

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

Bruce Winney<sup>1</sup>, Abdelhamid Boumertit<sup>1</sup>, Tammy Day<sup>1</sup>, Dan Davison<sup>2</sup>, Chikodi Echeta<sup>1</sup>, Irina Evseeva<sup>3</sup>, Katarzyna Hutnik<sup>1</sup>, Stephen Leslie<sup>1</sup>, Kristin Nicodemus<sup>4</sup>, Ellen C. Royrvik<sup>1</sup>, Susan Tonks<sup>5</sup>, Xiaofeng Yang<sup>1</sup>, James Cheshire<sup>6</sup>, Paul Longley<sup>6</sup>, Pablo Mateos<sup>6</sup>, Alexandra Groom<sup>7</sup>, Caroline Relton<sup>7</sup>, D. Tim Bishop<sup>8</sup>, Kathryn Black<sup>8</sup>, Emma Northwood<sup>8</sup>, Louise Parkinson<sup>8</sup>, Timothy M. Frayling<sup>9</sup>, Anna Steele<sup>10</sup>, Julian R. Sampson<sup>11</sup>, Turi King<sup>12</sup>, Ron Dixon<sup>13</sup>, Derek Middleton<sup>14</sup>, Barbara Jennings<sup>15</sup>, Rory Bowden<sup>16</sup>, Peter Donnelly<sup>2</sup>, Walter Bodmer<sup>1</sup>

1. Department of Oncology, University of Oxford, ORCRB, Old Road Campus, Headington, Oxford, OX3 7DQ, UK.
2. The Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford, OX3 7BN, UK.
3. Anthony Nolan, 2-3 Heathgate Place, 75-87 Agincourt Road, London, NW3 2NU, UK.
4. MRC Functional Genomics Unit, Department of Physiology, Anatomy and Genetics, Le Gros Clark Building, South Parks Road, Oxford, OX1 3QX, UK.
5. Clinical Trials and Research Governance, Manor House, John Radcliffe Hospital, Headington, Oxford, OX3 9DZ, UK.
6. Department of Geography, University College London, Gower Street, London WC1E 6BT, UK.
7. Institute of Human Genetics, Newcastle University, International Centre for Life, Central Parkway, Newcastle upon Tyne, NE1 3BZ, UK.
8. Section of Epidemiology and Biostatistics, Leeds Institute of Molecular Medicine, Wellcome Trust Brenner Building, St James's University Hospital, Leeds, LS9 7TF, UK.
9. Peninsula College of Medicine & Dentistry, St. Lukes Campus, Magdalen Road, University of Exeter, EX1 2LU, UK.
10. Peninsula NIHR Clinical Research Facility, Peninsula Medical School (University of Exeter), Barrack Road, Exeter, EX2 5DW, UK.
11. Department of Medical Genetics, Haematology & Pathology, Cardiff University School of Medicine, Institute of Medical Genetics Building, Heath Park, Cardiff,

- CF14 4XN, UK.
12. Department of Genetics, University of Leicester, Leicester, LE1 7RH, UK.
  13. School of Natural & Applied Sciences, Faculty of Health & Life Sciences, University of Lincoln, Brayford Pool, Lincoln, LN6 7TS, UK.
  14. Transplant Immunology, Royal Liverpool University Hospital, Liverpool, L7 8XP, UK.
  15. School of Medicine, Health Policy and Practice, University of East Anglia, Norwich, NR4 7TJ, UK.
  16. Department of Statistics, University of Oxford, South Parks Road, Oxford, OX1 3TG, UK.

**Corresponding author:** Bruce Winney, Department of Oncology, University of Oxford, ORCRB, Old Road Campus, Headington, Oxford, OX3 7DQ, UK.

Tel: 01865 617007

Fax: 01865 617100

E-mail: [bruce.winney@clinpharm.ox.ac.uk](mailto:bruce.winney@clinpharm.ox.ac.uk)

**Running Title:** A control population for the UK

## **ABSTRACT**

There is a great deal of interest in 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 play 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. Here 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 fine scale genetic information on the British population.

