European Journal of Human Genetics (2008) 16, 1413–1429

www.nature.com/ejhg

ARTICLE

Investigation of the fine structure of European populations with applications to disease association studies

Simon C Heath*[,1], Ivo G Gut[1], Paul Brennan[2], James D McKay[2], Vladimir Bencko[3], Eleonora Fabianova[4], Lenka Foretova[5], Michel Georges[6], Vladimir Janout[7], Michael Kabesch[8], Hans E Krokan[9], Maiken B Elvestad[9], Jolanta Lissowska[10], Dana Mates[11], Peter Rudnai[12], Frank Skorpen[13], Stefan Schreiber[14], José M Soria[15], Ann-Christine Syvänen[16], Pierre Meneton[17], Serge Herçberg[18], Pilar Galan[18], Neonilia Szeszenia-Dabrowska[19], David Zaridze[20], Emmanuel Génin[21], Lon R Cardon[22] and Mark Lathrop[1,23]

[1]*Centre National de Genotypage, Institut Genomique, Commissariat à l'énergie Atomique, Evry, France;* [2]*International Agency for Research on Cancer (IARC), Lyon, France;* [3]*First Faculty of Medicine, Institute of Hygiene and Epidemiology, Charles University in Prague, Prague, Czech Republic;* [4]*Specialized Institute of Hygiene and Epidemiology, Banska Bystrica, Slovakia;* [5]*Department of Cancer Epidemiology and Genetics, Masaryk Memorial Cancer Institute, Brno, Czech Republic;* [6]*Unit of Animal Genomics, Faculty of Veterinary Medicine, GIGA-Research and Department of Animal Sciences, University of Liège, Liège, Belgium;* [7]*Department of Preventive Medicine, Palacky University, Olomouc, Czech Republic;* [8]*University Children's Hospital, Ludwig Maximilian's University Munich, Munich, Germany;* [9]*Faculty of Medicine, Department of Cancer Research and Molecular Medicine, Norwegian University of Science and Technology, Trondheim, Norway;* [10]*M. Sklodowska-Curie Memorial Cancer Center and Institute of Oncology, Warsaw, Poland;* [11]*Institute of Public Health, Bucharest, Romania;* [12]*National Institute of Environmental Health, Budapest, Hungary;* [13]*Faculty of Medicine, Department of Laboratory Medicine, Children's and Women's Health, Norwegian University of Science and Technology, Trondheim, Norway;* [14]*Institute for Clinical Molecular Biology, PopGen biobank, Christian-Albrechts-University, Kiel, Germany;* [15]*Unitat de Genómica de Malaties Complexes, Institut de Recerca de l'Hospital de la Santa Creu i Sant Pau, C/Sant Antoni M[a] Claret, 167, Barcelona, Spain;* [16]*Molecular Medicine, Department of Medical Sciences, Uppsala University, Uppsala, Sweden;* [17]*INSERM U872, Centre de Recherche des Cordeliers, Paris, France;* [18]*INSERM U557 et Unité de Recherche de Epidémiologie Nutritionnelle, 74 rue Marcel Cachin, Bobigny Cedex, France;* [19]*Department of Epidemiology, Institute of Occupational Medicine, Lodz, Poland;* [20]*Cancer Research Centre, Institute of Carcinogenesis, Moscow, Russia;* [21]*INSERM U794, Fondation Jean Dausset-CEPH, Paris, France;* [22]*GlaxoSmithKline, 709 Swedeland Road, King of Prussia, Pennsylvania, USA;* [23]*Fondation Jean Dausset-CEPH, Paris, France*

An investigation into fine-scale European population structure was carried out using high-density genetic variation on nearly 6000 individuals originating from across Europe. The individuals were collected as control samples and were genotyped with more than 300 000 SNPs in genome-wide association studies using the Illumina Infinium platform. A major East–West gradient from Russian (Moscow) samples to Spanish samples was identified as the first principal component (PC) of the genetic diversity. The second PC identified a North–South gradient from Norway and Sweden to Romania and Spain. Variation of frequencies at markers in three separate genomic regions, surrounding *LCT*, *HLA* and *HERC2*, were strongly associated with this gradient. The next 18 PCs also accounted for a significant proportion of genetic

*Correspondence: Dr SC Heath, Centre National de Génotypage, 2, Rue Gaston Crémieux, 154 rue du Fbg. St Denis, Evry 91000, France.
Tel: +160878402; Fax: +160878485; E-mail: simon.heath@gmail.com
Received 8 August 2008; revised 8 October 2008; accepted 9 October 2008

diversity observed in the sample. We present a method to predict the ethnic origin of samples by comparing the sample genotypes with those from a reference set of samples of known origin. These predictions can be performed using just summary information on the known samples, and individual genotype data are not required. We discuss issues raised by these data and analyses for association studies including the matching of case-only cohorts to appropriate pre-collected control samples for genome-wide association studies.

European Journal of Human Genetics (2008) **16,** 1413–1429; doi:10.1038/ejhg.2008.210

Keywords: PCA; GWAS; European; population; association; structure

Introduction

The genetic structure of populations is very important both from the population genetics viewpoint of understanding past and current relationships between populations, and from the genetic epidemiological viewpoint of avoiding spurious associations between diseases and genetic markers caused by differences in the population structure of cases and controls in association studies.[1–3] This study was undertaken to address both of these aspects, to describe the relationships between approximately 6000 European control samples from 13 different populations ranging geographically from Spain to Russia using data on 300k SNPs, and to provide guidelines for the use of these and other pre-collected European control samples in genome-wide association studies when only cases have been collected.

There are several recent studies that address similar questions about the relationships between European populations.[4–7] This study can be distinguished from these earlier studies in part by the type and quantity of the data used for the analysis. All samples studied were typed on the same set of >300 000 SNPs. Most of the population studies had >100 samples (the minimum having 76 samples), with the median number of samples per population being 374 and the maximum being >1300. All samples were from current European residents, which could be expected to give clearer results than using individuals of European descent who are more likely to show the evidence of recent admixture. In addition, the large number of different populations spread geographically over a large part of Europe allows more information to be obtained about genetic differences across the continent than if only a small number of very different populations were sampled.

The most common methods used by these studies on human population structure are clustering approaches[8,9] or principal component analysis (PCA).[10,11] The clustering approaches work on the basis of the presence of distinct genetic groups, and the probability of group membership of samples or, at a finer level, of chromosome blocks, can be estimated. These approaches and, in particular, the Bayesian clustering methods[8] have been widely used in population genetic studies because of the detailed information they provide on group membership and individual admixture. However, these approaches tend to be computationally intensive, and are in practice not suited to the large numbers of markers present in genome-wide data sets. A second problem with the cluster-based methods is that they work best when the study population is a mixture of distinct populations that is, European, African and Asian, and are less well suited to the situation where sample populations are overlapping. New implementations are making the computational aspect less of a problem;[12] however, the difficulty with overlapping populations makes these approaches inappropriate for this study where the aim is to describe the relationships between a set of closely related populations within Europe. The PCA approach was therefore mainly used for this study, as it is possible to apply the technique to large data sets with many thousands of individuals and hundreds of thousands of markers, and with overlapping populations.

Materials and methods

The study was performed using control samples collected and genotyped for association or population studies; no genotyping was performed specifically for this study. For all studies, permission to use the samples was obtained from the original investigators. All samples were genotyped on the Illumina HumanHap 300 arrays or on its derivatives.

