Title

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

Permalink

https://escholarship.org/uc/item/75k9h2jm

Journal

European Journal of Human Genetics, 22(10)

ISSN

1018-4813

Authors

Huckins, Laura M

Boraska, Vesna

Franklin, Christopher S

et al.

Publication Date

2014-10-01

DOI

10.1038/ejhg.2014.1

Copyright Information

Peer reviewed

EJHG *Open*

ARTICLE

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

Laura M Huckins[*,1], Vesna Boraska[1,2], Christopher S Franklin[1], James AB Floyd[1], Lorraine Southam[1], GCAN[5], WTCCC3[6], Patrick F Sullivan[3], Cynthia M Bulik[3], David A Collier[4], Chris Tyler-Smith[1], Eleftheria Zeggini[1,7] and Ioanna Tachmazidou[1,7]

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 list 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.

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[2] approach uses genomic 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*[3] 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.[4–6] 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*[9] 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.

[1]The Wellcome Trust Sanger Institute (WTSI), Hinxton, UK; [2]University of Split School of Medicine, Split, Croatia; [3]University of North Carolina, Chapel Hill, NC, USA; [4]King's College, London, UK
[5]GCAN members are listed before the references.
[6]WTCCC3 members are listed before the references.
[7]These 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
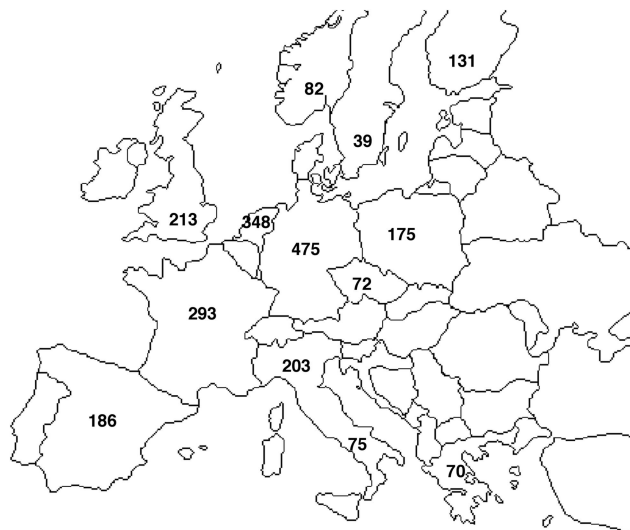

Figure 1 Geographical distribution of samples across Europe.

Table 1 Sample sizes per population

Population	Abbreviation	Sample size
Canada	CA	54
Czech Republic	CZ	72
Finland	FI	131
France	FR	293
Germany	DE	475
Greece	GR	70
North Italy	NIT	203
Netherlands	NL	348
Norway	NO	82
Poland	PL	175
South Italy	SIT	75
Spain	ES	186
Sweden	SE	39
UK	UK	213
USA	USA	491

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;[4] 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 package[16] 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 R[19] with 1000 permutations (as recommended).

FST
Tian et al[20] 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.

Informativeness was calculated according to Rosenberg et al,[8] using the formula below:

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

where p_j is the mean frequency of allele j over all populations, p_j 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 Raaum et al.[23]