So far, some 4,000 samples have been collected, the majority of which fit the criteria of coming from a rural area and having all four grandparents from approximately the same area. Analysis of the first 3,865 samples that have been geocoded indicates that 75% have a mean distance between grandparental places of birth of 37.3km, and that about 70% of grandparental places of birth can be classed as rural. Preliminary genotyping of 1,057 samples demonstrates the value of these samples for investigating fine scale population structure within the UK, and shows how this can be enhanced by the use of surnames.

**Keywords:** People of the British Isles, population structure, control population, admixture, genetics, surnames

## **INTRODUCTION**

During the last ten years there has been much interest in fine scale population structure, particularly in the UK, both as a signature of historical immigration events<sup>1-6</sup> and because of the effect population structure may have on disease association studies<sup>7, 8</sup>, though this depends on the magnitude of the associations<sup>9</sup>. 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 likely to be from Anglo-Saxon, Norse and Norman admixture<sup>10</sup>. Although population structure appears to have a minor impact on the current generation of genome-wide association studies<sup>9</sup>, it is likely to play a significant part in the next generation of studies designed to search for rare variants<sup>11, 12</sup>. It is, therefore, important that suitable control population cohorts are available for such studies. In this paper we describe the collection and preliminary analysis of a set of carefully chosen samples, to represent the areas of the UK from which they come.

A powerful way of detecting fine scale population structure is to control and document carefully the provenance of the samples involved. This can be done 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, since 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 utilise surnames when analysing the genetic data<sup>3, 4, 13</sup>. The distribution of surnames has been remarkably stable over at least the last 130 years (GB Names Profiler, [gbnames.publicprofiler.org](http://gbnames.publicprofiler.org)<sup>14</sup>), supporting the notion that the rural British population has been quite sedentary until relatively recently. Though evidence based on studies of testimonials<sup>15</sup> 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 10km, 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), [www.peopleofthebritishisles.org](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 region of the UK.

## 1. MATERIALS AND METHODS

### *Sample Collection*

Approximately 4,000 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 60km 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.

1a. As approved by the Research Ethics Committee, samples were anonymised upon collection so, 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 subject to UK standard research ethical consent procedures (Leeds (West) REC – 05/Q1205/35).

Twenty millilitres of blood were 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 Virus<sup>16</sup> 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 (98.5%). DNA was prepared from the 10ml 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 3,865 of the samples collected have had their birthplaces geocoded by assigning longitude and latitude coordinates. From these coordinates, the mean distance (MD) between the known grandparental birthplaces of each volunteer who gave details of all four grandparents was calculated (see Supplementary Information).

The geocoded place names make it possible to estimate, for any given set of volunteers, what proportion of their grandparents were born in a rural or in an urban area. For this analysis, the extent of UK urban areas was derived from a map layer provided by ESRI ([www.esri.com](http://www.esri.com)). For each sample, the mean geographical position (MGP) of the grandparental birthplaces was mapped using the ArcGIS 9.3 package ([www.esri.com](http://www.esri.com)). To determine if a MGP was rural, the distance to the fringe of the nearest urban area was calculated based on the straight line to the closest point in the fringe. MGPs were then assigned as rural if they were greater than a defined distance away from the edge of that urban area of a given population size, based on the 2001 census. A range of values of the distances and sizes of urban populations for this definition of rural was investigated.

#### *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 datasets are advantageous because they are less affected by recent migrations. The digitisation of the 1881 Census of Great Britain (UK Data Archive, [www.data-archive.ac.uk](http://www.data-archive.ac.uk)) provides an invaluable resource for the definition of area of origin. Although it is not the earliest available Census, it remains the one that has been digitally encoded (by the Church of Jesus Christ of the Latter-day Saints) to the highest quality. It provides the names and place of enumeration (Parish and Registration District) for twenty nine million people, with a total of 425,000 unique surnames, approximately 49,000 of which occur in more than 20 individual Census records. These data have been geocoded to

Registration District (RD) level (mean population 4,900) and linked to a shapefile containing the historical boundary data<sup>17</sup>.