A total of 5847 individuals from across Europe, all genotyped on >300 000 SNPs, were used for the study. The samples came from 13 different countries, with sample origin being taken as the geographic location where the sample was collected, and consisted of eight sample sets. In addition to these samples, the 210 unrelated HapMap[13] population samples: 60 CEPH samples (parents), Utah residents with ancestry from northern and western Europe (CEU), 60 Yoruba samples (parents) from Ibadan, Nigeria (YRI), 45 Han-Chinese samples from Beijing, China (CHB) and 45 Japanese samples from Tokyo, Japan (JPT), were included in the analyses. The first sample set contained 2016 control individuals from six different eastern European populations collected for a GWA (Genome-Wide Association) study on lung cancer;[14] 620 from Poland, 560 from Russia, 374 from the Czech Republic, 209 from Hungary, 145 from Slovakia and 108 from Romania. The next set contained 1228 population samples from France.

The third set had 1385 samples from the UK 1958 Birth Cohort from the Wellcome Trust Case Control Consortium.[15,16] The fourth set had 506 German and 52 UK control samples from a GWA study on Asthma.[17] The fifth set had 234 Belgian control samples from the GWA on Crohn's disease.[18] The sixth set had 95 Swedish population samples from the Uppsala Family Study.[19] The seventh set had 108 Norwegian control samples. The final set had 147 additional German population samples[20] and 76 Spanish control samples.

All samples were passed through the standard QC procedures followed at the Centre National de Genotypage for GWA studies. Samples with genotyping success rates $<95\%$ were removed, as were male samples with $>0.5\%$ or female samples with $<20\%$ heterozygous markers on the X chromosome. A check for closely related individuals was carried out within each study population by calculating average IBS (identity by state) scores for all pairs of individuals. Each marker that was successfully typed for the two individuals was scored as 0, 1 or 2 depending on the number of alleles in common between the samples. The mean and standard deviation of this score for all autosomal markers were calculated for each pair, and a scatter plot produced of the mean against the standard deviation. Outlying points owing to related pairs were identified and the relevant individuals were excluded. Apart from the family-based studies (the Spanish cohort and the small UK cohort (52 samples)), only a small percentage of individuals (0–1%) had to be removed from each cohort because of close relatedness. In addition to identifying related pairs, the IBS analysis can also detect individuals who are 'less' related to the rest of the population than would be expected if the samples were homogenous. This is because of the individuals in question having either a different ethnic background or a problem in the quality of their genotypes. Such individuals were also excluded from further analyses. The sample numbers reported above are the final numbers used for the analyses after all QC steps were completed.

All samples used in the study were unrelated; in the case that the original data contained related individuals, an unrelated subset was selected using the IBS analysis to identify unrelated pairs. For the Spanish samples, which consisted of extended families, a graph was constructed with samples as nodes and edges joining unrelated samples (as estimated from the IBS analysis). A maximal unrelated set of individuals was then found as the maximal clique from this graph, a problem for which efficient approximation algorithms exist.[21]

Five marker panels were used for the statistical analyses. Panel 1 contained 129 673 autosomal SNPs selected from the Illumina HumanHap 300 panel to have a very high genotyping success rates ($\geq 98\%$) and high informativity (minor allele frequency (MAF) ≥ 0.05). In addition, SNPs in linkage disequilibrium (LD) ($r^2 \geq 0.1$) with other SNPs on the panel were removed. To investigate the effect of the marker allele frequency spectrum on the analyses, a panel of low-frequency SNPs (panel 2) and common SNPs (panel 3) were selected as having success rates $\geq 95\%$ and with the low-frequency SNP panel having $0 < \text{MAF} < 0.05$, and the common SNP panel having $\text{MAF} > 0.485$. The cutoff for the common SNP panel was selected to give a similar number of markers in both panels (8412 low-frequency SNPs and 8734 common SNPs). Panel 4 was constructed by selecting markers that were significantly correlated with the population membership to produce an estimate of a minimal marker panel to distinguish the different European populations; this panel contained 391 SNPs. The final panel, panel 5, contained 48 587 SNPs and was constructed from the intersection between the autosomal markers from the Affymetrix Mapping 500k and Illumina HumanHap 300 panels, selected to have success rates $\geq 95\%$ and $\text{MAF} \geq 0.01$.

Detecting population differences

The relationship between the different populations was initially investigated by calculating the Fst statistic for each pair of populations using the markers in panel 1. The population structure was then investigated in more detail using PCA on the individual samples. Following Patterson *et al*,[10] a scaled genotype matrix G was generated with rows indexed by individuals and columns by polymorphic (autosomal) SNPs; hence, G is of size $n \times m$ where m is the number of SNPs and n is the number of individuals. Each element $g_{i,j}$ contains the normalized genotype for individual i at marker j, and is calculated from the frequency of variant alleles (0, 0.5 or 1) $x_{i,j}$ for an individual genotype by subtracting the variant allele frequency p_j and dividing by the standard deviation. In the case of missing genotype data, the corresponding element was set to zero that is, to the population mean for the marker.

The matrix A of size $n \times n$ was then constructed as:

$$A = \frac{1}{n} GG'$$

and the eigenvalues and eigenvectors (principal components (PCs)) of A calculated. A has a maximum of $n-1$ non-zero eigenvalues, but we considered only the k largest eigenvalues and associated vectors. Both PCA and Fst statistics were calculated using the EIGENSTRAT[10,22] software package.

Identification of SNPs or genomic regions that were correlated with a given PCs used the SNP weights, which were calculated as follows: let L be a diagonal $k \times k$ matrix where diagonal element L_i is the ith largest eigenvalue, and V be the $n \times k$ matrix containing the k PCs associated with the eigenvalues in L. Let W be the $k \times m$ matrix of SNP weights for each component:

$$W = V'G$$

Rather than directly using the SNP weights in W, the correlations $r_{i,j}$ for PCs i and SNP j between the genotype

vectors and individual component weights were calculated as follows: let $\boldsymbol{v}_i$ be the ith column of V and, therefore, the vector of component weights for PCs i, and $\boldsymbol{g}_j$ the jth column of G and, therefore, the vector of normalized genotypes for SNP j. The correlation $r_{i,j}$ was then calculated as:

$$r_{i,j} = \frac{v'_i, .g_j, .}{\sqrt{(v'_i, .v_i, .)(g'_j, .g_j, .)}}$$

$$r_{i,j} = \frac{w_{i,j}}{\sqrt{(v'_i, .v_i, .)(g'_j, .g_j, .)}}$$

The significance of the correlations was assessed by calculating the test statistic $(n-1)r^2_{i,j}$, which has a χ^2_1 distribution under the null hypothesis of no correlation. Note that this is equivalent (up to a factor of $(n-1)/n$) to the commonly used score test for the association of a quantitative trait with the PCs as the outcome variable. The EIGENSTRAT package was used to calculate the SNP weights for each eigenvector.

It is interesting to consider the minimum number of markers required to reproduce the most important set of components (ie, those corresponding to the largest eigenvalues and, therefore, explaining most of the variance). For a given component, the correlations between the SNPs and the component can be used to select a small set of SNPs that can accurately predict the component weights.

A simple step-up strategy was used, adding markers to the model one at a time and selecting at each step the marker with the highest residual correlation that was not already in the model, until no markers with test statistics over the genome-wide significance level remained. To model multiple components (ie, the k largest), this procedure was carried out sequentially, starting with the largest component. This strategy was intended to obtain a small (although not necessarily minimal) set of markers that can recover the largest features detected by the PCA using marker panel 1. This method was used to select a set of 391 markers (panel 4) that were significantly correlated with the first two components from the PCA (using panel 1) of the European populations.