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 was 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.[21,22] 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 12.9% 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 two top 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-southeast (NNW/SSE, $-11°$, $\rho = 0.91$). This is remarkably similar to the $-16°$ angle cited by Novembre et al.[4] We see no significant correlation between PC3 and PC4 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 inkilometres 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,[8] for 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 (eg, 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 $\rho = 0.97$, although it may be noted that the correlation is slightly lower ($\rho = 0.907$) 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 Raaum et al[23] to select PCAIMs; the top 25 are shown in Supplementary Table 3. We noted
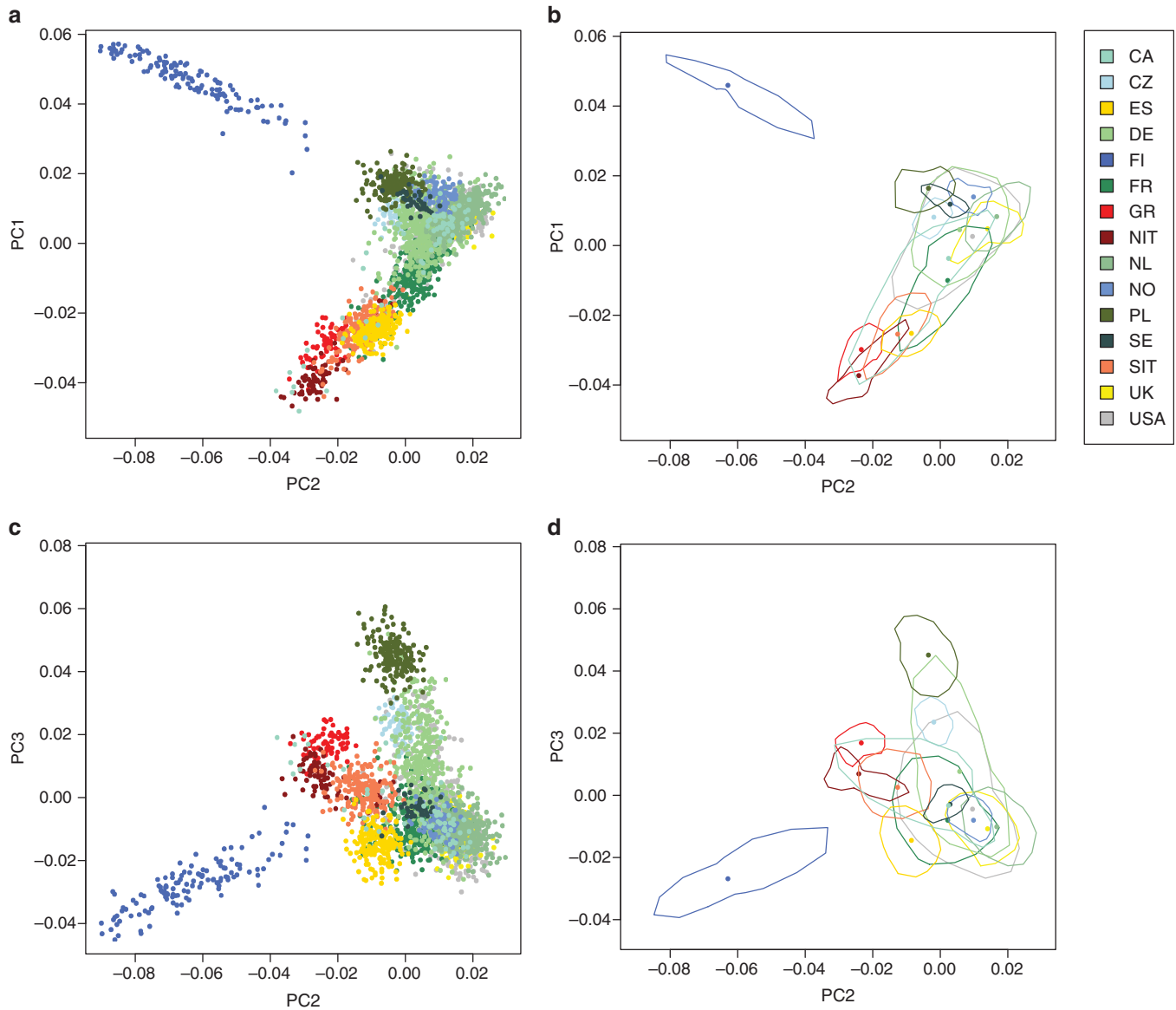
Figure 2 Fine structure between the 15 European populations studied. (**a**) Fine structure across all populations: PC1 *versus* PC2. (**b**) The distribution of samples is shown for each population. Outlying samples (deviating in location by more than 3 SDs from the mean) were excluded. A three-point moving average filter was used to smooth outlines. (**c**) Fine structure across all populations: PC2 *versus* PC3. (**d**) The distribution of samples is shown for each population, calculated as in (**b**). CA, Canada; CZ, Czech Republic; DE, Germany; ES, Spain; FI, Finland; FR, France; GR, Greece; NIT, North Italy; NL, Netherlands; NO, Norway; PL, Poland; SE, Sweden; SIT, South Italy; UK, United Kingdom; USA, United States of America

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 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 ($\rho = 0.98$). 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.

Figure 4c shows the proportion of samples correctly assigned for Dutch populations as a function of the number of markers used. Note that samples are predominantly assigned to neighbouring populations when using PCAIMs, especially Germany and France. A large proportion are assigned correctly, to the Netherlands. When using AIMs, the majority of samples are assigned to Germany, while only a small number are assigned to the Netherlands, and a similar number are left unassigned.

