People of the British Isles: preliminary analysis of genotypes and surnames in a UK-control population

Bruce Winney, Abdelhamid Boumertit, Tammy Day, Dan Davison, Chikodi Echeta, Irina Evseeva, Katarzyna Hutnik, Stephen Leslie, Kristin Nicodemus, Ellen C Royrvik, Susan Tonks, Xiaofeng Yang, James Cheshire, Paul Longley, Pablo Mateos, Alexandra Groom, Caroline Relton, D Tim Bishop, Kathryn Black, Emma Northwood, Louise Parkinson, Timothy M Frayling, Anna Steele, Julian R Sampson, Turi King, Ron Dixon, Derek Middleton, Barbara Jennings, Rory Bowden, Peter Donnelly and Walter Bodmer

There is a great deal of interest in a fine-scale population structure in the UK, both as a signature of historical immigration events and because of the effect population structure may have on disease association studies. Although population structure appears to have a minor impact on the current generation of genome-wide association studies, it is likely to have a significant part in the next generation of studies designed to search for rare variants. A powerful way of detecting such structure is to control and document carefully the provenance of the samples involved. In this study, we describe the collection of a cohort of rural UK samples (The People of the British Isles), aimed at providing a well-characterised UK-control population that can be used as a resource by the research community, as well as providing a fine-scale genetic information on the British population.

INTRODUCTION

During the last 10 years there has been much interest in a fine-scale population structure, particularly in the UK, both as a signature of historical immigration events and because of the effect population structure may have on disease association studies, although this depends on the magnitude of the associations. Fine-scale population structure is principally the outcome of historical movements of people into Britain following the last ice age about 10,000 years ago, with the major subsequent detectable influences being due to Anglo-Saxon, Norse and Norman admixture. Although population structure appears to have a minor impact on the current generation of genome-wide association studies, it is likely to have a significant part in the next generation of studies designed to search for rare variants. It is, therefore, important that suitable control population cohorts are available for such studies. In this study we describe the collection and preliminary analysis of a set of carefully chosen samples, to represent the areas of the UK from where they have come.

A powerful way of detecting a fine-scale population structure is to control and document carefully the provenance of the samples involved. This can be carried out by, for example, ensuring that volunteers are chosen for whom all four grandparents were born in the same rural area. This approach should maximise the probability of recruiting individuals whose families have been stable inhabitants of the area for many generations, as most recent migration has been into larger towns and cities. Genotyping a collection of such samples from throughout the UK should then allow identification of high-quality ancestrally informative markers and enable a detailed analysis of population structure. These samples can then be used to assess the impact of population structure on disease and other phenotype association studies, particularly when searching for rare variants. The resulting body of data will also provide an excellent basis for...
relating population structure to the known history and archaeology of the UK population.

A further way to investigate and refine the genetic signals of population structure is to use surnames when analysing the genetic data.1,4,13 The distribution of surnames has been remarkably stable over at least the last 130 years (GB Names Profiler, gbnames.public-profiler.org14), supporting the notion that the rural British population has been quite sedentary until relatively recently. Although evidence based on studies of testimonials15 suggests that there has been a great deal of movement, this is mainly over short distances. Thus, 75% of reported residential mobility was less than 10 km, with women historically averaging greater distances than men. Classification of surnames into those that have markedly local distributions, in contrast to those with wider, more national distributions, should help to enhance the signals of population structure.

Here we describe the collection of a cohort of samples carefully chosen using the above considerations, and present a preliminary analysis of some genotype and surname data on a small pilot subset of these samples. These are part of a much larger ongoing UK-wide project (The People of the British Isles (PoBI), http://www.peopleofthebritishisles.org), funded by the Wellcome Trust, to set up a well characterised and carefully collected UK-control population as a resource that can be used by the research community. Preliminary data analysis demonstrates that population structure can be detected within the UK even with a limited number of samples and loci, and that the analysis can be enhanced by using information on surnames. Here a population refers to a County or a region of the UK.