An investigation into possible differences in the LD patterns between populations was then carried out by estimating the extent of LD as measured by r^2 as a function of physical distance in the different populations. Using the markers in panel 1, estimates of r^2 for all pairs of markers closer than 10 Mb apart were obtained from the maximum likelihood (ML) estimates of the pairwise haplotype frequencies. For computational efficiency, only observations where both markers were typed were used, which allows a closed form solution for the ML estimates. Pairs of markers were grouped into bins of width 1 kb on the basis of the distance between them, and an average r^2 was calculated for each bin.

Predicting sample origin

Methods such as PCA or Bayesian clustering approaches,[8,9] which can be used to show the population substructure, can also be used to predict the genetic origin of unknown samples given their data. For example, if the PCA on a set of samples of known origin showed a separation between the samples from different origins, then it is possible to take samples of unknown origin and find the population with which they have the closest resemblance.

This can be carried out by performing a PCA on all samples, both of known and unknown origin. Using the known samples only, country-specific mean and variances for each PC showing a separation between the countries are calculated, and, using these, the relative probability of a new sample being in each of the possible candidate populations is calculated for each component used in the model assuming independent normal distributions for the weights from each component. These relative probabilities for each population are then multiplied across components to give the final probability distribution.

To show this method, the country of origin of 20% of the sample was ignored to form a test group and the remaining 80% of the sample was used to estimate the country-specific means and variances for the four most significant PCs. The probability of membership to each target population for each member of the test group was then calculated on the basis of this model.

The same analysis was also performed using the Bayesian clustering program STRUCTURE,[8] in which 80% of the model was marked as having a known origin and 20% was of unknown origin, and hence the origin had to be predicted. STRUCTURE is a Markov chain Monte Carlo (MCMC) sampling-based approach, and is computationally more intensive than the PCA approaches when used with large numbers of markers. It was not possible to run the analysis using the full set of markers, and instead both the PCA-based approach and STRUCTURE were run using the 'minimal' set of 391 markers, selected for their ability to predict the first two PCs in the European analysis. The STRUCTURE analysis was run for 100 000 iterations with 10 000 iterations of burn-in, and visual inspection of the likelihood at each sampling iteration indicated that convergence was reached after $\sim$20 000 iterations. The output from the STRUCTURE analysis is an estimated posterior probability distribution across the possible populations for each unknown sample, which can be compared with the probabilities obtained from the PCA approach.

One drawback of both approaches described above to predict sample origins is that if a sample comes from a country that is not represented in the original data set, it will still be classified as being a member of one of the original countries. To avoid this, some means of evaluating model fit must be used to identify samples that do not come from any of the proposed possibilities. A simple approach is to calculate a distance measure d_i for individual

i from a sample to the center of the population to which they have been assigned:

$$d_i = \sum_j \left(\frac{v_{i,j} - \mu_j}{\sigma_j} \right)$$

where $v_{i,j}$ is the component weight for individual i at component j, and μ_j and σ_j are the mean and standard deviation of component j for the population assigned to individual i.

Although this classification procedure is very useful, there is a drawback in that it is necessary to perform the PCA using both the known and unknown samples, which is computationally intensive, as it will require that the PCA be re-run each time new samples need to be processed. In addition, this requires access to the individual genotype data for the known samples, which may not always be available. However, given the SNP weight matrix W, the eigenvalues and the allele frequencies from the original population used for the PCA, we can calculate a normalized genotype matrix H for the new samples, using the allele frequencies from the original population to perform the normalization. Given W, L (the diagonal matrix with the eigenvalues on the diagonal) and the original genotype matrix G, the eigenvector matrix V for the original samples can be calculated as follows.

From the properties of eigen decompositions, we can write

$$V'A = LV'$$

Given that $A = (1/n)\ GG'$, we can substitute for A to give:

$$(1/n)V'GG' = LV'$$

The matrix of SNP weights, W, was defined as $W = V'G$, so we can re-write this as:

$$(1/n)WG' = LV'$$

The transpose of V can then be obtained from:

$$V' = (1/n)L^{-1}WG'$$

This is not interesting in itself, as we already needed V to calculate the SNP weights W. However, if we replace G by H, the normalized genotype matrix for the new samples, then it is possible to calculate Q, a vector of 'pseudo-eigenvectors' for the new samples:

$$Q = (1/n)L^{-1}WH'$$

The assumption here is that if we had performed the joint PCA with the new and old samples together, the significant components (and the SNP weights associated with them) would be close to those calculated in the original PCA with only the old samples.

To test this, the above procedure was carried out for the 210 HapMap samples from all four populations, estimating the eigenvectors for the HapMap samples based on the SNP weights and eigenvalues calculated from the European control samples. All European control samples were then used as the training set for the population classification of the HapMap samples.

Results

Table 1 shows the Fst statistic calculated using marker panel 1 for all population pairs using the 5847 European samples along with the 210 HapMap samples. Not unexpectedly, the differences between the African, Asian and European populations are much greater than the differences seen within Europe. However, substructure within Europe is clearly indicated by Table 1. Note that although the values of Fst between the European populations are small (from 0.006 between Spanish and Russian

Table 1 Fst statistics calculated between each pair of countries: Spain (Sp), France (Fr), Belgium (Be), Sweden (Sw), Norway (No), Germany (Ge), Romania (Ro), Czech (Cz), Slovakia (Sl), Hungary (Hu), Poland (Po), Russia (Ru), and the four HapMap cohorts CEU, CHB, JPT and YRI

	Sp	*Fr*	*Be*	*UK*	*Sw*	*No*	*Ge*	*Ro*	*Cz*	*Sl*	*Hu*	*Po*	*Ru*	*CEU*	*CHB*	*JPT*
Fr	0.0008															
Be	0.0015	0.0002														
UK	0.0024	0.0006	0.0005													
Sw	0.0047	0.0023	0.0018	0.0013												
No	0.0047	0.0024	0.0019	0.0014	0.0010											
Ge	0.0025	0.0008	0.0005	0.0006	0.0011	0.0016										
Ro	0.0023	0.0017	0.0018	0.0028	0.0041	0.0044	0.0016									
Cz	0.0033	0.0016	0.0013	0.0014	0.0016	0.0024	0.0003	0.0016								
Sl	0.0034	0.0017	0.0015	0.0017	0.0019	0.0026	0.0005	0.0014	0.0001							
Hu	0.0030	0.0015	0.0013	0.0016	0.0020	0.0026	0.0004	0.0011	0.0001	0.0001						
Po	0.0053	0.0032	0.0028	0.0027	0.0023	0.0034	0.0012	0.0028	0.0004	0.0004	0.0006					
Ru	0.0059	0.0037	0.0034	0.0032	0.0025	0.0036	0.0016	0.0030	0.0008	0.0007	0.0009	0.0003				
CEU	0.0026	0.0008	0.0005	0.0002	0.0011	0.0012	0.0006	0.0028	0.0014	0.0016	0.0016	0.0026	0.0031			
CHB	0.1096	0.1094	0.1093	0.1096	0.1073	0.1081	0.1085	0.1047	0.1080	0.1069	0.1058	0.1086	0.1036	0.1095		
JPT	0.1118	0.1116	0.1114	0.1117	0.1095	0.1103	0.1107	0.1068	0.1102	0.1091	0.1079	0.1108	0.1057	0.1117	0.0069	
YRI	0.1460	0.1493	0.1496	0.1513	0.1524	0.1531	0.1502	0.1463	0.1503	0.1498	0.1490	0.1520	0.1504	0.1510	0.1901	0.1918

samples down to 0.00008 between Czech and Slovak samples), the standard errors of the estimates are such that these are all significantly different from zero with $P<0.05$.

A PCA was performed on the same set of samples and on the same marker panel. The first 110 PCs were significant at the 5% level as evaluated using the Tracy–Widom test on the eigenvalues.[10] However, the first few components stand out from the others, with the graph of percentage of variance contributed per component flattening out after the first four (Figure 1).