Finally, we considered the assignment of Swedish samples (Figure 4d). This population was not included at all when originally calculating AIMs and PCAIMs; thus, these samples provided an opportunity to see how well a 'new' population could be assigned using the derived AIMs and PCAIMs. Using PCAIMs, all samples

Table 2 Significance of principal components[a]

Principal component	% Variance explained	Tracy–Widom statistic	P-value
1	0.14	1333.1	$<1E-300$
2	0.09	603.3	$<1E-300$
3	0.07	294.9	$<1E-300$
4	0.06	121.2	$<1E-300$
5	0.05	100.9	$1.40E-295$
6	0.05	43.7	$9.79E-86$
7	0.05	10.9	$3.20E-12$
8	0.05	10.0	$5.08E-11$
9	0.05	10.2	$3.24E-11$
10	0.05	6.9	$5.30E-07$

The Tracy–Widom statistic is calculated using the smartpca software package.[15]
[a]Proportion of variance explained by the top 10 principal components.

Table 3 Pair-wise Fst calculated between all populations

	CZ	DE	ES	FI	FR	GR	NIT	NL	NO	PL	SIT	UK
CZ												
DE	0											
ES	0.003	0.002										
FI	0.006	0.007	0.011									
FR	0.001	0.001	0.001	0.008								
GR	0.004	0.004	0.003	0.013	0.003							
NIT	0.005	0.004	0.002	0.014	0.003	0.001						
NL	0.001	0.001	0.003	0.007	0.001	0.005	0.006					
NO	0.002	0.001	0.004	0.006	0.002	0.007	0.007	0.001				
PL	0	0.001	0.005	0.006	0.003	0.006	0.007	0.002	0.003			
SIT	0.003	0.002	0.001	0.011	0.001	0.001	0.001	0.003	0.004	0.004		
UK	0.001	0	0.002	0.007	0	0.005	0.005	0	0.001	0.002	0.002	
USA	0.001	0	0.002	0.007	0	0.004	0.004	0	0.001	0.002	0.002	0

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.
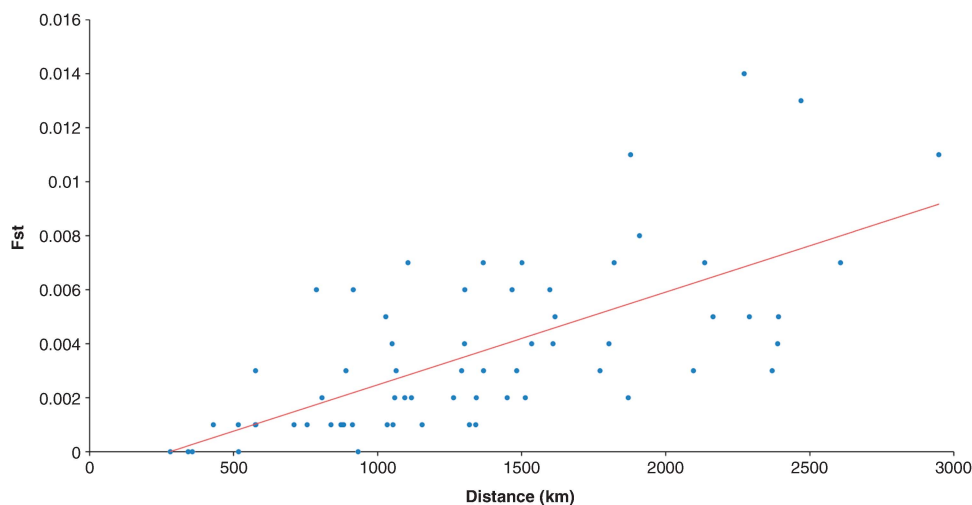
Figure 3 Genetic distance correlates with geographical distance. We computed pair-wise Fst between all populations, and compared this to the geographic distance in kilometres between the midpoints of each population. $R^2 = 0.465$.

Table 4 Correlation between AIMs when calculated using all 13 populations, and when leaving one population out

Missing population	ρ
CZ	0.9773
DE	0.9810
ES	0.9635
FI	0.9070
FR	0.9795
GR	0.9605
NIT	0.9564
NL	0.9760
NO	0.9713
PL	0.9682
SIT	0.9745
UK	0.9774
USA	0.9771

We calculated AIMs for 13 sets of 12 populations, and computed the Spearman's rank correlation coefficient (ρ) in each instance.

were assigned to geographically close populations, including Germany, the Netherlands and France. We were not able to assign all the samples using AIMs. Further, one sample was assigned to South Italy. All other samples were assigned to geographically close populations using AIMs.