Some surname distributions are very localised (eg. Grahamslaw, Forster or Pedlar, Supplementary Figure 1), while other surnames are much more prevalent throughout the UK (eg. Smith or Gray). 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 done 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<sup>18</sup>, for example a county or district versus Great Britain as a whole. It is defined as follows:

$$LQ_{ij} = \frac{A_{ij} / \sum_{i=1}^n A_{ij}}{B_i / \sum_{i=1}^n B_i}$$

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  $LQ_{ij}$  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 than would be expected if the surname had a uniform distribution throughout 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 (eg. Pedlar MLQ=323). There are, however, some surnames with relatively lower MLQs that are



relatively common but still have, in essence, local distributions (eg. Forster MLQ=45). In order to investigate the effects of utilising surname localisation on the ability to detect genetic population structure, a range of MLQs was at first used as a cut off to define local versus non-local surnames. These were 19, 45 and 120, respectively the lower quartile, the median and the upper quartile of the distribution of the highest MLQs for each surname. The definitions of local and non-local were then refined according to whether there were two or more non-adjacent RDs with similarly high LQs in which case the surname was reclassified as non-local (eg. Wyer, MLQ=297). A further refinement was based on whether the MGP of the birthplaces of the four grandparents of a given individual was less than either 83km (the median of the distribution of the MDs) from the district with the MLQ for the given individual's surname, or less than 120km from the district (twice the maximum distance between birthplaces of the grandparents of a given individual aimed at when collecting samples). Only if both the MLQ and distance from the MLQ criteria were satisfied was the surname classified as local (Supplementary Figure 2).

### *Genotyping*

1,057 of the samples were used in an initial pilot genotyping project, which included cell lines from 99 Orcadian samples previously collected by the laboratory<sup>19</sup>. The samples were genotyped with a number of markers that were chosen because they have been used to differentiate populations by many different studies over the years. Specifically they were: *HLA*<sup>20-22</sup> (typed at a low-medium resolution, Table 2, Supplementary Table 1), MC1R (R151C (rs1805007) and R160W (rs1805008), the minor alleles of which are associated with red hair<sup>23</sup>)<sup>24, 25</sup>, *ABO*<sup>26, 27</sup> (rs7853989, the SNP that differentiates alleles A and B) and the Y-chromosome (*NR1*)<sup>1, 2, 28, 29</sup>. The six most common *NR1* haplogroups<sup>2</sup> were typed (Table 2) as defined by specific SNPs (R1a1 (rs3908), F(xI/J2/R1) (rs2032652), E (rs9306841), I (rs2032597), J2 (rs2032604), R1(xR1a1) (rs2032624)).

### *Assessment of allele frequency differences and calculation of $F_{ST}$*

In order 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<sup>10</sup>. 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 regional fine scale population structure.

Fisher's exact test was used to assess allele frequency differences using 2x2 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<sup>30</sup>.

#### *Admixture*

In order to investigate further signals of fine scale population structure within the UK, point estimates of admixture were calculated using a maximum likelihood approach<sup>31</sup> (see Supplementary Information). Autosomal admixture was estimated using the six most common *HLA-A*, *-B*, *-DRB1* haplotypes, together with only those *HLA* alleles not represented on any of those six haplotypes, and the *MC1R* and *ABO* SNPs.

## **RESULTS**

### *Sampling*

For the 3,865 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 16.05km (quartiles 2.96km and 44.85km), while it is slightly larger for the genotyped samples (16.31km, (3.72km and 48.92km)). 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 3,865 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 3,865 geocoded samples who were born in rural areas ranges from 0.375 (assuming the stringent criterion that people born within 10km 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 2km 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, that 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 MGPs 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 surnames, and Norfolk and Lincolnshire generally have the next highest proportions of local surnames. 845 of the geocoded samples, 824 of which had been successfully genotyped were used for the local classification of surnames.

