Using ancestry-informative markers to identify fine structure across 15 populations of European origin


Published in:
European Journal of Human Genetics

Document Version:
Publisher's PDF, also known as Version of record

Queen's University Belfast - Research Portal:
Link to publication record in Queen's University Belfast Research Portal

Publisher rights
©2014 Macmillan Publishers Limited All rights reserved
This work is licensed under a Creative Commons Attribution 3.0 Unported License. To view a copy of this license, visit http://creativecommons.org/licenses/by/3.0/which permits unrestricted use, distribution and reproduction in any medium, provided the author and source are cited.

General rights
Copyright for the publications made accessible via the Queen's University Belfast Research Portal is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The Research Portal is Queen's institutional repository that provides access to Queen's research output. Every effort has been made to ensure that content in the Research Portal does not infringe any person's rights, or applicable UK laws. If you discover content in the Research Portal that you believe breaches copyright or violates any law, please contact openaccess@qub.ac.uk.
Using ancestry-informative markers to identify fine structure across 15 populations of European origin

Laura M Huckins*,1, Vesna Boraska1,2, Christopher S Franklin1, James AB Floyd1, Lorraine Southam1, GCAN3, WTCCC3, Patrick F Sullivan3, Cynthia M Bulik3, David A Collier4, Chris Tyler-Smith1, Eleftheria Zeggini1,7 and Ioanna Tachmazidou1,7

The Wellcome Trust Case Control Consortium 3 anorexia nervosa genome-wide association scan includes 2907 cases from 15 different populations of European origin genotyped on the Illumina 670K chip. We compared methods for identifying population stratification, and suggest lists of markers that may help to counter this problem. It is usual to identify population structure in such studies using only common variants with minor allele frequency (MAF) > 5%; we find that this may result in highly informative SNPs being discarded, and suggest that instead all SNPs with MAF > 1% may be used. We established informative axes of variation identified via principal component analysis and highlight important features of the genetic structure of diverse European-descent populations, some studied for the first time at this scale. Finally, we investigated the substructure within each of these 15 populations and identified SNPs that help capture hidden stratification. This work can provide information regarding the designing and interpretation of association results in the International Consortia.

European Journal of Human Genetics (2014) 22, 1190–1200; doi:10.1038/ejhg.2014.1; published online 19 February 2014

Keywords: population stratification; AIMs; principal component analysis

INTRODUCTION

Population stratification can be a major cause of concern in genetic association studies. Specifically, imperfect matching between cases and controls can lead to spurious associations, or failure to detect true associations.1 Several ways of accounting for hidden population stratification have been proposed (genomic control (GC) correction, adjusting for ancestry-informative principal components (PCs)), but these approaches are only applicable in genome-wide scale data. The GC approach uses genomics features of the samples to correct for stratification, and thus avoids inflation in the test statistic.1 Population stratification may lead to ‘overdispersion’ of the statistics used to test for association; by measuring several polymorphisms across the genome, the degree of this overdispersion may be estimated and taken into account. However, GC may not perform well with too few loci, or may overcorrect and lead to a substantial loss in power.1 Menozzi et al.8 described the use of PC analysis (PCA) in human genetics in 1978. PCA summarizes high-dimensionality data by capturing the latent variables that best describe a data set, allowing simple visualization of allele frequency differences among populations. It is possible to correlate PCs of the data with meaningful geographic axes. For example, genetic variation in the first two PCs is closely associated with geographic alignment across Europe.1–6 As with GC, PCA may also be used to correct for population stratification when working with a very large number of markers, ideally genome-wide data sets. However, population stratification is much of a concern in replication studies or studies focusing on a smaller number of variants, in which GC or PCs cannot be readily calculated. To circumvent this problem, adjustment for the genotypes of ancestry-informative markers (AIMs) has been proposed as an alternative approach.

Shriver et al.7 proposed that certain markers with distinct frequency differences across populations may be highly informative for assigning ancestry. These markers are referred to as AIMs. A small number of these AIMs may be used to perform population clustering; between 40 and 80 loci, Rosenberg et al.8 demonstrates convergence to five broad continental clusters. Kidd et al.8 used 128 AIMs to characterize samples from 119 populations into 8 broad clusters, which agree with continental boundaries. Precalculated lists of AIMs are available, although these are mostly applicable only to cross-continental studies,10,11 or require a relatively large set of SNPs.12

A different way to derive AIMs is to identify SNPs that contribute highly to the significant PCs (PCAIMs), as first discussed by Paschou et al.13 SNPs that contribute heavily to the underlying axes of variation will be instrumental in clustering samples along population lines; it follows that these SNPs may be used to assign ancestry. A recent study has identified these PCAIMs for samples of North-Central European and Mediterranean origin, and has shown that they may be used to assign sample ancestry.14

In this work, we investigated the structure across closely related European populations. We discuss evidence for stratification using PCA and Fst, a measure of genetic distance among samples. Further, we identified lists of AIMs and PCAIMs, which are able to correct for stratification by using a small number of markers.