Scatter plots of the first two components (Figure 2a) show that these components explain the main differences between the Asian, European and African samples. Focusing just on the European samples (Figure 2b) shows some within-European structure, with Russian, UK and Spanish samples marking the 'east', 'north-west' and 'south-west' extremes of the European cluster. There is, however, no clear separation of the individual countries on this plot.

Repeating the PCA analysis omitting the African and Asian HapMap samples, it is possible to get a clearer view of the within Europe variation. With the European samples only, the number of significant PCs stayed at almost the same level (107 *vs* 110), with the first two PCs contributing 6.0 and 2.3%, respectively, of the variance because of the significant PCs (0.21 and 0.08% of the variance because of all components) (Figure 1). The scatter plot of the first two components (Figure 3) show a clear East–West and North–South gradient respectively, with a striking correlation between the geographic position of the countries and the position of the samples from each country on the PCA plot. Interestingly, the first two components of the European-only PCA are almost identical to components 3 and 4 from the PCA, including all four HapMap cohorts (the absolute value of the correlations between the relevant components from the two analyses are both 0.999, data not shown).

From the PCA with all samples, the first two components contribute $>32\%$ of the variance because of the 110 significant components. We can therefore estimate that roughly a third of the genetic variance in the samples is because of differences between the Asian, African and European samples, despite the much larger numbers of European samples when compared with Africa and Asia. It is important, however, not to regard the variation within Europe as insignificant as Figure 3 shows that there is a considerable European substructure.

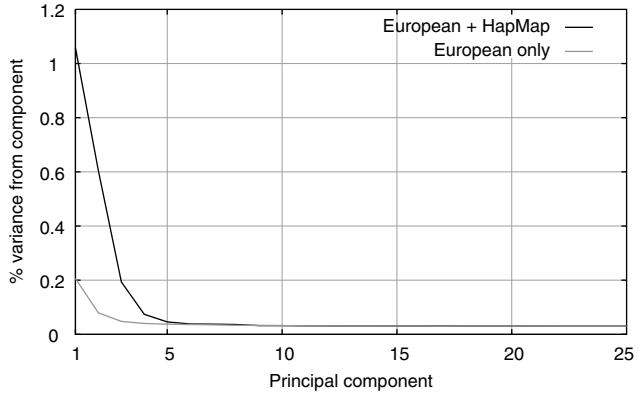

Figure 1 The percentage variance contributed by the first 25 PCs from the PCA of the European and HapMap populations combined and from the European populations alone.

The correlations between the SNPs and the largest components showed that large numbers of markers on all chromosomes were significantly correlated with the components. There were, however, some genomic regions that stood out. The SNP with the highest χ^2 for the first PC (showing an East–West gradient) was in the region of lactase (*LCT*), although there was a broad support for this component on all chromosomes (Figure 4a). The second PC (with the North–South gradient), however, had three genomic regions that stood out as being highly correlated to the PC (Figure 5b): *LCT*, *HLA* and *HERC2* (which is associated with iris color,[23–25]). Although the SNP test statistics indicate genomic regions that are more important than others in producing the gradients seen in Figure 3, it should be noted that there are $>24\,000$ SNPs scattered throughout the genome that achieve genome-wide significance level for correlation with the first two components ($\chi^2\geq 26$ for a genome-wide error rate of 0.05 after Bonferroni correction for multiple tests), with $>20\,000$ of these being for component 1. Indeed, the three genomic regions discussed above are not necessary to produce the pattern in Figure 3. To show this, the above PCA analysis was repeated after removing all markers from panel 1 within 10 Mb of each of *LCT*, *HLA* and *HERC2*, and the correlations between the first two PCs from the full marker panel analysis and the corresponding PCs from the reduced marker panel analysis were both ≥ 0.995 (Table 3). The scatter plot of the first two PCs from the reduced marker panel analysis was almost identical to Figure 3 (data not shown), as would be expected given the high correlations between the PC weights. This is not surprising as even after removing the markers within 10 Mb of the obvious peaks in Figure 4, the number of markers that achieve global significance only dropped by ~ 800.

A recent study using PCA on GWA data in European samples[5] observed a PC that appeared to be largely because of an inversion at around location 9 Mb on chromosome 8p. This region has been identified earlier[26,27] as containing an inversion, and has LD extending over a 4 Mb region around the inversion in the samples used for this study. The effects of this inversion were less apparent in this study because of the exclusion of markers in strong LD from the marker panel; before this filtering being applied the same observation was made of three clusters in a PC (fifth PC from an analysis with European samples only) strongly linked to chromosome 8p (data not shown). It should be noted that the division of samples into the three clusters was independent of population origin, indicating that the

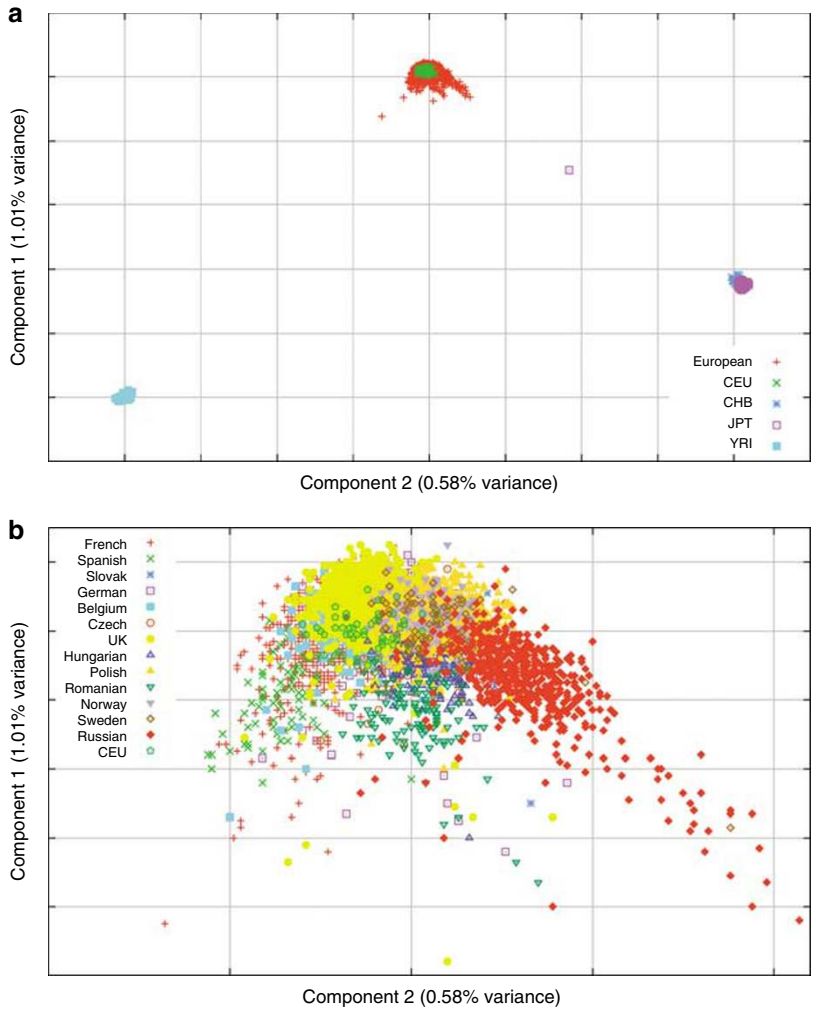

Figure 2 The scatter plot of the first two PCs from the PCA of the European and HapMap populations combined showing (**a**) all samples and (**b**) European samples only.

frequency of the inversion does not vary greatly between the European populations in this study.

