circulation 2002;105(12):1440–1445.

62 Wong WR, Hawe E, Li LK, Miller GJ, Nicaud V, Pennacchio LA, Humphries SE, Talmud PJ: Apolipoprotein AIV gene variant S347 is associated with increased risk of coronary heart disease and lower aoplipoprotein AIV plasma concentrations. Circ Res 2003;in press.

63 Humphries SE, Talmud PJ, Hawe E, Bolla M, Day IN, Miller GJ: Apolipoprotein E4 and coronary heart disease in middle-aged men who smoke: a prospective study. Lancet 2001;358 (9276):115–119.

64 Murphy M, McHugh B, Tighe O, Mayne P, O'Neill C, Naughten E, Croke DT: Genetic basis of transferase-deficient galactosaemia in Ireland and the population history of the Irish Travellers. Eur J Hum Genet 1999;7(5):549–554.