**MATERIALS AND METHODS**

**Sample collection**

Approximately 4000 rural samples from throughout the UK have so far been collected using the criteria that all four grandparents were born in the same rural area, defined as lying within 60 km linear map distance of each other. For each sample, a self-reported questionnaire was completed. Details requested included place and year of birth of grandparents, parents and the volunteer, place of residence, gender and surname at birth. As approved by the Research Ethics Committee, samples were anonymised upon collection therefore, for research undertaken outside the core research group, surname data and full date of birth will be excluded. During the period of sample collection, consent for genotyping has broadened (see Supplementary Information). The whole project was subjected to the UK standard research ethical consent procedures (Leeds (West) REC – 05/Q1205/35).

A volume of 20 μl of blood was collected from each volunteer and peripheral blood lymphocytes (PBLs) were harvested (see Supplementary Information). A number of the stored viable PBLs were subsequently transformed with Epstein Barr virus16 by the European Collection of Cell Cultures and the Avon Longitudinal Study of Parents and Children to check viability and to replenish some depleted DNA stocks, with a success rate of 531/539 (98.3%). DNA was prepared from the 10 ml of blood residue remaining after sterile separation (see Supplementary Information).

**Samples**

Basic information on numbers, gender and the age distribution of the total sample, and separately, of the sample used for the pilot genotyping is given in Table 1. At the time of this analysis, 3865 of the samples collected have had their geocoded data.3,4,13 The distribution of surnames has been remarkably stable over at least the last 130 years (GB Names Profiler, gbnames.public-profiler.org14), supporting the notion that the rural British population has been quite sedentary until relatively recently. Although evidence based on studies of testimonials15 suggests that there has been a great deal of movement, this is mainly over short distances. Thus, 75% of reported residential mobility was less than 10 km, with women historically averaging greater distances than men. Classification of surnames into those that have markedly local distributions, in contrast to those with wider, more national distributions, should help to enhance the signals of population structure.

Here we describe the collection of a cohort of samples carefully chosen using the above considerations, and present a preliminary analysis of some genotype and surname data on a small pilot subset of these samples. These are part of a much larger ongoing UK-wide project (The People of the British Isles (PoBI), http://www.peopleofthebritishisles.org), funded by the Wellcome Trust, to set up a well characterised and carefully collected UK-control population as a resource that can be used by the research community. Preliminary data analysis demonstrates that population structure can be detected within the UK even with a limited number of samples and loci, and that the analysis can be enhanced by using information on surnames. Here a population refers to a County or a region of the UK.

**Table 1 Basic information on numbers, gender and the age distribution of the total sample and, separately, of the sample used for the pilot genotyping is given in the top part of the table**

| Gender | Overall | Proportion | Pilot | Proportion |
|--------|---------|------------|-------|------------|
| M      | 1824    | 0.472      | 506   | 0.479      |
| F      | 1982    | 0.513      | 497   | 0.470      |
| Unknown| 59      | 0.015      | 54    | 0.051      |
| Total  | 3865    |            | 1057  |            |

**Age (in 2009)**

| Age group | Overall | Proportion | Pilot | Proportion |
|-----------|---------|------------|-------|------------|
| <20       | 8       | 0.002      | 0     | 0.000      |
| 20–29     | 82      | 0.021      | 13    | 0.012      |
| 30–39     | 180     | 0.047      | 33    | 0.031      |
| 40–49     | 462     | 0.120      | 66    | 0.062      |
| 50–59     | 688     | 0.178      | 172   | 0.163      |
| 60–69     | 1161    | 0.300      | 295   | 0.279      |
| 70–79     | 915     | 0.237      | 246   | 0.233      |
| 80–89     | 291     | 0.075      | 96    | 0.091      |
| 90–99     | 21      | 0.005      | 12    | 0.011      |
| >100      | 10      | 0.003      | 2     | 0.002      |
| Unknown   | 47      | 0.012      | 122   | 0.115      |
| Total     | 3865    |            | 1057  |            |