Substructure within populations

We investigated within-population substructure by performing PCA on each population individually. K-means clustering was then used to assign samples to separate subclusters (Supplementary Figure 2). We found evidence of subclustering in the USA and Canadian populations, and a small number of outliers in the Spanish population. USA samples cluster into three broad groups.

We further investigated substructure within the USA samples by testing nearest-neighbour assignments for the USA samples, using all markers (see Supplementary Material for methods). Each USA sample may then be assigned a 'nearest' European population, as shown in Supplementary Figure 3. We observe that the majority (74%) of the USA samples cluster with North-Central Europeans, while a further 25% cluster with the Mediterranean populations. We find only a very small number of samples (1%) assigned to Finland, as would be expected.

Substructure within Canadian samples is likely to be due to the large French-Canadian population component. We found that Canadian samples were divided into two groups: a tight cluster and a number of outliers. We plotted Canadian samples alongside French samples, and found that the tight Canadian cluster overlapped the French cluster; outlying samples, on the other hand, did not intersect the French cluster at all (Supplementary Figure 2A).

DISCUSSION

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 the GC approach, or by adjusting for PCs; however, these methods are only applicable on a genome-wide scale. An alternative approach to this problem is to correct for stratification using AIMs.

We investigated evidence of population stratification across 15 populations of European origin using genome-wide methods such as PCA and Fst. This represents one of the largest studies of this kind, and includes some populations that have not previously been used to

assign 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.[14] 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 Raaum'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;
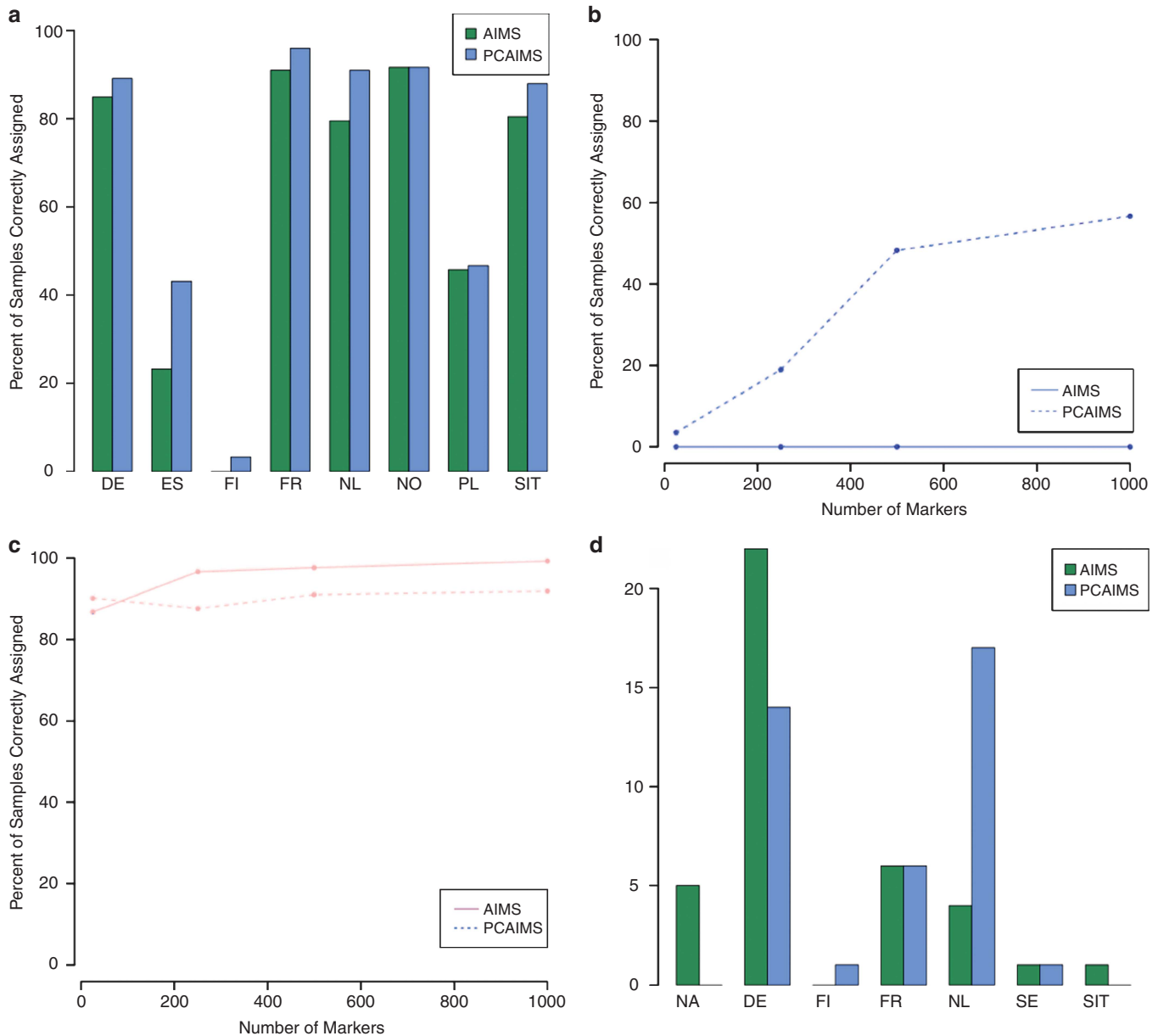
Figure 4 AIMs and PCAIMs are able to predict sample ancestry with high accuracy for most populations, even at small numbers of markers. (**a**) Percent of samples correctly assigned using 25 markers, across all populations. AIMs are shown in green, PCAIMs in blue. (**b**) Assignment of Finnish samples, 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.