1The Wellcome Trust Sanger Institute (WTSI), Hinxton, UK; 2University of Split School of Medicine, Split, Croatia; 3University of North Carolina, Chapel Hill, NC, USA; 4King’s College, London, UK
5GCAN members are listed before the references.
6WTCCC3 members are listed before the references.
7These authors contributed equally to this work.
*Correspondence: Ms LM Huckins, The Wellcome Trust Sanger Institute (WTSI), Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK. Tel: +44 (0)1223 834244; Fax: +44 (0)1223 496826; E-mail: lh10@sanger.ac.uk
Received 1 March 2013; revised 24 October 2013; accepted 9 November 2013; published online 19 February 2014
We investigated population stratification using data taken from the Wellcome Trust Case Control Consortium 3 anorexia nervosa (AN) genome-wide association scan, which includes 2907 cases from 15 different populations of European origin (unpublished data). Thirteen of these are European, and are divided between Scandinavian (Finland, Norway and Sweden), North-Central European (Czech Republic, France, Germany, the Netherlands, Poland and the United Kingdom) and Mediterranean populations (Greece, North Italy, South Italy and Spain). Two further populations of European origin included in this study are United States and Canada. Sample sizes range from 39 (Swedish samples) to 475 (Germany); numbers of samples are shown in Figure 1 and Table 1. Populations were genotyped on the Illumina 670K chip.

We discuss the fine structure within these populations, and identify a set of informative SNPs. We compare different methods of calculating these, and assess their usefulness in assigning samples to populations.

MATERIALS AND METHODS
Sample collection
We used samples that had been collected for an AN GWAS. The samples comprise 15 discovery data sets of European origin. All samples used were female. All samples met the DSM-IV diagnostic criteria for lifetime AN or lifetime 'eating disorder not otherwise specified', with the exception of the requirement for amenorrhoea. Samples with a lifetime history of bulimia nervosa were also included in the data set.

Genotyping
All cases were genotyped using the Illumina 660W-Quad arrays (Illumina Inc., San Diego, CA, USA) at the Wellcome Trust Sanger Institute. Quality control was performed individually on each of the 15 case–control subgroups (Supplementary Information).

PCA
We calculated PCs using the smartpca software (developed at Harvard School of Public Health, Boston, MA, USA).15 We identified the top PCs by selecting those components that explained the greatest variance. We used the Tracy–Widom (TW) statistic to assess the significance of each PC. The TW statistic tests whether the average eigenvector coordinates across all samples within each population differ significantly across components. We found that the first six PCs differ significantly (TW statistic >100, P<10-86).

Geographic relevance of PCs
We applied three different tests to calculate the geographic relevance of the PCs. To do this, we first computed the mean eigenvector coordinates of all samples within a population. We then compared these to the centre of genetic variance to the geographic centre. As our samples were obtained from tertiary referral centres, we define 'Geographic centre' as the geographical midpoint of the country from which the samples were taken. Coordinates were obtained in the same way by Novembre et al,3 the same coordinates are used here, with the exception of North Italy, which is assigned Verona as its geographic centre.

We then performed the following correlation tests:

1) We used a Spearman's rank correlation coefficient to test for significance of association. Spearman's rank correlations were computed using a standard R package.

2) We applied a Mantel test. This test calculates the correlation between the two distance matrices, and then computes an empirical P-value by randomly permuting the rows and columns of one matrix. We performed the Mantel test using the 'ape' R package16 and used 1000 permutations (as recommended).

3) We applied a Procrustes test. This works in the same way as the Mantel test, but is likely to be more sensitive.17,18 We performed the Procrustes test using the 'vegan' package in R19 with 1000 permutations (as recommended).

FST
Tian et al20 assign a threshold of Fst = 0.001, below which populations may not be said to be genetically distinct.

Fst values were computed using the smartpca software.15 To test the correlation between Fst (genetic distance) and geographic distance between population centres, we applied a Mantel test, as for the PCA data.

AIMs
AIMs are defined as markers that provide information as to the ancestry of a sample. Informativeness describes the amount of information that is imparted by the marker. We use a harmonized data set of 70 samples per population to calculate informativeness. We selected 70 samples per population to avoid any sample-size associated bias in the Informativeness calculation.

Samples were selected at random from all populations; note that Sweden (39 samples) and Canada (54 samples) were omitted owing to small population sizes. The remaining samples were designated as a testing set, to validate AIMs. The Swedish population was set aside to test the ability of AIMs (and PCAIMs) to assign ancestry of samples from a new population.

AIMs were thinned for LD using PLINK.21,22 A threshold of 0.8 was used.

Table 1 Sample sizes per population

<table>
<thead>
<tr>
<th>Population</th>
<th>Abbreviation</th>
<th>Sample size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Canada</td>
<td>CA</td>
<td>54</td>
</tr>
<tr>
<td>Czech Republic</td>
<td>CZ</td>
<td>72</td>
</tr>
<tr>
<td>Finland</td>
<td>FI</td>
<td>131</td>
</tr>
<tr>
<td>France</td>
<td>FR</td>
<td>293</td>
</tr>
<tr>
<td>Germany</td>
<td>DE</td>
<td>475</td>
</tr>
<tr>
<td>Greece</td>
<td>GR</td>
<td>70</td>
</tr>
<tr>
<td>North Italy</td>
<td>NIT</td>
<td>203</td>
</tr>
<tr>
<td>Netherlands</td>
<td>NL</td>
<td>348</td>
</tr>
<tr>
<td>Norway</td>
<td>NO</td>
<td>82</td>
</tr>
<tr>
<td>Poland</td>
<td>PL</td>
<td>175</td>
</tr>
<tr>
<td>South Italy</td>
<td>SIT</td>
<td>75</td>
</tr>
<tr>
<td>Spain</td>
<td>ES</td>
<td>186</td>
</tr>
<tr>
<td>Sweden</td>
<td>SE</td>
<td>39</td>
</tr>
<tr>
<td>UK</td>
<td>UK</td>
<td>213</td>
</tr>
<tr>
<td>USA</td>
<td>USA</td>
<td>491</td>
</tr>
</tbody>
</table>
Informativeness was calculated according to Rosenberg et al., using the formula below:

$$I = \sum_{j=1}^{N} \left( -p_j \log p_j + \frac{\sum_{i=1}^{K} p_{ij} \log p_{ij}}{K} \right)$$

where $p_j$ is the mean frequency of allele $j$ across all populations, $p_i$ is the relative frequency of allele $j$ in population $i$, and $K$ is the total number of populations.