**MD between grandparental birthplaces**

| MD (between grandparental birthplaces) | Overall | Proportion | Pilot | Proportion |
|---------------------------------------|---------|------------|-------|------------|
| Median (km)                           | 16.05   | 16.31      |       |            |
| 25% quartile (km)                     | 2.96    | 3.72       |       |            |
| 75% quartile (km)                     | 44.85   | 48.92      |       |            |
| n                                     | 3646    | 893        |       |            |
| No. missing                           | 219     | 65         |       |            |
| Orkney<sup>a</sup>                    | 0       | 99         |       |            |

The lower part of the table gives the median and 25 and 75% quartiles of the mean distance (MD) between grandparental birthplaces for volunteers who gave information for all four grandparents. 99% of the unknown age group in the pilot data are previously collected Orkney samples.<sup>19</sup> These are not included in the overall geocoded data set.

**Use of surnames to subdivide populations**

Surnames of the volunteers were routinely collected and this knowledge should allow a more detailed investigation of population structure. Individuals whose surnames are localised to an area are more likely to have ancestry from that area down the male lineage, and should be more representative of the region over a long time period. This should be backed up by the genetics. Although it is possible to determine a surname’s area of origin from contemporary data, historical data sets are advantageous because they are less affected by recent migrations. The digitisation of the 1881 Census of Great Britain (UK Data Archive, http://www.data-archive.ac.uk) provides the names and place of enumeration (Parish and Registration District) for 29 million people, with a total of 425 000 unique surnames, ~49 000 of which occur in more than 20 individual census records. These data have been geocoded to registration district (RD) level (mean population 4900) and linked to a shapefile containing the historical boundary data.<sup>17</sup>
Some surname distributions are very localised (e.g. Grahamslaw, Forster or Pedlar, Supplementary Figure 1), while other surnames are much more prevalent throughout the UK (e.g. Smith or Grey). The distribution of the frequencies of surnames in districts throughout the UK provides an approach to assessing how local a surname is. This can be carried out using the location quotient, which compares the relative frequency of a surname in a given region with the relative frequency of that surname at a more aggregate spatial level,\(^{18}\) for example a county or district versus Great Britain as a whole. It is defined as follows:

\[
\text{LQ}_i = \frac{A_{ij}}{\sum_i A_{ij}} \times \frac{B_{ij}}{\sum_j B_{ij}},
\]

where \(A_{ij}\) is the count of surname \(i\) in registration district (RD) \(j\), \(B_i\) is the count of surname \(i\) in Great Britain, \(n\) is the total number of surnames in Britain and \(LQ_i\) is the location quotient of surname \(i\) in region \(j\). LQ values greater than 1 indicate an RD with a higher concentration of the selected name that would be expected if the surname had a uniform distribution throughout the Britain.

The RDs with the three highest LQs for a given surname are taken to define the surname’s core locality. In many cases these are contiguous or at least very close to each other, and this is taken to indicate that the surname has a single core. If this is not the case, the surname may either have more than one core or a dispersed distribution.

The district with the maximum LQ (MLQ) can be used as a starting point for assigning a surname as local or non-local. In general it appears that surnames with high MLQs tend to be comparatively rare (Figure 1) and are more likely to have a local distribution (e.g. Pedlar MLQ=323). There are, however, some surnames with relatively lower MLQs that might be expected for the surname sample size (Jones, Davies, Evans, Thomas, Hughes, James and Phillips), which are established Welsh surnames. The surnames from Supplementary Figure 1 are also marked.

Assessment of allele frequency differences and calculation of \(F_{ST}\)

To conduct a meaningful analysis of population structure with the limited genotyping we have so far carried out on the pilot samples, these were pooled into groups based mainly on geographical association, but also to some extent using historical and archaeological criteria.\(^{10}\) We recognise that these distinctions are somewhat arbitrary and their effect will be investigated in more detail in the future work. Cornwall, Devon and Pembrokeshire were pooled to represent the South/West (SW) and the area that could be considered the closest surrogate to the Ancient British. Kent, Norfolk and Lincolnshire were pooled to represent the East (E) and the area most directly influenced by the Anglo-Saxon invasions. Cumbria, Yorkshire and the North East were pooled broadly to represent the North of England (N); Oxfordshire and the Forest of Dean were combined to represent the Central region of England (CN); and Orkney was kept separate from the others, largely because of the known substantial Norse Viking influence in Orkney. The aim was to achieve a grouping that, a priori and given the limitations of the sample size, would be most likely to reveal differences in a regional fine-scale population structure.