A large case–control study[15] on UK samples identified markers in 12 autosomal regions showing large allele frequency differences between individuals from 12 different UK regions. Of these 12 genomic regions, seven contain SNPs that were significantly associated with PCs 1 or 2 in this study, and a further four regions show SNPs that were just under the threshold for global significance (Table 2). This study therefore detected 11 out of 12 of the genomic regions identified as correlating with geographic regions within the United Kingdom as involved in East–West and North–South gradients covering all of Europe. In addition, this study finds many more regions correlated with these gradients. It is likely that the larger geographical spread of the samples used in this study gives increased power to detect such regions compared with earlier studies.

The PCA on the European samples only was repeated using the low-frequency and common SNP panels (panels 2 and 3) to investigate the effect of the marker allele frequency spectrum on the detected population structure. The plots of the first two PCs from the PCA using these two panels can be seen in Figures 5a and b. It can be seen that although the separation of the countries is less clear than with the full panel, both the low-frequency SNP and common SNP panels give the same overall picture with the first two components corresponding to an East–West and North–South geographic axis. The correlations between the larger PCs from the different marker panels range from 0.8 to 0.95 (Table 3), and again indicate that, at least for the first two components, the smaller panels are capturing much of the information. This does not, however, extend to the smaller components; the PCA with the low-frequency and high-frequency panels had only six and two significant PCs, respectively, compared with 107 with the full marker panel.

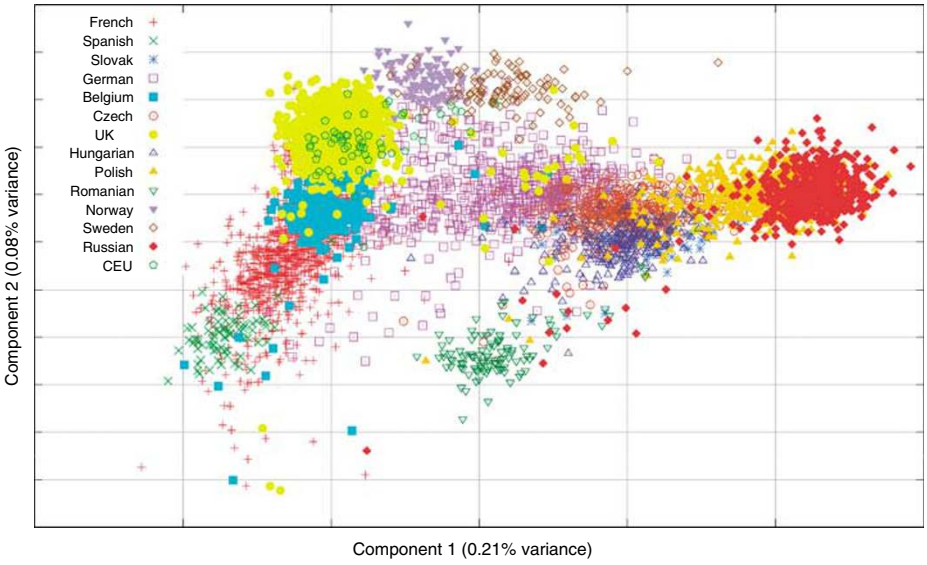

Figure 3 The scatter plot of the first two PCs from the PCA of the European populations only.

The PCA using the European samples was repeated using panel 4, with just 391 markers selected to predict just the first two PCs from the original European PCA. The scatter plot of the first two components (Figure 5c) shows that the general features of the two gradients in Figure 4 are recovered, but with a significant loss of resolution. The absolute value of the correlations between the first two components from the full panel and the 'minimal' panel were 0.95 and 0.80.

The final PCA was performed on the same set of European samples using marker panel 5, which had 48 587 SNPs that were in common between the Infinium Hapmap300k panel and the Affymetrix 500k panel. The correlations between the first PCs and second PCs from the two analyses were 0.99 and 0.98, and the scatter plots of the first two PCs were visually almost identical to Figure 3 (not shown), showing that a panel selected from the intersection between the two common genotyping platforms for Illumina and Affymetrix can detect most of the detail detected using all markers in panel 1.

Although the genetic variation within Europe is less than the variation between samples from different continents (Table 1), the variation within individual countries is still great enough to potentially give false-positive results in association studies if not taken into account. For example, there were 252 samples from Dresden and 222 samples from Munich among the samples analyzed that were plotted together as German samples. Figure 6 shows the plots of the first two PCs showing the distribution of the samples from these two cities. A case–control study drawing cases from Dresden and controls from Munich is at risk of false-positive results because of the differences between samples from the two cities. Considering the Dresden and Munich samples as cases and controls, respectively, the median single marker χ^2 statistic is 0.493 compared with an expected 0.456; hence the Genomic control lambda parameter[28,29] is 1.08 indicating an inflation that would need to be accounted for in an association study. This indicates that even if both cases and controls are collected from within a single country, it might be necessary to make corrections for population stratification. This effect would be much stronger if samples were collected from different countries, even if all samples were correctly classified as being European/Caucasian.

The extent of LD as a function of distance in the sample populations was investigated using the r^2 measure in different populations. It is known that there is a bias in the estimation of LD with low sample sizes;[30,31] small sample sizes lead to an overestimation of r^2 when LD is low. This bias can be seen in Figure 7a, which shows the decay curves estimated from subsets of 25, 50, 100, 500 and 1000 samples selected from the French cohort. The smaller sample sizes show a slower decay, which is because of an overestimation of r^2 when LD is low with small sample sizes. The curves for 500 and 1000 samples are very close, showing that this effect is small for larger samples. To avoid the sample size effect from influencing the population comparison, only the five largest populations (UK, France, Germany, Poland and Russia) were examined in detail for LD, with a random subset of 500 samples being taken from each population to equalize the sample sizes. The estimated decay curves show little difference between the five populations (Figure 7c and d), and the empirical 95% distributions of the LD estimates are large such that the differences are not significant (data not shown). More variance was seen between different chromosomes than