Figure 1 shows, for each surname, a plot of the MLQ against the surname population size as given in the 1881 UK census. There are a few obvious outliers from the general distribution, which indicates that there are a few surnames with higher MLQs than would be expected from their abundance, with MLQs ranging from 23 to 42. These surnames are almost exclusively established Welsh surnames (Jones, 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 83km or 120km 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*

1,019 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 *NR1*) 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 calculated using the *NR1* data were greater than those for the autosomal markers.

The aim of dividing the samples into those with local as opposed to non-local surnames, was to see whether this would accentuate regional divergence and so reveal a greater extent of population substructure. The procedures described in the methods section for distinguishing between local and non-local surnames enable a hierarchical classification of the samples based on a combination of MLQ values and distance constraints. This ranges, as described above, from no constraint (no splitting between local and non-local) to the maximum locality constraint of an  $MLQ > 120$  and distance  $< 83\text{km}$ , with lower LQ cut offs and the lesser distance cut offs lying somewhere between these two extremes. Pairwise  $F_{ST}$  values calculated from different degrees of locally defined surname samples still did not reveal any consistent patterns (Supplementary Table 5).

Given that the  $F_{ST}$  analysis was clearly not powerful enough to detect population structure in our pilot sample, we decided to see whether an analysis of population admixture might be more revealing. For this we first assumed that the central population was a simple mixture between two source populations, namely the South West, a surrogate for the Ancient British, and the Eastern, a surrogate for the Anglo-Saxons. Using only local

samples of each of the population groups to estimate the admixture, by the maximum likelihood procedure, the autosomal data with the most stringent thresholds (MLQ>120, distance<83km) 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, distance<120km), 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 intervals were very large (0.620, 0.000-1.000). At face value these data suggest first of all that there is measurable population sub-structure, in contrast to the  $F_{ST}$  calculations. Secondly they suggest a very substantial contribution to the central population from the East, putatively the Anglo-Saxons. Intriguingly, the difference between the autosomal and *NRY* analysis suggests that the male Eastern contribution may be less than the female. However, the *NRY* CIs are large.

The Orcadian population is thought to be a mixture of Norse Vikings and, mostly, the Ancient British<sup>1, 28</sup>. Because our Norse population surrogate was based on limited published Norwegian data, we used only a subset of the autosomal data (*HLA-A*, *-B*, *-C*, *-DQB1*, *MC1R*, rs7853989) for the admixture analysis. The source populations were the South Western set, as before as a proxy for Ancient British ancestry, and published Norwegian (or Swedish) data as a proxy for Norse Viking ancestry. The estimate of Norse ancestry was 0.375 (0.331-0.419) for the local surnames, rising just slightly to 0.405

(0.357-0.453) when non-local surnames were used at the highest stringency. These estimates were 0.315 (0.266-0.364) and 0.420 (0.375-0.465) at a lower stringency. The *NRV* estimate of Norse ancestry was 1.000 (0.139-1.000), again with a very wide CI.

We repeated the analysis on the Orcadian samples using the Eastern set instead of Norway. This comparison showed a lower admixture from the East for the local than the non-local samples, especially using the less stringent criteria. This may well be because the non-local samples are ‘contaminated’ with some Viking admixture, though possibly mainly from the Danish Vikings, who must have been very closely related to the Anglo-Saxons since they came from essentially the same geographical area. Using the most stringent criteria for local, the estimates of admixture from West versus East and Norse versus West match remarkably well, suggesting in both cases a nearly 50% contribution from Ancient British to Orcadian ancestry, with a likely higher Norse contribution from males than females. There can be no doubt that the admixture analysis is much more sensitive for the detection of population structure in these rather closely related populations, and that the use of local surnames, does affect the analysis and helps to create a finer population subdivision.

## **DISCUSSION**

The PoBI samples represent a very carefully recruited set of rural volunteers with the intention that they can be used as a standard UK Control population. The main advantage of the samples is that the provenance of the four grandparents is known, reaching further into the past than by simply using the volunteer’s place of birth. This greatly improves the chance that the volunteers are locally representative samples and avoids recent admixture events as far as possible.