**PCAIMs**

PCAIMs were selected using a weighting system as outlined by Raanum et al. SNP contributions to each PC were calculated using smartpca.

Contributions of each SNP to each PC were normalized to the maximum weight, so that the SNPs that contributed most to a PC were given a weight of 1. These weights were multiplied by the corresponding eigenvector. To get a rank for each SNP, weights were summed across all PCs.

AIMs were thinned for LD using PLINK. A threshold of 0.8 was used.

**K-nearest neighbour**

K-nearest-neighbour assignments were used to assess how well AIMs and PCAIMs were able to assign a sample to a certain population. (Here, we used $K = 5$). The KNN algorithm identifies the K-nearest genetic neighbours by computing Euclidean distances between samples. We used PLINK to find each sample’s K-nearest genetic neighbours, based on only a given number of AIMs. Clustering samples that are ‘closest’ together according to a genetic similarity measure, derived by AIMs or PCAIMs, implies that the nearest neighbours share common ancestry with the sample in question. The ancestry of the nearest neighbours was used as a ‘majority vote’ to determine the ancestry of the sample.

In cases where the five nearest neighbours did not reach a majority vote, only the four nearest were selected, and a majority vote again taken. If this was still unsuccessful, only the top three were used. If still no majority vote was reached, the sample was classed as ‘unassigned’.

Ancestry was assigned to a sample based on the result of the majority vote. Each sample was considered correctly assigned if the result of the majority vote was either the true ancestry of the sample or a population with a pair-wise Fst < 0.001 with the true population.

**RESULTS**

**Evidence of structure among populations**

We performed PCA on the 15 population sets, and plotted the PCs for all populations as shown in Figure 2. The first two PCs accounted for 25.2 and 12.99% of the variation in the data, as shown in Table 2. We used the proportion of variance explained, along with the TW statistic as shown in Table 2, to identify significant PCs.

We tested the geographic relevance of the PCs by calculating the correlation between PC magnitude and latitude and longitude, obtained using the geographic centre of each nation, shown in Supplementary Table 1. Canadian and USA samples were not included in this aspect of the study, owing to the difficulty of assigning meaningful geographic locations. We found that the top two PCs were correlated with perpendicular geographical axes ($\rho = 0.90$ for PC1 versus latitude, $\rho = 0.59$ for PC2 versus longitude). After rotation, PC1 aligns north-northwest/south-southwest (NNW/SSE, $-11^\circ$, $\rho = 0.91$). This is remarkably similar to the $-16^\circ$ angle cited by Novembre et al.4 We see no significant correlation between PCs and geographical axes. We tested for significance between PC locations and geographic centres, and found that this was significant for the first and second PCs ($P < 1e^{-300}$ for PC1, $P = 0.036$ for PC2, using a Mantel test; $P = 0.001$ for PC1, $P = 0.015$ for PC2, using a Procrustes test).

Figure 2 presents the first three PCs of the data. Populations form three overlapping subclusters: Finland, central European and Southern or Mediterranean populations. Samples form tight subclusters along population lines, implying that even closely related neighbouring populations are genetically distinct.

USA samples cluster loosely across North-Central European and Scandinavian populations, with some samples clustering with the Mediterranean population. As expected, we see little overlap between Finnish and USA samples. Canadian samples tend to cluster with North-Central European and Scandinavian populations. We performed a PCA using only USA, Canadian, North-Central and Scandinavian populations (therefore removing Mediterranean and Finnish samples), to illustrate this more clearly, as shown in Supplementary Figure 1. This figure confirms the substantial overlap between USA, Canadian and North-Central and Scandinavian populations.

We calculated genetic distance among populations by means of the Fst statistic (Table 3). Fst correlated well with distance in kilometres between populations (Figure 3) when using the geographic centres of the populations given in Supplementary Table 1. We found a significant correlation between distance in kilometres and Fst (using a Mantel test, $P < 1e^{-300}$).

It is clear from Table 3 that a number of pair-wise comparisons between populations show only a very low Fst value. We used a threshold Fst value of 0.001 to identify pairs of populations that are not genetically distinct; this may be owing to recent admixture or shifting of national borders. Pairs of populations that fall below this threshold are shaded in Table 3.

**AIM derivation**

We extracted a list of AIMs using Rosenberg’s informativeness calculation, using a harmonized data set of 70 samples per population (for a brief description see the Materials and Methods section). We used 70 samples per population to avoid over-representing populations with larger sample sizes. Populations with fewer than 70 samples were not used to calculate AIMs.

We calculated AIMs using all SNPs with average minor allele frequency (MAF) across all populations > 1%. Although it is usual to take 5% as a lower boundary, we find that this risks removing highly informative markers. For example, consider the ‘perfect’ marker, which appears in every sample of one population, and not at all in others. For the harmonized set of 13 populations, this marker would have an average MAF of 3.8% across all populations, and would be dismissed under a 5% threshold. We show the top 25 most informative markers in Supplementary Table 2, along with their average MAF. Note that 7 out of these top 25 markers have an average MAF < 5%.