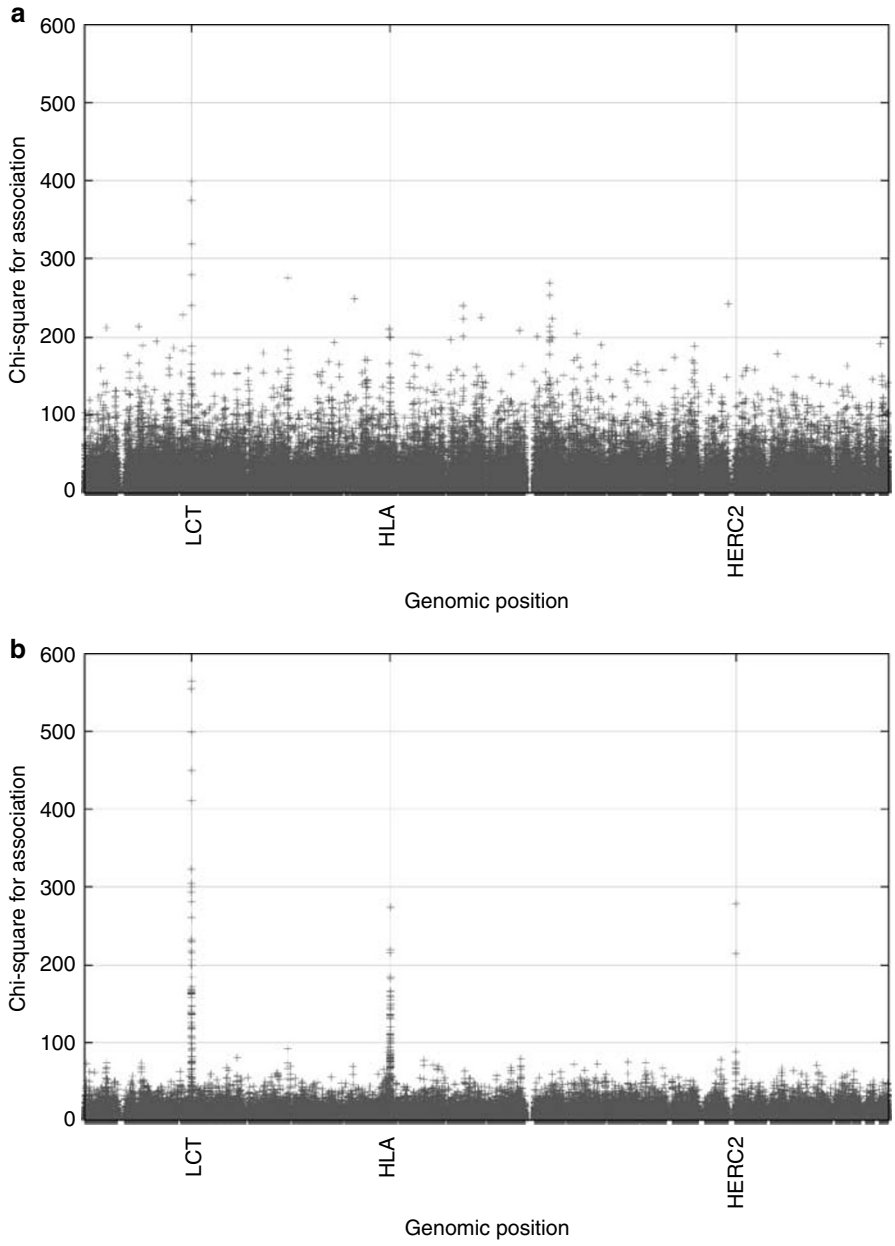

Figure 4 Graph showing the association along the genome with the first and second PCs (panels **a** and **b**, respectively).

between the different populations; for example, chromosome 22 shows a faster decay in LD with distance than most of the other chromosomes. This is shown in Figure 7b for the French population, but the same holds true for the other populations.

Predicting sample origin

The method to predict the origin of unknown samples using the PCA was tested with the European samples by using 80% of the samples selected at random to generate a model for the populations, and using this to predict the relative probabilities of population membership for the remaining 20% of the samples.

A summary of this analysis is presented in Table 4, which shows the relative probabilities of samples being in a given target population, averaged over all test samples originating from each population. In all cases (averaged over samples with the same origin), the target population with the highest posterior probability is that corresponding to the origin of the sample, with neighboring countries typically having the next highest probabilities. For example, for test samples from UK, the posterior probability of a

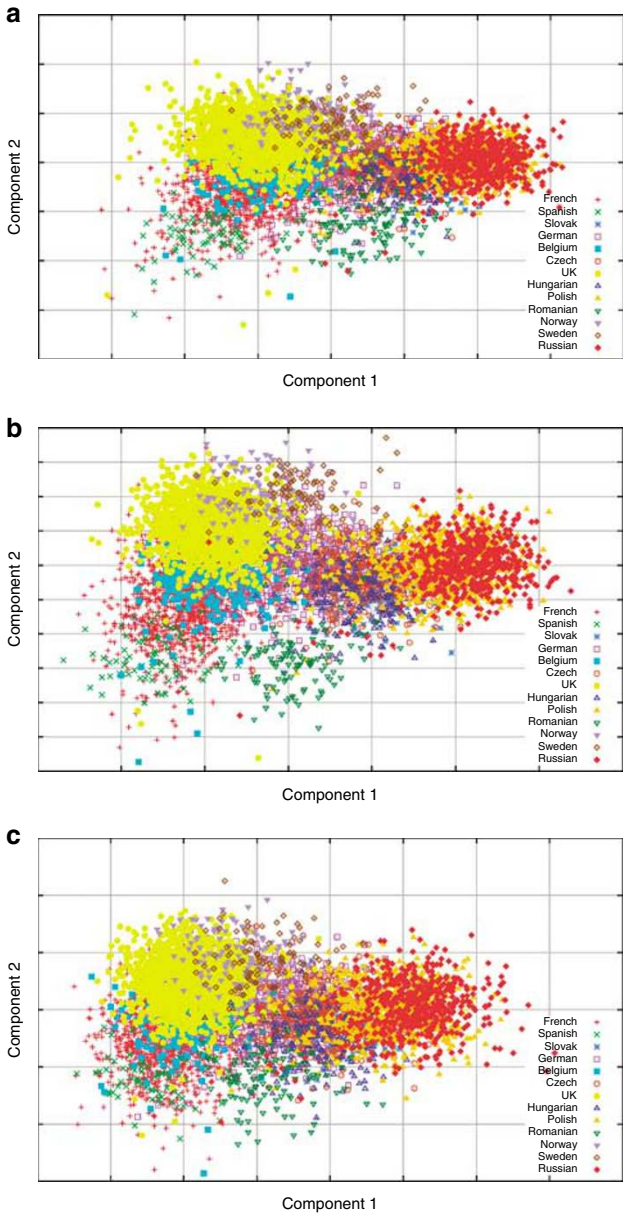

Figure 5 The scatter plots of the first two PCs from the PCA of the European populations using (**a**) the low frquency marker panel, (**b**) the common marker panel and (**c**) the minimal 391 marker panel.

UK origin was 0.95 with the next highest scoring target populations being Belgian, German and French; the probability for Slovak samples being Slovak was 0.45, whereas the probability of the Slovak samples being either Czech or Hungarian was 0.47.

Table 4 also shows the median and empirical 95% CI of the distance measure for test samples from each origin population. The distance measure is a measure of the distance in standard deviations from a sample to the center of the closest matching population. The median distance measure for all groups is low, as is to be expected since an appropriate target population is available for each test sample. This will not always be the case, and it can arise that a test sample may come from a population not represented in the training set. To test the behavior of the method in this case, a second analysis was performed taking each population in turn, removing the samples belonging to the test population from the training set, generating the model for the remaining populations, and then estimating the probabilities of population membership from the resulting model. The results for this analysis are shown in Table 5. The European samples all show the highest posterior probabilities for populations geographically close to the 'true' origin, with the median distance measure still being low (although higher than shown in Table 4). This indicates that the method works even when there is not an exact population match present in the training set.

Having to re-run the PCA with both old and new samples together has the disadvantages of being computationally intensive, especially if the training set is large, and requires that the individual genotypes of all individuals in the training set are available, which may not always be the case. Using the approach outlined here to estimate the component weights for new samples, the origins of the HapMap samples were estimated using the European samples as the training set. The results of this analysis are also given in Table 5. The HapMap CEU samples have, on average, a probability of 0.72 of being from the United Kingdom, with Germany and Belgium and, to a lesser extent, Norway and Sweden, accounting for the remaining cases. None of the HapMap CEU samples appeared to have a non-European origin. For the other HapMap populations, the classification procedure assigned 100% of the YRI samples to France, and almost 100% of the CHB and JPT samples to Russia. However, the distribution of the distance measure for the four populations was quite different. For the CEU samples, the median and 95% CI of the distance measure were 0.41 (0.11–1.01), whereas for the YRI, CHB and JPT populations, the median and 95% CIs were 19.3 (18.0–20.6), 17.7 (15.9–19.3) and 18.0 (15.4–19.6), respectively. Therefore, although attempting to classify the origin of a sample that comes from a population not contained in the original PCA can give a false classification, the distance measure makes it clear that the sample is far from the other known members of that population.