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

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

V Boraska[1,2]
CS Franklin[1]
JAB Floyd[1,3]
LM Thornton[4]
LM Huckins[1]
L Southam, BSc[1]
N William Rayner[1,5,6]
I Tachmazidou[1]
KL Klump[7]
J Treasure[8]
CM Lewis[9]
U Schmidt[8]
F Tozzi[4]
K Kiezebrink[10]
J Hebebrand[11]
P Gorwood[12,13]
RAH Adan[14,15]
MJH Kas[14]
AFavaro[16]
P Santonastaso[16]
F Fernández-Aranda[17,18]
M Gratacos[19,20,21,22]
F Rybakowski[23]
M Dmitrzak-Weglarz[24]
J Kaprio[25,26,27]
A Keski-Rahkonen, MPH[25]
A Raevuori[25,28]
EF Van Furth[29,30]
MCT Slof-Op t Landt[29,31]

JI Hudson[32]
T Reichborn-Kjennerud[33,34]
GPS Knudsen[33]
P Monteleone[35,36]
AS Kaplan[37,38]
A Karwautz[39]
H Hakonarson[40,41]
WH Berrettini[42]
Y Guo[40]
D Li[40]
NJ Schork[43]
G Komaki[44,45]
T Ando[44]
H Inoko[46]
T Esko[47]
K Fischer[47]
K Männik[48,49]
A Metspalu[47,48]
JH Baker[4]
RD Cone[50]
J Dackor[51]
JE DeSocio[52]
CE Hilliard[4]
JK O'Toole[53]
J Pantel[54]
JP Szatkiewicz[51]
C Taico[4]
S Zerwas[4]
SE Trace[4]

OSP Davis[9,55]
S Helder[9]
K Bühren[56]
R Burghardt[57]
M de Zwaan[58,59]
K Egberts[60]
S Ehrlich[61,62]
B Herpertz-Dahlmann[56]
W Herzog[63]
H Imgart[64]
A Scherag[65]
S Scherag[11]
S Zipfel[66]
C Boni[12]
N Ramoz[12]
A Versini[12]
MK Brandys[14,15]
UN Danner[15]
C de Kovel[67]
J Hendriks[14]
BPC Koeleman[67]
RA Ophoff[68,69]
E Strengman[67]
AA van Elburg[15,70]
A Bruson[71]
M Clementi[71]
D Degortes[16]
M Forzan[71]
E Tenconi[16]
E Docampo[19,20,21,22]
G Escaramís[19,20,21,22]
S Jiménez-Murcia[17,18]
J Lissowska[72]
A Rajewski[73]
N Szeszenia-Dabrowska[73]
A Slopien[24]
J Hauser[24]
L Karhunen[74]
I Meulenbelt[31]
PE Slagboom[31,75]
A Tortorella[35]
M Maj[35]
G Dedoussis[76]
D Dikeos[77]
F Gonidakis[78]
K Tziouvas[76]
A Tsitsika[79]
H Papezova[80]
L Slachtova[81]
D Martaskova[80]
JL Kennedy[37,38]
RD Levitan[37,38]
Z Yilmaz[4,37]
J Huemer[39]
D Koubek[39]
E Merl[39]
G Wagner[39]
P Lichtenstein[82]
G Breen[9]