Fisher’s exact test was used to assess allele frequency differences using \(2 \times 2\) tables of allele counts to split the data in three ways (see Supplementary Information) and \(F_{ST}\) was calculated using Weir and Cockerham’s method.\(^{30}\)

RESULTS

Sampling

For the 3865 of the samples that have been geocoded the distances between birthplaces could be accurately and consistently calculated. Of these, 958 were genotyped for this study. The distribution in England and Wales of the MGP of each individual’s grandparents birthplace is shown in Figure 2. The data on distances between grandparental birthplaces, given in Table 1, show that the median of the MD between grandparental birthplaces for all the geocoded samples is 16.05 km (quartiles 2.96 and 44.85 km), while it is slightly larger for the genotyped samples (16.31 km, 3.72 and 48.92 km). The overall distribution of these distances is skewed towards the lower values (Supplementary Figure 3). The individuals who did not know where all their grandparents were born, and the 99 genotyped Orkney samples for whom this information was not available, are excluded from these calculations. Overall, 219 out of the 3865 geocoded samples were excluded from further analysis using distance information.

Using the approaches discussed in the methods section for the definition of rural versus urban, the proportion of grandparents from the 3865 geocoded samples who were born in rural areas ranges from 0.375 (assuming the stringent criterion that people born within 10 km
of small towns of 20,000 people (as of 2001), such as Penzance, or any towns larger than this, count as urban) to 0.859 (assuming the much less stringent criterion that only those born within 2 km of large cities of 300,000 or more, such as Southampton, count as urban, Supplementary Table 2). Choosing a definitive cut off population size for the distinction between rural and urban is difficult, but from Figure 3, (Supplementary Table 2) plotting the proportion of rural samples against population size for different distances, there seems to be a definite discontinuity at around population size 125,000 (eg Doncaster). Choosing this size as the threshold that distinguishes rural from urban gives estimates of the proportion of rural volunteers, for all geocoded samples, which range from 0.726 to 0.757, depending on the distance from the urban area. In the geocoded samples, there are 683 (4.5%) grandparental birthplaces that were given simply as a county and 365 (2.4%) that were unknown. The corresponding numbers for the genotyped data are 120 (3.1%) and 94 (2.5%).

Local classification by surname

Surnames of individuals in the pilot set were classified as local using a combination of five different MLQ thresholds and two different thresholds for distances between the MGP and the district with the MLQ for the individual’s surname (Table 3). The proportion of surnames classified as local ranged from 0.034 (Cumbria and Yorkshire with a threshold LQ of 300) to 0.767 (Cornwall with a threshold MLQ of 19). Cornwall and Kent/Sussex generally had, respectively, the highest and second highest proportions of local...
Davies, Evans, Thomas, Hughes, James and Phillips), surnames that are distinctive, but at a scale that is region specific. There are also some surnames that were not classified as local despite having a high MLQ. This is either because they had a multi-centre distribution or the average grandparental birthplace was further than 83 or 120 km from the district with the MLQ. The proportion excluded from the local classification for these reasons ranged from 0 (several populations for which high MLQ thresholds were used) to 0.385 (Pembrokeshire, MLQ>19, Supplementary Table 3).

Genotypes

In all, 1019 of the pilot samples were successfully genotyped and the genotype data for the loci typed are given, by region, in Table 2 (Supplementary Figure 4). Only HLA alleles with a frequency greater than 7.5% in at least one population are shown here. The full HLA allele data set is given in Supplementary Table 1. All autosomal loci were in Hardy–Weinberg equilibrium.