To compare the PCA-based method with existing Bayesian clustering methods, a similar analysis to that above was performed using the program STRUCTURE.[8] To avoid the computational difficulties of running STRUCTURE with marker panel 1, both the PCA-based method and STRUCTURE were run using the 391 marker panel (panel 4). The results of these analyses are presented in Table 6. It can be seen that while the PCA method with panel 4 did not perform as well as the full panel, over all test samples the true country of origin always had the

Table 2 SNPs with maximum correlations to PC 1 or PC 2 in genomic regions identified in Consortium WTCC[15] as containing markers with significant allele frequency differences between different geographic regions in the United Kingdom

Chromosome	*Genes*	*Region (Mb)*	*SNP*	*Position*	*PC*	P-*value*
2q21	LCT	135.16–136.82	rs1446585	136 123 949	2	8.1E−125
4p14	TLR1, TLR6 and TLR10	38.51–38.74	rs6531684	38 617 025	1	1.0E−41
4q28		137.97–138.01	rs2612131	137 977 142	2	8.2E−01
6p25	IRF4	0.32–0.42	rs1473602	373 722	1	9.6E−21
6p21	HLA	31.1–31.55	rs2844513	31 496 193	2	6.3E−49
9p24	DMRT1	0.86–0.88	rs7047524	864 129	1	2.0E−05
11p15	NAV2	19.55–19.7	rs10741780	19 556 547	1	6.5E−09
11q13	NADSYN1 and DHCR7	70.78–70.93	rs3794060	70 865 327	1	1.7E−08
12p13	DYRK4, AKAP3, NDUFA, RAD51AP1 and GALNT8	4.37–4.82	rs11063148	4 400 998	1	3.6E−18
14q12	HECTD1, AP4S1 and STRN3	30.41–31.03	rs7157080	30 623 471	1	1.8E−06
19q13	GIPR, SNRPD2, QPCTL, SIX5, DMPK and DMWD	50.84–51.09	rs4803866	51 026 795	1	3.8E−06
20q12		38.3–38.77	rs6029180	38 612 337	2	2.6E−06

Table 3 Pearson correlations between the first and second principal components calculated with the full marker panel and with the alternate panels

	Reduced SNP panel		*Rare SNP panel*		*Common SNP panel*		*Joint Illumina/Affymetrix SNP panel*	
	PC 1	*PC 2*	*PC 1*	*PC 2*	*PC 1*	*PC 2*	*PC 1*	*PC 2*
Full panel								
PC 1	1.000	−0.180	0.934	−0.062	0.963	−0.001	0.995	−0.108
PC 2	−0.161	0.995	−0.110	0.806	−0.159	0.878	−0.154	0.981

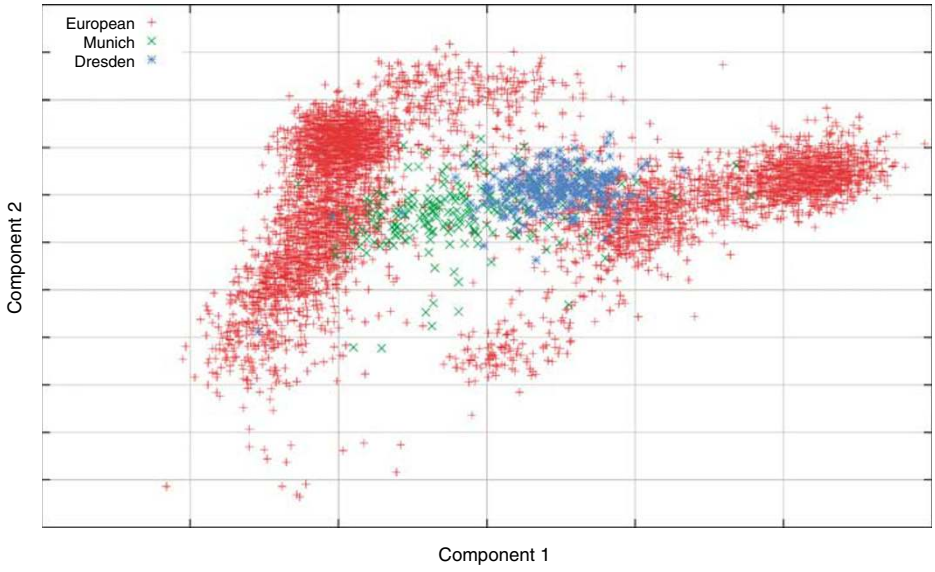

Figure 6 The scatter plot of the first two PCs of the European populations highlighting the samples from Dresden and Munich.

highest posterior probability, and the posterior was concentrated in the true country of origin and close neighbors. For example, the average probability that UK test samples were classified as being from the United Kingdom was 55% rather than 92% with the full panel. However, the smaller panel assigned a high probability to a larger region around the true country of origin. For example, the probability of the UK samples being French, Belgian, UK or Norwegian was >93%.

STRUCTURE also correctly identified the true region for the unknown samples (Table 6). However, it performed less well than the PCA approach in that the probability distribution for all populations was more diffuse, with a lower probability assigned to the true country of origin.

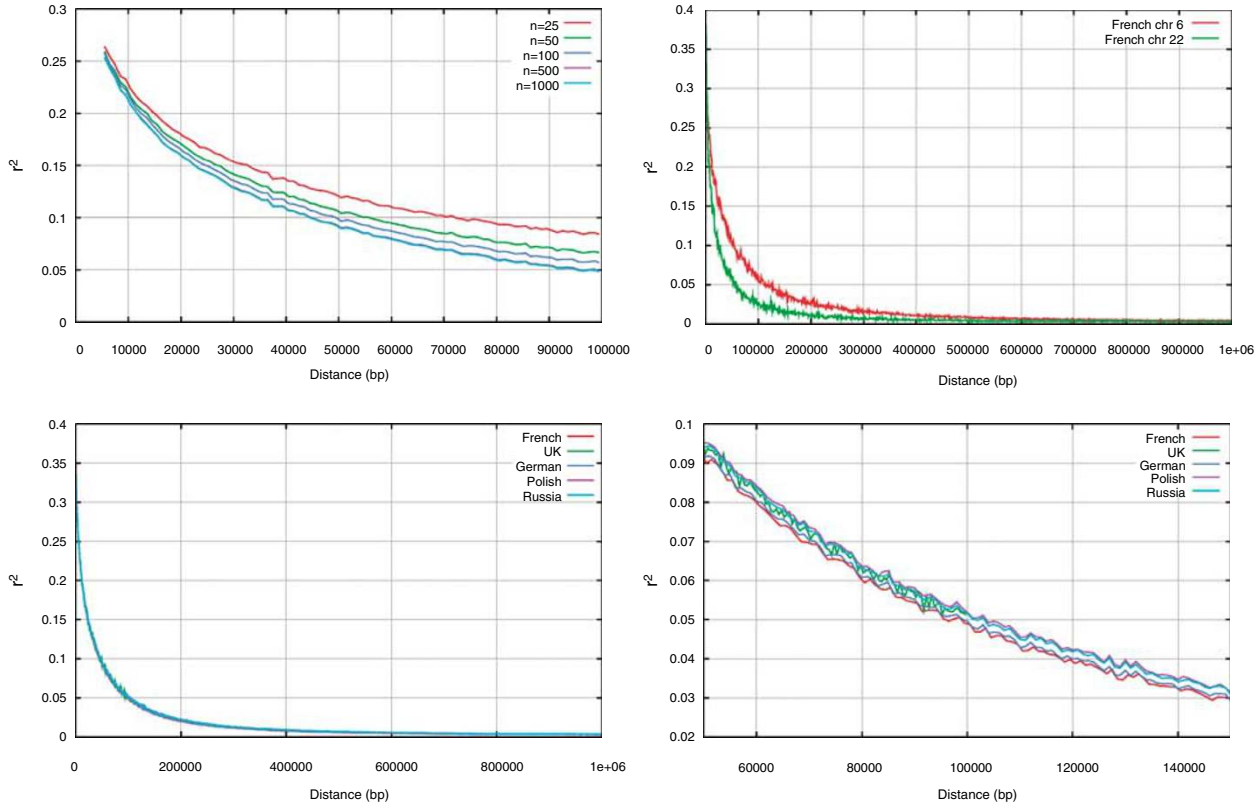

Figure 7 Graphs showing the decay in LD as a function of genomic distance for (**a**) differently sized subsets of the French cohort, (**b**) chromosomes 6 and 22 from the French cohort, (**c**) and (**d**) random subsets of size 500 from the five largest cohorts.