S Cohen-Woods[9]
A Farmer[9]
P McGuffin[9]
S Cichon[83,84,85]
I Giegling[86]
S Herms[83,85]
D Rujescu[86]
S Schreiber[87]
H-E Wichmann[88,89]
C Dina[90]
R Sladek[91]
G Gambaro[92]
N Soranzo[1]
A Julia[93]
S Marsal[93]
Ra Rabionet[19,20,21,22]
V Gaborieau[94]
DM Dick[95]
A Palotie[1,96,97]
S Ripatti[96,98]
E Widén[96,98]
OA Andreassen[99]
T Espeseth[99,100]
A Lundervold[101,102,103]
I Reinvang[100]
VM Steen[104,105]
S Le Hellard[104,105]
M Mattingsdal[99]
I Ntalla[76]
V Bencko[106]
L Foretova[107]
V Janout[108]
M Navratilova[107]
S Gallinger[109]
D Pinto[110]
SW Scherer[111]
H Aschauer[112]
L Carlberg[112]
A Schosser[112]
L Alfredsson[113]
B Ding[113]
L Klareskog[114]
L Padyukov[114]
C Finan[1]
G Kalsi[9]
M Roberts[9]
DW Logan[1]
L Peltonen[1]
GRS Ritchie[1,115]
P Courtet[116,117]
S Guillame[116,117]
I Jaussent[116,117]
JC Barrett[1]
X Estivill[19,20,21,22]
A Hinney[11]
PF Sullivan, FRANZCP[4,51]
DA Collier[9,118]
E Zeggini[1]
CM Bulik[4,119]

GCAN
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, NCNP, 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, Universitätsklinikum 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

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 Anderson[1], Jeffrey C Barrett[1], James AB Floyd[1], Christopher S Franklin[1], Ralph McGinnis[1], Nicole Soranzo[1], Eleftheria Zeggini[1].

UK Blood Services Controls: Jennifer Sambrook[2], Jonathan Stephens[2], Willem H Ouwehand[2].

1958 Birth Cohort Controls: Wendy L McArdle[3], Susan M Ring[3], David P Strachan[4].

Management Committee: Graeme Alexander[5], Cynthia M Bulik[6], David A Collier[7], Peter J Conlon[8], Anna Dominiczak[9], Audrey Duncanson[10], Adrian Hill[11], Cordelia Langford[1], Graham Lord[12], Alexander P Maxwell[13], Linda Morgan[14], Leena Peltonen[1], Richard N Sandford[15], Neil Sheerin[12], Nicole Soranzo[1], Fredrik O Vannberg[11], Jeffrey C Barrett[1] (chair).

DNA, Genotyping, and Informatics Group: Hannah Blackburn[1], Wei-Min Chen[16], Sarah Edkins[1], Mathew Gillman[1], Emma Gray[1], Sarah E Hunt[1], Cordelia Langford[1], Suna Onengut-Gumuscu[16], Simon Potter[1], Stephen S Rich[16], Douglas Simpkin[1], Pamela Whittaker[1].

1. The Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, UK

2. Division of Transfusion Medicine, Department of Haematology, University of Cambridge, NHSBT Cambridge Centre, Long Road, Cambridge CB2 0PT, UK