The most challenging aspect of this project has been to collect samples from volunteers who fit the stringent selection criteria. A number of methods were used to recruit the volunteers through a collaboration with 10 groups spread throughout the country and it took a full five years to collect the current 4,000 PoBI samples. This is largely due to the

fact that, from our experience, a small proportion of people (probably less than 5% of the population in general) fit the criteria. Indeed, the age range of the samples, with the majority being over 60, suggests that there is likely to have been more movement in recent years, and so in the future, fewer people will fit these criteria. It should, however, be borne in mind that this bias in the ages will also, to some extent, be dictated by availability of volunteers to attend events because of restrictions caused by work and some self-selection of volunteers with an interest in family history. Our volunteers are older than the population average (the average age for starting genealogical research is likely to be 40 ([psych.fullerton.edu/genealogy/#elderly](http://psych.fullerton.edu/genealogy/#elderly))), but the older age distribution has the advantage of giving a greater time depth to the set of samples. Numbers of individuals who fit the criteria will continue to decline as a result of an increasingly mobile global community and so now may well be the last opportunity to collect such samples with relative ease.

The majority of the samples collected did fit the criteria required. Analysis of the first 3,865 samples that have been geocoded indicates that 75% have an MD between grandparental birthplaces of 37.3km (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 than can be done using other available control sample collections.

Preliminary genotyping of 1,057 samples, using nine loci, demonstrates the value of these samples for investigating fine scale population structure within the UK. The use of traditional methods such as pairwise estimation of  $F_{ST}$ , 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 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).

The project has now collected about 4,000 samples that are available for further analysis. Just under three thousand of the samples have recently been genotyped as replication controls for WTCCC2 on both the Illumina 1.2M and Affymetrix v6.0 whole-genome SNP platforms and these data will facilitate a more detailed investigation of UK population structure. These genotype data should be further enhanced by using surnames to improve the sample localisation, together with a careful geocoding analysis using the detailed knowledge of the grandparental birthplaces. One hundred of the samples, split between regions likely to be most representative of the Ancient Britons and the subsequent Anglo-Saxon and Norse Viking incursions (Cornwall, Kent, Orkney, West Scotland), are being sequenced as part of the 1,000 Genomes Project<sup>12</sup>. These will also be available to the research community.

2.

We believe our method of selecting volunteers is a powerful way to collect a set of samples that can be used for high quality analysis of fine scale population structure in the UK. Subsequent localisation using surnames can sharpen the results of the structure analysis. Even with limited data, an appropriate admixture analysis can give a much more refined result than use of  $F_{ST}$ , or PCA and similar structure detection analyses.

## REFERENCES

- 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 U S A* 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, Schurer KE, Jobling MA: Genetic signatures of coancestry within surnames. *Curr Biol* 2006; **16**: 384-388.
- 4 Bowden GR, Balaesque 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.
- 5 Novembre J, Johnson T, Bryc K et al.: Genes mirror geography within Europe. *Nature*. 2008; **456**: 98-101.
- 6 O'Dushlaine CT, Morris D, Moskvina V et al.: Population structure and genome-wide patterns of variation in Ireland and Britain. *Eur J Hum Genet* 2010; **18**: 1248–1254.
- 7 Cardon LR, Bell JI: Association study designs for complex diseases. *Nat Rev Genet* 2001; **2**: 91-99.
- 8 Marchini J, Cardon LR, Phillips MS, Donnelly P: The effects of human population structure on large genetic association studies. *Nat Genet* 2004; **36**: 512-517.
- 9 Wellcome Trust Case Control Consortium: Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 2007; **447**: 661-678.
- 10 Miles D: The Tribes of Britain: who are we and where do we come from? London, Weidenfeld and Nicolson, 2005.
- 11 Bodmer W, Bonilla C: Common and rare variants in multifactorial susceptibility to common diseases. *Nat Genet* 2008; **40**: 695-701.