Evidence for population structure

Pairwise $F_{ST}$ values, calculated separately for each marker, showed no obvious consistent patterns, apart from the suggestion at three loci (HLA-B, rs7853989 and NRY) that the Orcadian samples appear to be significantly different from the rest (Supplementary Table 4). As may be expected from a marker with a lower effective population size, $F_{ST}$ values still did not reveal any consistent patterns, apart from the suggestion at three loci (MLQ 4, 120 and distance 83 km) suggested that most of the contribution was from the Eastern population (0.945 East (0.895–0.995), Table 4). When only non-local samples are used for the analysis, there was a substantial contribution from both source populations (0.630 East (95% CI 0.591–0.669), Table 4). Using a much lower stringency (MLQ>19, distance120 km), the estimates suggested that there was again a major contribution from the Eastern population (0.900, 0.829–0.971) and again, when non-local samples are used, there was a substantial contribution from both source populations (0.525, 0.482–0.568). The NRY sample sizes were too small to allow analysis of subdivided data. Using all the available male samples, the Eastern contribution to the Central population was still substantially greater than the Western contribution, although the confidence...
The two main criteria were a minimum location quotient (LQ) of the district with the highest LQ (MLQ) and maximum distance of the mean grandparental place of birth (MGP) from that district for each sample. When no distance is given, the distance constraint was not used. A number of samples were further excluded because of observed multiple peaks or broad geographic surname distributions (see Supplementary Table 3). These exclusions are incorporated into the proportions here.

Table 3 Proportion of surnames classified as local depending on different exclusion criteria

| Population  | Proportion | MLQ>19, dist<83 km | MLQ>19, dist<120 km | MLQ>19 | MLQ>45, dist<83 km | MLQ>45, dist<120 km | MLQ>45 | MLQ>120, dist<83 km | MLQ>120, dist<120 km | MLQ>120 | MLQ 120 | MLQ 200 | MLQ 300 |
|-------------|------------|---------------------|---------------------|-------|---------------------|---------------------|-------|---------------------|---------------------|---------|---------|---------|---------|
| Cornwall    | 0.550      | 0.583               | 0.767               | 0.467  | 0.483               | 0.533               | 0.417  | 0.433               | 0.467               | 0.267   | 0.217   |
| Cumbria     | 0.345      | 0.397               | 0.552               | 0.293  | 0.293               | 0.328               | 0.190  | 0.190               | 0.190               | 0.086   | 0.034   |
| Devon       | 0.316      | 0.354               | 0.684               | 0.316  | 0.342               | 0.456               | 0.253  | 0.266               | 0.316               | 0.152   | 0.076   |
| Forest of Dean | 0.164  | 0.299               | 0.478               | 0.149  | 0.209               | 0.239               | 0.090  | 0.134               | 0.149               | 0.119   | 0.045   |
| Kent/Sussex | 0.469      | 0.469               | 0.653               | 0.429  | 0.429               | 0.490               | 0.388  | 0.367               | 0.408               | 0.204   | 0.122   |
| Lincolnshire | 0.367      | 0.433               | 0.667               | 0.367  | 0.400               | 0.567               | 0.267  | 0.267               | 0.333               | 0.133   | 0.067   |
| North East  | 0.324      | 0.382               | 0.588               | 0.309  | 0.338               | 0.485               | 0.096  | 0.103               | 0.154               | 0.088   | 0.044   |
| Norfolk     | 0.430      | 0.440               | 0.700               | 0.400  | 0.410               | 0.520               | 0.230  | 0.240               | 0.270               | 0.150   | 0.120   |
| Pembroke/shire | 0.436    | 0.487               | 0.590               | 0.231  | 0.256               | 0.359               | 0.103  | 0.103               | 0.128               | 0.051   | 0.051   |
| Oxfordshire | 0.278      | 0.316               | 0.582               | 0.241  | 0.266               | 0.380               | 0.190  | 0.203               | 0.266               | 0.165   | 0.101   |
| Yorkshire   | 0.372      | 0.414               | 0.621               | 0.248  | 0.269               | 0.379               | 0.090  | 0.103               | 0.138               | 0.083   | 0.034   |
| All populations | 0.363     | 0.411               | 0.625               | 0.309  | 0.333               | 0.431               | 0.186  | 0.200               | 0.236               | 0.131   | 0.077   |

The majority of the samples collected did fit the criteria required. Analysis of the first 3865 samples that have been geocoded indicates that 75% have an MD between grandparental birthplaces of 37.3 km (Table 1), and about 70% of grandparental birthplaces could be classed as rural, although this does depend on the criteria used. These figures emphasise the quality of the samples collected, which gives the potential for a finer-scale analysis of the UK population that can be carried out using other available control sample collections.