One caveat when using AIMs is that populations might not contribute evenly to the choice of markers. A large number of our samples originated from central Europe; although these are classified into distinct populations, we have already shown that some of these populations are very closely related (e.g. France and Germany); meanwhile, there were a smaller number of samples from an outlying population (Finland). To ensure that AIMs were chosen evenly to represent all populations, we computed the AIMs using only 12 of 13 populations. We repeated this 13 times, leaving a different population out each time. For each new set of AIMs, we computed the Spearman’s rank correlation coefficient with the original list (Table 4). We found an average $p = 0.97$, although it may be noted that the correlation is slightly lower ($p = 0.907$) for the set excluding Finland. The high correlations indicate that no single population is over-represented. The lower correlation when excluding the Finnish samples is owing to the greater genetic distance between Finland and other populations.

We use a weighting system as discussed by Raanum et al. to select PCAIMs; the top 25 are shown in Supplementary Table 3. We noted...
that a number of these SNPs fall into clusters (15 of the top 25 cluster on chr. 2, 4 cluster on chr. 15). These locations are associated with geographically restricted positive selection throughout Europe, implying that many of these SNPs may be reflecting the same past event, and may thus not be truly independent. To select SNPs that provide the maximum possible information, we selected only the most informative SNP from each cluster, as shown in Supplementary Table 4.

Validation of AIMS/PCAIMS
We validated the top AIMs and PCAIMs by testing their ability to assign ancestry to new samples. We used the samples not included in the 70-sample per population harmonized data set; any population with more than 10 samples remaining was included in the validation set.

We used K-nearest-neighbour algorithms to identify possible ancestry of the samples (for a brief description see the Materials and Methods section.

Both AIMs and PCAIMs were able to assign ancestry to samples with a high accuracy, even at small numbers of markers. For example, both AIMs and PCAIMs predicted about 90% of the total samples correctly using only 25 markers, although some populations are not predicted well (Spain, Finland and Poland) (Figure 4a).

It may be noted that PCAIMS predict outlying populations better than AIMS. A key example of this is the performance of both sets of markers when predicting Finnish samples (Figure 4b); AIMs predict no samples correctly, even at larger numbers of markers. This failure is due to the way in which AIMs are assigned. We observe high genetic similarity between some central European populations, for example, Czech Republic, France, Germany and Netherlands (as
illustrated by low pair-wise Fst values in Table 3). This indicates that a marker that predicts a French sample well will also predict a German sample well. As a sample is considered to be correctly assigned if the final assignment is the original population, or a population with pair-wise Fst $<0.001$, markers that predict French samples well will also predict German samples well, and will thus increase the number of samples correctly assigned for these populations. In this way, we effectively have 280 samples contributing to 'Czech/French/German/Dutch' ancestry, as opposed to just 70 Finnish samples. This ties in well with Table 4, as removing any of these four populations still gives a very high correlation of AIMS ($r = 0.98$). PCAIMs, on the other hand, predict Finnish samples better as they take into account the underlying variation of the data, rather than just the entropy of allelic frequency across samples.

Table 3 Pair-wise Fst calculated between all populations

<table>
<thead>
<tr>
<th></th>
<th>CZ</th>
<th>DE</th>
<th>ES</th>
<th>FI</th>
<th>FR</th>
<th>GR</th>
<th>NIT</th>
<th>NL</th>
<th>NO</th>
<th>PL</th>
<th>SIT</th>
<th>UK</th>
</tr>
</thead>
<tbody>
<tr>
<td>CZ</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DE</td>
<td>0.003</td>
<td>0.002</td>
<td></td>
<td>0.007</td>
<td>0.011</td>
<td>0.008</td>
<td>0.001</td>
<td>0.001</td>
<td>0.003</td>
<td>0.013</td>
<td>0.003</td>
<td>0.001</td>
</tr>
<tr>
<td>ES</td>
<td>0.006</td>
<td>0.007</td>
<td>0.011</td>
<td>0.008</td>
<td>0.001</td>
<td>0.003</td>
<td>0.003</td>
<td>0.001</td>
<td>0.003</td>
<td>0.003</td>
<td>0.001</td>
<td>0.003</td>
</tr>
<tr>
<td>FI</td>
<td>0.004</td>
<td>0.004</td>
<td>0.003</td>
<td>0.013</td>
<td>0.003</td>
<td>0.001</td>
<td>0.005</td>
<td>0.006</td>
<td>0.007</td>
<td>0.007</td>
<td>0.001</td>
<td>0.007</td>
</tr>
<tr>
<td>FR</td>
<td>0.005</td>
<td>0.002</td>
<td>0.003</td>
<td>0.002</td>
<td>0.014</td>
<td>0.003</td>
<td>0.001</td>
<td>0.003</td>
<td>0.003</td>
<td>0.003</td>
<td>0.001</td>
<td>0.003</td>
</tr>
<tr>
<td>GR</td>
<td>0.001</td>
<td>0.001</td>
<td>0.003</td>
<td>0.007</td>
<td>0.003</td>
<td>0.001</td>
<td>0.005</td>
<td>0.006</td>
<td>0.007</td>
<td>0.007</td>
<td>0.001</td>
<td>0.007</td>
</tr>
<tr>
<td>NIT</td>
<td>0.002</td>
<td>0.001</td>
<td>0.004</td>
<td>0.006</td>
<td>0.002</td>
<td>0.003</td>
<td>0.007</td>
<td>0.007</td>
<td>0.007</td>
<td>0.007</td>
<td>0.001</td>
<td>0.007</td>
</tr>
<tr>
<td>NL</td>
<td>0.002</td>
<td>0.001</td>
<td>0.005</td>
<td>0.006</td>
<td>0.003</td>
<td>0.006</td>
<td>0.007</td>
<td>0.007</td>
<td>0.007</td>
<td>0.007</td>
<td>0.001</td>
<td>0.007</td>
</tr>
<tr>
<td>NO</td>
<td>0.001</td>
<td>0.002</td>
<td>0.001</td>
<td>0.011</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.003</td>
<td>0.004</td>
<td>0.001</td>
<td>0.003</td>
</tr>
<tr>
<td>PL</td>
<td>0.001</td>
<td>0.002</td>
<td>0.001</td>
<td>0.011</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.003</td>
<td>0.004</td>
<td>0.001</td>
<td>0.003</td>
</tr>
<tr>
<td>SIT</td>
<td>0.001</td>
<td>0.002</td>
<td>0.001</td>
<td>0.011</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.003</td>
<td>0.004</td>
<td>0.001</td>
<td>0.003</td>
</tr>
<tr>
<td>UK</td>
<td>0.001</td>
<td>0.002</td>
<td>0.007</td>
<td>0.008</td>
<td>0.002</td>
<td>0.003</td>
<td>0.005</td>
<td>0.005</td>
<td>0.005</td>
<td>0.005</td>
<td>0.001</td>
<td>0.004</td>
</tr>
<tr>
<td>USA</td>
<td>0.001</td>
<td>0.002</td>
<td>0.007</td>
<td>0.008</td>
<td>0.002</td>
<td>0.003</td>
<td>0.005</td>
<td>0.005</td>
<td>0.005</td>
<td>0.005</td>
<td>0.001</td>
<td>0.004</td>
</tr>
</tbody>
</table>