3. Department of Social Medicine, University of Bristol, Bristol BS8 2BN, UK

4. St George's University, Division of Community Health Sciences, London SW19 0RE, UK

5. Department of Hepatology, Cambridge University Hospitals NHS Foundation Trust, Cambridge CB2 0QQ, UK

6. Department of Psychiatry, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

7. Institute of Psychiatry, King's College London, London SE5 8AF, UK

8. Department of Nephrology, Beaumont Hospital, Dublin, Ireland and Royal College of Surgeons Dublin, Dublin, Ireland

9. BHF Glasgow Cardiovascular Research Centre, University of Glasgow, Glasgow G12 8TA, UK

10. Gibbs Building, 215 Euston Road, London NW1 2BE, UK

11. Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX1 2JA, UK

12. MRC Centre for Transplantation, King's College London, London SE1 9RT, UK

13. Belfast City Hospital, Lisburn Road, Belfast BT9 7AB, UK

14. School of Molecular Medical Sciences, University of Nottingham, Nottingham NG7 2UH, UK

15. Academic Department of Medical Genetics, Cambridge University, Cambridge CB2 0QQ, UK

16. Center for Public Health Genomics, University of Virginia, Charlottesville, VA, USA.

1 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.
2 Devlin B, Roeder K: Genomic control for association studies. *Biometrics* 1999; **55**: 997–1004.
3 Menozzi P, Piazza A, Cavalli-Sforza L: Synthetic maps of human gene frequencies in Europeans. *Science* 1978; **201**: 786–792.
4 Novembre J, Johnson T, Bryc K et al: Genes mirror geography within Europe. *Nature* 2008; **456**: 98–103.
5 Lao O, van Duijn K, Kersbergen P, de Knijff P, Kayser M: Proportioning whole-genome single-nucleotide polymorphism diversity for the identification of geographic population structure and genetic ancestry. *Am J Hum Genet* 2006; **78**: 680–690.
6 Nelis M, Esko T, Mägi R et al: Genetic structure of europeans: a view from the North–East. *PLoS One* 2009; **4**: e5472.
7 Shriver MD, Parra EJ, Dios S et al: Skin pigmentation, biogeographical ancestry and admixture mapping. *Hum Genet* 2003; **112**: 387–399.
8 Rosenberg NA, Li LM, Ward R, Pritchard JK: Informativeness of genetic markers for inference of ancestry. *Am J Hum Genet* 2003; **73**: 1402–1422.
9 Kidd JR, Friedlaender FR, Speed WC, Pakstis AJ, De La Vega FM, Kidd KK: Analyses of a set of 128 ancestry informative single-nucleotide polymorphisms in a global set of 119 population samples. *Invest Genet* 2011; **2**: 1–13.
10 Kim H, Hysi PG, Pawlikowska L et al: Population stratification in a case–control study of brain arteriovenous malformation in Latinos. *Neuroepidemiology* 2008; **31**: 224–228.
11 Nassir R, Kosoy R, Tian C et al: An ancestry informative marker set for determining continental origin: validation and extension using human genome diversity panels. *BMC Genet* 2009; **10**: 39.
12 Price AL, Patterson N, Hancks DC et al: Effects of cis and trans genetic ancestry on gene expression in African Americans. *PLoS Genet* 2008; **4**: e1000294.
13 Paschou P, Drineas P, Lewis J et al: Tracing sub-structure in the European American population with PCA-informative markers. *PLoS Genet* 2009; **4**: e1000114.
14 Drineas P, Lewis J, Paschou P 2010Inferring Geographic Coordinates of Origin for Europeans Using Small Panels of Ancestry Informative Markers. *PLoS One* 2009; **5**: e11892.
15 Patterson N, Price A, Reich D: Population structure and eigenanalysis. *PLoS Genet* 2006; **2**: e190.
16 Paradis E, Claude J, Strimmer K: APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* 2004; **20**: 289–290.
17 Peres-Neto PR, Jackson DA: How well do multivariate data sets match? The advantages of a Procrustean superimposition approach over the Mantel test. *Oecologia* 2001; **129**: 169–178.
18 Mardia KV, Kent JT, Bibby JM: *Multivariate Analysis*. New York, NY: Academic Press, 1979.
19 Oksanen J, Blanchet FG, Kindt R et al: vegan: Community Ecology Package, R package version 1.17-3, 2010. Available at: http://CRAN.R-project.org/package= vegan (last accessed August 2013).
20 Tian C, Plenge RM, Ransom M et al: Analysis and application of European genetic substructure using 300K SNP information. *PLoS Genet* 2008; **4**: e4.
21 Purcell S: PLINK (Version 1.07) [Software], 2013. Available at: http://pngu.mgh.harvard.edu/purcell/plink/ (last accessed August 2013).
22 Purcell S, Neale B, Todd-Brown K et al: PLINK: a toolset for whole-genome association and population-based linkage analysis. *Am J Hum Genet* 2007; **81**: 559.
23 Raaum RL, Wang AB, Al-Meeri AM, Mulligan CJ: Efficient population assignment and outlier detection in human populations using biallelic markers chosen by principal component-based rankings'. *BioTechniques* 2010; **48**: 449–454.

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