- 12 The 1000 Genomes Project Consortium: A map of human genome variation from population-scale sequencing. *Nature* 2010; **467**: 1062-1073.
- 13 King TE, Jobling MA: Founders, drift and infidelity: the relationship between Y chromosome diversity and patrilineal surnames. *Mol Biol Evol* 2009; **26**: 1093-1102.
- 14 Cheshire J, Mateos P, Longley PA: Creating a Regional Geography of Britain through the Spatial Analysis of Surnames, *Geoforum* in press.
- 15 Whyte ID: Migration in early-modern Scotland and England; in Pooley CG, Whyte ID (eds): *Migrants Emigrants and immigrants*. Routledge, 1991.
- 16 Sly WS, Gurbax SS, Kennett R, Bodmer WF, Bodmer JG: Permanent lymphoid lines from genetically marked lymphocytes: success with lymphocytes recovered from frozen storage. *Tissue Antigens* 1976; **7**: 165-172.
- 17 Woollard M, Allen M: 1881 Census for England, Wales, The Channel Islands and the Isle of Man: Introductory user Guide V 0.4. Distributed by the History Data Service, UK Data Archive. 1991.
- 18 Burt J, Barber G, Rigby D: *Elementary Statistics for Geographers*. Third Edition. The Guilford Press, London, 2009.
- 19 Bodmer JG, De Luca M, Moses JH, Heyes JM, Marsh SGE: HLA Class I and II study in an Orcadian population. *Hum Immunology* 1996; **47**: 51.
- 20 Bodmer WF, Bodmer JG: Evolution and function of the HLA system. *Br Med Bull* 1978; **34**: 309-316.
- 21 Solberg OD, Mack SJ, Lancaster AK et al.: Balancing selection and heterogeneity across the classical human leukocyte antigen loci: a meta-analytic review of 497 population studies. *Hum Immunol* 2008; **69**: 443-464.

- 22 Evseeva I, Nicodemus KK, Bonilla C, Tonks S, Bodmer WF: Linkage disequilibrium and age of HLA region SNPs in relation to classic HLA gene alleles within Europe. *E J Hum Gen* 2010; **18**: 924-932.
- 23 Rees JL: The melanocortin 1 receptor (*MC1R*): more than just red hair. *Pigment Cell Res* 2000; **13**: 135-140.
- 24 Harding RM, Healy E, Ray AJ C et al.: Evidence for variable selective pressures at *MC1R*. *Am J Hum Genet* 2000; **66**: 1351-1361.
- 25 Gerstenblith MR, Goldstein AM, Fargnoli MC, Peris K, Landi MT: Comprehensive evaluation of allele frequency differences of *MC1R* variants across populations. *Hum Mutat* 2007; **28**: 495-505.
- 26 Mourant AE, Kopeć AC, Domaniewska-Sobczak K: The distribution of human blood groups. Oxford University Press, 1976.
- 27 Garratty G, Glynn SA, McEntire R; Retrovirus Epidemiology Donor Study: ABO and Rh(D) phenotype frequencies of different racial/ethnic groups in the United States. *Transfusion* 2004; **44**: 703-706.
- 28 Wells RS, Yuldasheva N, Ruzibakiev R et al.: The Eurasian heartland: a continental perspective on Y-chromosome diversity. *Proc Natl Acad Sci USA* 2001; **98**: 10244-10249.
- 29 Y Chromosome Consortium: A nomenclature system for the tree of human Y-chromosomal binary haplogroups. *Genome Res* 2002; **12**: 339-348.
- 30 Weir BS, Cockerham CC: Estimating F-statistics for the analysis of population structure. *Evolution* 1984; **38**: 1358–1370.
- 31 Winney B, Litton C, Parkin D, Feare C: The subspecific origin of the inland breeding colonies of the cormorant *Phalacrocorax carbo* in Britain. *Heredity* 2001; **86**: 45-53.