Swedish and Canadian samples are not included here owing to small sample sizes. Population pairs falling below the Fst $<0.001$ threshold are in pink.

Figure 3 Genetic distance correlates with geographical distance. We computed pair-wise Fst between all populations, and compared this to the geographic distance in kilometres between the midpoints of each population. $R^2 = 0.465$. 
Population stratification can have a major negative impact on genetic association studies, whether by creating spurious results or by obscuring true associations. This stratification may be corrected using AIMs, only applicable on a genome-wide scale. An alternative approach to assigning AIMs (such as Canada and the Czech Republic). Further, these populations are more closely related than those used previously and span a wider geographic range than those seen in recent studies. For example, we include two Scandinavian populations (Norway and Sweden) and two eastern European populations (Czech and Polish), which are usually clustered into one population. We saw a geographical alignment of our first three PCs. Further, populations cluster along meaningful geographic and cultural lines. We see three broad clusters consisting of Finland, North-Central Europe and Scandinavia, and Mediterranean populations. USA samples cluster largely with North-Central European and Scandinavian samples, with a few clustering with Italian samples, consistent with migratory patterns from Europe to North America.

It appears that Canadian samples cluster closely with French samples; we investigated this in more detail and found that Canadian samples fell into two groups: a tight cluster, which corresponded with the French samples, and a loose cluster, which did not lie close to French samples. This is consistent with some of our samples being of French-Canadian heritage, rather than simply of central European backgrounds.

We also found evidence for substructure within the USA population. We found three broad clusters when performing a PCA plot. We found that most samples cluster with the North-Central European populations (likely to correspond to the largest cluster on our PCA plot), but that there is also a distinct group stemming from Mediterranean populations. This is likely to be due to immigration patterns to the United States. Our third and smallest cluster on the PCA plot is likely to represent a mix of Finnish samples and samples with joint Scandinavian and North-Central European heritage.

We found a correlation between genetic distance, Fst, and the geographic distance between populations. This fits well with the clusters obtained using PCA, and is likely due to admixture between neighbouring populations. In addition, we see very low Fst values between certain pairs of populations, for example, France, Germany and the Netherlands. It is likely that this is due to a lack of significant geographical boundaries in these regions, for example, the Pyrenees or the Alps, and due to shared territories and shifting empire boundaries.

We obtained two lists of AIMs: one list was calculated using Rosenberg’s informativeness calculation, and the other using Raum’s PCAIMs. Our initial list of 25 PCAIMs shows that SNPs cluster around three loci, corresponding to lactase and pigmentation-associated loci, HERC2 and OCA2. These genes are classic examples of positively selected genes in European populations, indicating that some of our PCAIMs are picking up high levels of differentiation due to geographically restricted positive selection, rather than due to neutral genetic drift.

Using only a small number of markers, both AIMs and PCAIMs were able to predict sample origin accurately. A key difference between the two sets is the ability to predict ancestry of outlying populations; in this case, PCAIMs outperform AIMs. This is likely to be due to how AIMs and PCAIMs are identified. For example, PCAIMs are chosen to represent the underlying variance of all samples; for our data set, a large part of this variance exists between central European populations and outlying populations (eg, Finland and Spain). As PCAIMs are chosen to explain this variance, even a small number of markers are able to predict outlying populations well.

AIMs, on the other hand, are chosen from markers with a high variance across populations. In this instance, we treat individual populations as independent, and select markers, which explain equally well the difference between all these populations. This is obviously a problem with closely related populations; we can see from PCA graphs that central European populations are in fact not independent;
further, we have a much larger number of central European populations than outlying populations, causing a skew towards markers that predict central European populations well. This difference between the two sets becomes more pronounced when looking at larger numbers of markers. For example, using 500 or 1000 AIMs performs better than PCAIMs in predicting central European nations (ie, in very fine detail), but lag significantly in predicting the ancestry of outlying populations.