Preliminary genotyping of 1057 samples, using nine loci, demonstrates the value of these samples for investigating a fine-scale population structure within the UK. The use of traditional methods such as pairwise estimation of FST, PCA and STRUCTURE (PCA and STRUCTURE were both applied but showed no patterns) failed to
detect any structure in this pilot project, probably because the sample sizes and numbers of loci used are too small to detect such differences. Instead, we have used an admixture analysis, based on historical priors, to investigate whether a fine-scale structure in the UK could be detected in these samples and to see if partitioning the samples by surnames, an important asset of our PoBI cohort, enhances the power to detect structure. Simple point admixture estimates, based on linear combinations of contributions from ancestral populations, did reveal the expected population structure. This was more finely dissected using the surname data to further stratify the samples by local and non-local surnames. In particular, for both the high and low stringencies, there is a significant difference between admixture estimates for the local versus non-local surnames in both the CN (Central) and Orkney populations when the Eastern and Western populations are used as parental populations (Table 4).

Table 4 Maximum likelihood admixture estimates for the most stringent and the least stringent criteria used to define local and non-local surnames

| 'Admixed' population | Parental populations | Local (L) or non-local (N) | Proportion East | −95% CI | +95% CI |
|----------------------|----------------------|---------------------------|----------------|---------|---------|
| CN > 120, distance < 83 km² | West vs East | L | 0.945 | 0.895 | 0.995 |
| CN > 120, distance < 83 km² | West vs East | N | 0.630 | 0.591 | 0.669 |
| OR > 19, distance < 120 km² | West vs East | L | 0.550 | 0.488 | 0.614 |
| OR > 19, distance < 120 km² | West vs East | N | 0.695 | 0.630 | 0.760 |

The contributions of the putative ancestral populations (East, West and Norse) to the putative admixed population (Central (CN) or Orkney (OR)) were estimated for either the local surnames (L) alone or only the non-local (N) surnames. For the Orkney analysis, all Orcadian samples were compared with either local or non-local stratified PoBI samples.

ACKNOWLEDGEMENTS
The authors declare no conflict of interest.

CONFLICT OF INTEREST
The authors declare no conflict of interest.

ACKNOWLEDGEMENTS
This project was funded by the Wellcome Trust, and we would like to acknowledge Lon Cardon and David Bicknell for their help in getting the grant application together. Sorrel May and Nick Godwin helped to kick start the large-scale recruitment of volunteers, while Chris Abell, Tanefa Apekey, Carolina Bonilla, Julie Burton, John Burns, Joy Hadfield, Christopher Hand, Amanda Howe, Pat Jonas, Ciaran Kilkelly, Donald Lehman, Julie Lewis, Louise Lynagh, Shan Owens, Jan Tawn, Malcolm Taylor and Stan Urbaniaik were heavily involved in recruiting. This project was supported by the Peninsula NIHR Clinical Research Facility (University of Exeter) and the Wellcome Trust Clinical Research Facility (Manchester). Finally, we would like to thank all the volunteers themselves and the many friends and colleagues who helped to recruit them.

1. 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: 5078-5083.
2. Capelli C, Redhead N, Abernethy J et al. A Y chromosome census of the British Isles. Curr Biol 2003; 13: 979-984.
3. King TE, Ballereau SJ, Schurter KE, Jobling MA: Genetic signatures of coancestry within surnames. Curr Biol 2006; 16: 384-388.
4. Bowden GR, Balaresque P, King TE et al. Excavating past population structures by surname-based sampling: the genetic legacy of the Vikings in northwest England. Mol Biol Evol 2008; 25: 301-309.