## **ACKNOWLEDGEMENTS**

This project was funded by the Wellcome Trust and the authors would like to acknowledge Lon Cardon and David Bicknell for their help in putting the grant application together. Sorrel May and Nick Godwin helped to kick start the large-scale recruitment of volunteers, whilst 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 Urbaniak, 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.

## **CONFLICT OF INTEREST**

The authors declare no conflict of interest.

## TITLES AND LEGENDS TO FIGURES

**Figure 1.** Graph of the Log(MLQ) of the RD with the highest LQ for each surname (y-axis) against Log(surname population size) in the 1881 census (x-axis). There are a number of surnames (circled) with a higher MLQ than 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.

**Figure 2.** Distribution of mean geographic position (MGP) of grandparental birthplaces of the 3,646 volunteers for whom there was information for all four grandparents. Dots mark the MPG for individual volunteers. The populations from which samples were taken for the genotyping are marked on the inset map.

**Figure 3.** Percentage of volunteers with all four grandparents classed as rural according to their distance (2km, 5km or 10km) from an urban area (y axis) of a given population size (x-axis). Estimates are made for all the geocoded samples (All samples) and those genotyped (Pilot samples).

	<i>Overall</i>	<i>Proportion</i>	<i>Pilot</i>	<i>Proportion</i>
<i>Gender</i>				
<i>M</i>	1824	0.472	506	0.479
<i>F</i>	1982	0.513	497	0.470
<i>Unknown</i>	59	0.015	54	0.051
<i>Total</i>	3865		1057	
<i>Age (in 2009)</i>				
<i>&lt;20</i>	8	0.002	0	0.000
<i>20-29</i>	82	0.021	13	0.012
<i>30-39</i>	180	0.047	33	0.031
<i>40-49</i>	462	0.120	66	0.062
<i>50-59</i>	688	0.178	172	0.163
<i>60-69</i>	1161	0.300	295	0.279
<i>70-79</i>	915	0.237	246	0.233
<i>80-89</i>	291	0.075	96	0.091
<i>90-99</i>	21	0.005	12	0.011
<i>&gt;100</i>	10	0.003	2	0.002
<i>unknown</i>	47	0.012	122 <sup>1</sup>	0.115
<i>Total</i>	3865		1057	
<i>MD between grandparental birthplaces</i>				
<i>Median (km)</i>	16.05		16.31	
<i>25% Quartile (km)</i>	2.96		3.72	
<i>75% Quartile (km)</i>	44.85		48.92	
<i>n</i>	3646		893	
<i>No. missing</i>	219		65	
<i>Orkney<sup>1</sup></i>	0		99	

**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. 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.

<sup>1</sup>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.

	NRY		R1		F		E	n
	R1a1	I	J2	xR1a1	xI/J2/R1			
SW	0.013	0.150	0.025	0.750	0.038	0.025	80	
CN	0.053	0.140	0.018	0.684	0.105	0.000	57	
E	0.035	0.161	0.023	0.598	0.138	0.046	87	
N	0.022	0.202	0.033	0.656	0.071	0.016	183	
OR	0.342	0.079	0.000	0.579	0.000	0.000	38	

  

	HLA-A				HLA-B				HLA-C							
	01	02	03	11	2n	07	08	15	35	44	2n	03	05	06	07	2n
SW	0.169	0.307	0.156	0.055	326	0.151	0.170	0.077	0.071	0.106	311	0.123	0.141	0.083	0.368	326
CN	0.191	0.270	0.157	0.071	267	0.139	0.147	0.053	0.090	0.109	266	0.167	0.104	0.089	0.333	270
E	0.172	0.304	0.175	0.052	326	0.134	0.130	0.103	0.103	0.090	322	0.175	0.089	0.092	0.316	326
N	0.177	0.271	0.151	0.076	661	0.186	0.105	0.060	0.081	0.124	651	0.131	0.122	0.087	0.366	666
OR	0.183	0.291	0.091	0.080	175	0.222	0.090	0.084	0.030	0.204	167	0.171	0.114	0.091	0.381	176

  