We used our lists of markers to assign ancestry to samples from a new population (Sweden), and assessed the ability of our markers to assign ancestry to these samples. Both sets of markers performed well, although PCAIMs perform better than AIMs.

A small proportion of Swedish samples are unassigned using AIMs, whereas all are assigned using PCAIMs. This is likely to be due to the fact that AIMs have been chosen to explain specific differences between a certain set of populations – they may be thought of as discrete measures of differences between populations. PCAIMs, on the other hand, are chosen to represent the continuum of variation. In this respect, we conclude that PCAIMs are better able to explain the ancestry of a new population, as long as it lies on the same continuum.

It is worth bearing in mind the intrinsic limitations of our data set, which consists of clinical samples, obtained by the WTCCC3 for an AN GWAS. Although we have a large number of samples, these have been collected for clinical purposes, rather than for use in population genetics. For this reason, detailed information on ancestry is not always available. Further, samples have been accepted, or excluded, based on clinical relevance and guidelines, rather than based on information about their ancestry. For these reasons, our data may not

---

**Figure 4** AIMs and PCAIMs are able to predict sample ancestry with high accuracy for most populations, even at small numbers of markers. (a) Percent of samples correctly assigned using 25 markers, across all populations. AIMs are shown in green, PCAIMs in blue. (b) Assignment of Finnish samples, for varying numbers of markers. AIMs are shown as a solid line and PCAIMs as a dashed line. (c) Assignment of German samples, with increasing numbers of markers. (d) Assignment of Swedish samples, using 25 markers; AIMs are shown in green and PCAIMs in blue.
be as evenly distributed or as well defined as that used in previous population differentiation studies, in which it is usually required that all four grandparents of the sample are also from the region. Further, many anthropological studies focus on rural samples, whereas our samples are statistically more likely to be urban rather than rural. This can also be considered a strength of the study, showing the power of the method to assign ancestry even in a clinically based sample series, which perhaps would not be expected to display the population structure seen in grandparental sampling schemes.

In summary, we derive a set of 25 PCAIMs that can be used to adjust for population stratification within European samples. By genotyping these markers in replication experiments of large-scale genetic association studies, spurious associations arising owing to ancestry differences can be identified and corrected.

CONFLICT OF INTEREST
The authors declare no conflict of interest.

ACKNOWLEDGEMENTS
LMH, VB, CSF, JABF, LS, CTS, IT and EZ were supported by the Wellcome Trust (098051). LMH is also supported by the MRC (MR/J500355/1), VB is also supported by Unity Through Knowledge Fund CONNECTIVITY PROGRAM (‘Gaining Experience’ Grant 2A), The National Foundation for Science, Higher Education and Technological Development of the Republic of Croatia (BRAIN GAIN- Postdoc fellowship) and CMB is supported by the Foundation of Hope.
Affiliations

1 Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK
2 University of Split School of Medicine, Split, Croatia
3 William Harvey Research Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, John Vane Science Centre, Charterhouse Square, London, UK
4 Department of Psychiatry, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
5 Wellcome Trust Centre for Human Genetics (WTCHG), University of Oxford, Oxford, UK
6 Oxford Centre for Diabetes, Endocrinology and Metabolism (OCDEM), Oxford, UK
7 Department of Psychology, Michigan State University, East Lansing, MI, USA
8 Section of Eating Disorders, Institute of Psychiatry, King's College London, London, UK
9 Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, King's College London, London, UK
10 Health Services Research Unit, University of Aberdeen, Aberdeen, UK
11 Department of Child and Adolescent Psychiatry, Psychosomatics and Psychotherapy, Universitätsklinikum Essen, University of Duisburg-Essen, Essen, Germany
12 INSERM U894, Centre of Psychiatry and Neuroscience, Paris, France
13 Sainte-Anne Hospital (CMME), University of Paris-Descartes, Paris, France
14 Brain Center Rudolf Magnus, Department of Translational Neuroscience, University Medical Center Utrecht, Utrecht, The Netherlands
15 Altrecht Eating Disorders Rintveld, Zeist, The Netherlands
16 Department of Neurosciences, University of Padova, Padova, Italy
17 Department of Psychiatry and CIBERON, University Hospital of Bellvitge-IDIBELL, Barcelona, Spain
18 Department of Clinical Sciences, School of Medicine, University of Barcelona, Barcelona, Spain
19 Genomics and Disease Group, Centre for Genomic Regulation (CRG), Barcelona, Spain
20 Universitat Pompeu Fabra (UPF), Barcelona, Spain
21 Centro de Investigación Biomédica en Red en Epidemiología y Salud Pública (CIBERESP), Barcelona, Spain
22 Hospital del Mar Medical Research Institute (IMIM), Barcelona, Spain
23 Department of Child and Adolescent Psychiatry, Institute of Psychiatry and Neurology, Warsaw, Poland
24 Department of Child and Adolescent Psychiatry, Department of Psychiatry, Poznan University of Medical Sciences, Poznan, Poland
25 Hjelt Institute, University of Helsinki, Helsinki, Finland
26 Institute of Molecular Medicine, University of Helsinki, Helsinki, Finland
27 Department of Mental Health and Substance Abuse Services, National Institute for Health and Welfare, Helsinki, Finland
28 Department of Adolescent Psychiatry, Helsinki University Central Hospital, Helsinki, Finland
29 Center for Eating Disorders Ursula, Leidschendam, The Netherlands
30 Leiden University Medical Centre, Department of Psychiatry, Leiden, The Netherlands
31 Leiden University Medical Centre, Molecular Epidemiology Section (Department of Medical Statistics), Leiden, The Netherlands
32 Department of Psychiatry, McLean Hospital/Harvard Medical School, Belmont, MA, USA
33 Department of Genetics, Environment and Mental Health, Norwegian Institute of Public Health, Oslo, Norway
34 Institute of Clinical Medicine, University of Oslo, Oslo, Norway
35 Department of Psychiatry, University of Naples SUN, Naples, Italy
36 Chair of Psychiatry, University of Salerno, Salerno, Italy
37 Centre for Addiction and Mental Health, University of Toronto, Toronto, Canada
38 Department of Psychiatry, University of Toronto, Toronto, Canada
39 Eating Disorders Unit, Department of Child and Adolescent Psychiatry, Medical University of Vienna, Vienna, Austria
40 The Center for Applied Genomics, The Children's Hospital of Philadelphia, Philadelphia, PA, USA
41 The Division of Human Genetics, Department of Pediatrics, The Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
42 Department of Psychiatry, University of Pennsylvania, Philadelphia, PA, USA
43 Department of Molecular and Experimental Medicine and the Scripps Translational Science Institute, The Scripps Research Institute, La Jolla, CA, USA
44 Department of Psychosomatic Research, National Institute of Mental Health, NCPN, Tokyo, Japan
45 School of Health Sciences at Fukuoka, International University of Health and Welfare, Fukuoka, Japan
46 Department of Molecular Life Sciences, Tokai University School of Medicine, Kanagawa, Japan
47 Estonian Genome Center, University of Tartu, Tartu, Estonia
48 Institute of Molecular and Cell Biology, University of Tartu, Tartu, Estonia
49 Center for Integrative Genomics, University of Lausanne, Lausanne, Switzerland
50 Department of Molecular Physiology and Biophysics, Vanderbilt University School of Medicine, Nashville, TN, USA
51 Department of Genetics, The University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
52 Seattle University College of Nursing, Seattle, WA, USA
53 Kartini Clinic, Portland, OR, USA
54 Centre de Psychiatrie et Neurosciences – Inserm U894, Paris, France
55 UCL Genetics Institute, Department of Genetics, Evolution and Environment, University College London, London, UK
56 Department of Child and Adolescent Psychiatry, Psychosomatics and Psychotherapy, University Clinics RWTH Aachen, Aachen, Germany
57 Department of Child and Adolescent Psychiatry, Psychosomatics and Psychotherapy, Charité, Berlin, Germany
58 Department of Psychosomatic Medicine and Psychotherapy, Hannover Medical School, Hannover, Germany
59 Department of Psychosomatic Medicine and Psychotherapy, University of Erlangen-Nuremberg, Erlangen, Germany
60 Department of Child and Adolescent Psychiatry, Psychosomatics and Psychotherapy, University Würzburg, Würzburg, Germany
61 Department of Child and Adolescent Psychiatry, University Hospital Carl Gustav Carus, Dresden University of Technology, Dresden, Germany
62 Massachusetts General Hospital/Harvard Medical School, Athinoula A. Martinos Center for Biomedical Imaging, Psychiatric Neuroimaging Research Program, Charlestown, MA, USA
63 Departments of Psychosocial and Internal Medicine, Heidelberg University, Heidelberg, Germany
64 Parklandklinik, Bad Wildungen, Germany
65 Institute for Medical Informatics, Biometry and Epidemiology, Universitätssklinikum Essen, University of Duisburg-Essen, Essen, Germany
66 Department of Internal Medicine VI, Psychosomatic Medicine and Psychotherapy, University Medical Hospital Tübingen, Tübingen, Germany
67 Department of Medical Genetics, University Medical Center Utrecht, Utrecht, The Netherlands
68 Center for Neurobehavioral Genetics, University of California, Los Angeles, Los Angeles, CA, USA
69 Brain Center Rudolf Magnus, Department of Psychiatry, University Medical Center Utrecht, The Netherlands
70 Department of Child and Adolescent Psychiatry, University Medical Center Utrecht, Utrecht, The Netherlands
71 Clinical Genetics Unit, Department of Woman and Child Health, University of Padova, Padova, Italy
72 M. Sklodowska-Curie Cancer Center and Institute of Oncology, Warsaw, Poland
73 Department of Epidemiology, Institute of Occupational Medicine, Department of Epidemiology, Lodz, Poland
74 Department of Clinical Nutrition, Institute of Public Health and Clinical Nutrition, University of Eastern Finland, Kuopio, Finland
75 Netherlands Consortium for Healthy Ageing, Leiden University Medical Center, The Netherlands
76 Department of Nutrition and Dietetics, Harokopio University, Athens, Greece
77 1st Department of Psychiatry, Athens University Medical School, Athens, Greece
78 Eating Disorders Unit, 1st Department of Psychiatry, Athens University Medical School, Athens, Greece
79 Adolescent Health Unit (AHU), 2nd Department of Pediatrics – Medical School, University of Athens ‘P & A Kyriakou’ Children’s Hospital, Athens, Greece
80 Department of Psychiatry, 1st Faculty of Medicine, Charles University, Prague, Czech Republic
81 Department of Pediatrics, 1st Faculty of Medicine, Charles University, Prague, Czech Republic
82 Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden
83 Institute of Human Genetics, Department of Genomics, Life & Brain Center, University of Bonn, Bonn, Germany
84 Institute of Neuroscience and Medicine (INM-1), Research Center Jülich, Jülich, Germany
85 Division of Medical Genetics, Department of Biomedicine, University of Basel, Basel, Switzerland
86 Martin-Luther-Universität Halle-Wittenberg, Klinikum der Medizinischen Fakultät, Halle/Saale, Germany
87 Institute of Clinical Molecular Biology, University of Kiel, Kiel, Germany
88 Institute of Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany
89 Institute of Medical Informatics, Biometry and Epidemiology, Ludwig-Maximilians-University, Munich, Germany
90 CNRS 8090-Institute of Biology, Pasteur Institute, Lille, France
91 McGill University and Genome Quebec Innovation Centre, Montreal, QC, Canada
92 Division of Nephrology, Department of Internal Medicine and Medical Specialties, Columbus-Gemelly Hospitals, Catholic University, Rome, Italy
93 Unitat de Recerca de Reumatologia (URR), Institut de Recerca Hospital Universitari Vall d’Hebron, Barcelona, Spain
94 Genetic Epidemiology Group, International Agency for Research on Cancer (IARC), Lyon, France
95 Virginia Institute for Psychiatric and Behavioral Genetics, Department of Psychiatry, Virginia Commonwealth University, Virginia, VA, USA
96 The Finnish Institute of Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland
97 The Program for Human and Population Genetics, The Broad Institute of MIT and Harvard, Cambridge, MA, USA
98 Finnish Institute of Occupational Health, Province of Southern Finland, Helsinki, Finland
99 NORMENT, KG Jebsen Centre for Psychosis Research, Division of Mental Health and Addiction, Oslo University Hospital & Institute of Clinical Medicine, University of Oslo, Oslo, Norway
100 Department of Psychology, University of Oslo, Oslo, Norway
101 Department of Biological and Medical Psychology, University of Bergen, Bergen, Norway
102 Kavli Research Centre for Aging and Dementia, Haraldsplass Deaconess Hospital, Bergen, Norway
103 K.G. Jebsen Centre for Research on Neuropsychiatric Disorders, University of Bergen, Bergen, Norway
104 KG Jebsen Centre for Psychosis Research, Norwegian Centre For Mental Disorders Research (NORMENT), Department of Clinical Science, University of Bergen, Bergen, Norway
105 Dr Einar Martens Research Group for Biological Psychiatry, Center for Medical Genetics and Molecular Medicine, Haukeland University Hospital, Bergen, Norway
106 Institute of Hygiene and Epidemiology, 1st Faculty of Medicine, Charles University, Prague, Czech Republic
107 Department of Cancer Epidemiology and Genetics, Masaryk Memorial Cancer Institute, Brno, Czech Republic
108 Palacky University, Olomouc, Czech Republic
109 University Health Network and Mount Sinai Hospital, Toronto General Hospital, and Samuel Lunenfeld Research Institute, Toronto, ON, Canada
110 Departments of Psychiatry, and Genetics and Genomic Sciences, Seaver Autism Center, and the Mindich Child Health and Development Institute, Mount Sinai School of Medicine, New York, NY, USA
111 The Centre for Applied Genomics and Program in Genetics and Genome Biology, The Hospital for Sick Children, Toronto, ON, Canada
112 Department of Psychiatry and Psychotherapy, Medical University Vienna, Vienna, Austria
113 The Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden
114 Rheumatology Unit, Department of Medicine at the Karolinska University Hospital, Solna, Sweden
115 European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK
Markers for distinguishing European populations