	HLA-DRB1				HLA-DQB1								
	01	03	04	07	11	13	15	2n	02	03	05	06	2n
SW	0.106	0.156	0.240	0.100	0.065	0.109	0.122	321	0.225	0.338	0.184	0.225	320
CN	0.105	0.165	0.173	0.102	0.094	0.109	0.132	266	0.229	0.342	0.150	0.259	266
E	0.079	0.142	0.195	0.145	0.085	0.101	0.123	318	0.280	0.341	0.137	0.213	314
N	0.116	0.144	0.177	0.147	0.055	0.090	0.164	654	0.262	0.308	0.150	0.253	652
OR	0.061	0.141	0.160	0.184	0.043	0.086	0.209	163	0.267	0.320	0.111	0.273	172

  

	HLA-A, -B, DRB1 Haplotypes						
	01-08-03	03-07-15	02-44-04	02-07-15	02-15-04	29-44-07	2n
SW	0.076	0.023	0.030	0.020	0.030	0.003	304
CN	0.088	0.023	0.012	0.012	0.008	0.008	260
E	0.077	0.026	0.016	0.026	0.019	0.016	310
N	0.062	0.019	0.021	0.015	0.011	0.011	623
OR	0.051	0.013	0.044	0.051	0.013	0.044	158

  

	MC1R (rs1805007)		MC1R (rs1805008)		ABO (rs7853989)	
	C	2n	C	2n	C	2n
SW	0.906	340	0.945	328	0.929	328
CN	0.924	264	0.909	264	0.893	270
E	0.908	326	0.932	310	0.920	326
N	0.910	652	0.934	664	0.914	660
OR	0.887	194	0.906	192	0.828	192

**Table 2.** Allele and haplotype frequency data. The *NRY* haplogroups are those that are the most common in Europe, whilst the *HLA* alleles (low, allele group, resolution) are those that have a frequency of >7.5% in at least one region. The estimated frequencies of the six most common *HLA-A, -B, -DRB1* haplotypes are also shown. Only the major allele frequencies are presented for the *MC1R* and *ABO* SNPs. Populations are grouped into regions as defined in the main text. The regions are: SW (Cornwall, Devon and Pembrokeshire), CN (Oxfordshire and the Forest of Dean), E (Sussex, Kent, Norfolk and Lincolnshire), N (Cumbria, Yorkshire and the North East) and OR (Orkney).



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

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

'Admixed' population	Parental populations	Local (L) or Non-local (N)	Proportion East	-95% CI	+95% CI
<i>LQ&gt;120, Distance&lt;83km<sup>1</sup></i>					
CN	West vs East	L	0.945	0.895	0.995
CN	West vs East	N	0.630	0.591	0.669
OR	West vs East	L	0.550	0.488	0.614
OR	West vs East	N	0.695	0.630	0.760
<i>LQ&gt;19, Distance&lt;120km<sup>2</sup></i>					
CN	West vs East	L	0.900	0.829	0.971
CN	West vs East	N	0.525	0.482	0.568
OR	West vs East	L	0.360	0.265	0.455
OR	West vs East	N	0.815	0.761	0.869
'Admixed' population	Parental populations	Local (L) or Non-local (N)	%Norse	-95% CI	+95% CI
<i>LQ&gt;120, Distance&lt;83km<sup>1</sup></i>					
OR	Norse vs West	L	0.375	0.331	0.419
OR	Norse vs West	N	0.405	0.357	0.453
<i>LQ&gt;19, Distance&lt;120km<sup>2</sup></i>					
OR	Norse vs West	L	0.315	0.266	0.364
OR	Norse vs West	N	0.420	0.375	0.465

**Table 4.** Maximum likelihood admixture estimates for the <sup>1</sup>most stringent and the <sup>2</sup>least stringent criteria used to define local and non-local surnames. 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 *NRV*, all the males were used as the sample sizes were too small to divide into local and non-local. For the Orkney analysis, all Orcadian samples were compared to either local or non-local stratified PoBI samples.