LM Hawkins et al

European Journal of Human Genetics

116 Inserm, U1061, Université Montpellier 1, Montpellier, France
117 Department of Emergency Psychiatry, CHU Montpellier, Montpellier, France
118 Eli Lilly and Company Ltd, Erl Wood Manor, Windlesham, Surrey, UK
119 Department of Nutrition, The University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

WTCCC3

Data Analysis Group: Carl A Anderson1, Jeffrey C Barrett1, James AB Floyd1, Christopher S Franklin1, Ralph McGinnis1, Nicole Soranzo1, Eletheria Zeggini1
UK Blood Services Controls: Jennifer Sambrook2, Jonathan Stephens2, Willem H Ouwehand2
1958 Birth Cohort Controls: Wendy L McArdle3, Susan M Ring3, David P Strachan4
Management Committee: Graeme Alexander5, Cynthia M Bulik6, David A Collier7, Peter J Conlon8, Anna Dominiczak9, Audrey Duncanson10, Adrian Hill11, Cordelia Langford11, Graham Lord12, Alexander P Maxwell11, Linda Morgan14, Leena Peltonen1, Richard N Sandford11, Neil Sheeran12, Nicole Soranzo1, Fredrik O Vannberg1, Jeffrey C Barrett1 (chair).

DNA, Genotyping, and Informatics Group: Hannah Blackburn1, Wei-Min Chen11, Sarah Edkins1, Mathew Gillman1, Sarah E Hunt1, Cordelia Langford1, Graham Lord12, Floyd1, Christopher S Franklin1, Ralph McGinnis1, Nicole Soranzo1, Not

15. Academic Department of Medical Genetics, Cambridge University, Cambridge CB2 0QQ, UK
16. Center for Public Health Genomics, University of Virginia, Charlottesville, VA, USA.


This work is licensed under a Creative Commons Attribution 3.0 License. To view a copy of this license, visit http://creativecommons.org/licenses/by/3.0/