Taking the same example as for the PCA approach, the probability of UK samples being assigned to the United Kingdom was just 24%, and their probability of being either French, Belgian, UK or Norwegian was 61%.

For a final example, we return to the German samples that were discussed earlier. These samples were a part of an Asthma GWA study[17] and in addition to the control samples used in this study, there were also 676 German cases. The above procedure was used to generate the component weights for the case samples without running the PCA with the case samples included, allowing the case samples to be added to the plot in Figure 6. The sample classification was also performed; this predicted that 60% of the case samples were German with the others coming from other central or western countries, which is close to the figure for the German controls in Table 4. The plot of the first two components for the control samples and the case samples is shown in Figure 8, and confirms the classification results.

Discussion

For case–control studies to be effective, it is important in general that the cases and controls are matched as far as possible with respect to their genetic backgrounds. A striking feature of the samples used for this study is how well the geographic origin of the samples appears to correlate with the genetic origin, so that separating the samples by country of origin or on the basis of genetic measures gives similar results. The only major deviation from this pattern is with the Romanian samples that appear to be closer to the Spanish samples (further 'west') than their geographic position would indicate. This could be because of the historical close ties between Romania and Italy, but further studies would be required to confirm this. The relatively compact form of most of the individual country clusters in Figure 3 and the overall compact nature of the pan-European cluster in Figure 2 show the high quality of the geographic origin information in its ability to predict genetic relatedness. This is not always the case, and in situations where sample origin information is unreliable or missing, the ability to use the genetic information to match cases and controls would be very valuable.

The fine population structure within Europe, which can be detected using PCA techniques (Figure 3), is notable for the close correspondence with the geographical location of the sample origins. The two largest PCs from the European-only analysis are closely correlated with an East–West and a North–South geographic gradient, respectively. The

Table 4 Each horizontal line in the table shows the proportions of test samples originating from a given country that were assigned to each possible target country

Populations	*Distance*	*Spain*	*France*	*Belgium*	*United Kingdom*	*Norway*	*Sweden*	*Romania*	*Germany*	*Hungary*	*Slovakia*	*Czech*	*Poland*	*Russia*
Spain	1.13 (0.12–2.32)	0.945	0.055	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
France	0.85 (0.20–2.62)	0.085	0.515	0.270	0.105	0.000	0.000	0.004	0.014	0.007	0.000	0.000	0.000	0.000
Belgium	0.56 (0.13–2.20)	0.000	0.086	0.854	0.059	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
United Kingdom	0.67 (0.20–2.67)	0.000	0.009	0.027	0.947	0.000	0.000	0.000	0.017	0.000	0.000	0.000	0.000	0.000
Norway	0.87 (0.30–2.78)	0.000	0.000	0.000	0.000	0.991	0.010	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Sweden	0.73 (0.28–2.13)	0.000	0.000	0.000	0.000	0.099	0.901	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Romania	0.60 (0.22–1.90)	0.000	0.000	0.000	0.000	0.000	0.000	0.960	0.000	0.040	0.000	0.000	0.000	0.000
Germany	0.78 (0.22–3.27)	0.000	0.000	0.102	0.004	0.029	0.022	0.008	0.644	0.003	0.003	0.177	0.008	0.000
Hungary	0.68 (0.22–1.71)	0.000	0.000	0.000	0.000	0.000	0.000	0.022	0.051	0.546	0.292	0.090	0.000	0.000
Slovakia	0.78 (0.20–3.10)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.077	0.220	0.453	0.250	0.000	0.000
Czech	0.65 (0.16–2.28)	0.000	0.000	0.000	0.000	0.000	0.000	0.038	0.052	0.161	0.205	0.484	0.062	0.000
Poland	0.74 (0.14–2.26)	0.000	0.000	0.000	0.000	0.000	0.000	0.008	0.002	0.009	0.025	0.021	0.802	0.134
Russia	0.65 (0.13–3.01)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.008	0.008	0.000	0.040	0.944

Overall, 20% of the samples from each country were treated as test samples; these samples were not used to generate the population model.

Table 5 Each horizontal line in the table shows the proportion of samples originating from a given country that were assigned to each possible target country

Populations	*Distance*	*Spain*	*France*	*Belgium*	*United Kingdom*	*Norway*	*Sweden*	*Romania*	*Germany*	*Hungary*	*Slovakia*	*Czech*	*Poland*	*Russia*
Spain	1.33 (0.47–2.26)		0.987	0.013	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
France	1.28 (0.28–3.80)	0.188		0.658	0.134	0.001	0.000	0.002	0.011	0.004	0.001	0.001	0.001	0.000
Belgium	0.89 (0.25–2.62)	0.015	0.783		0.173	0.000	0.000	0.000	0.030	0.000	0.000	0.000	0.000	0.000
United Kingdom	1.39 (0.53–3.52)	0.000	0.233	0.730		0.013	0.001	0.000	0.020	0.001	0.001	0.002	0.000	0.000
Norway	1.25 (0.41–4.33)	0.000	0.000	0.000	0.158		0.823	0.000	0.019	0.000	0.000	0.000	0.000	0.000
Sweden	1.90 (0.49–6.17)	0.000	0.000	0.000	0.000	0.335		0.000	0.654	0.000	0.000	0.000	0.000	0.011
Romania	3.77 (1.52–7.65)	0.000	0.010	0.000	0.000	0.000	0.000		0.916	0.074	0.000	0.000	0.000	0.000
Germany	1.35 (0.41–3.90)	0.000	0.015	0.164	0.065	0.051	0.072	0.025		0.019	0.004	0.579	0.007	0.000
Hungary	0.84 (0.22–2.83)	0.000	0.000	0.000	0.000	0.000	0.000	0.025	0.074		0.688	0.208	0.005	0.000
Slovakia	0.80 (0.24–2.92)	0.000	0.000	0.000	0.000	0.000	0.000	0.014	0.028	0.362		0.511	0.084	0.000
Czech	0.83 (0.21–3.05)	0.000	0.000	0.000	0.000	0.000	0.000	0.011	0.091	0.229	0.598		0.072	0.000
Poland	1.12 (0.39–2.70)	0.000	0.000	0.000	0.000	0.000	0.000	0.005	0.001	0.009	0.084	0.048		0.854
Russia	1.94 (0.49–8.37)	0.000	0.002	0.000	0.000	0.000	0.000	0.011	0.004	0.008	0.007	0.000	0.968	
CEU	0.41 (0.11–1.01)	0.000	0.000	0.106	0.724	0.028	0.017	0.000	0.124	0.000	0.000	0.000	0.000	0.000
CHB	17.7 (15.9–19.3)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000
JPT	18.0 (15.4–19.6)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.010	0.012	0.000	0.000	0.977
YRI	19.3 (18.0–20.6)	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

The table shows the results when perfect matches were not available, ie, the samples from each country were analyzed using a model created without any samples from that country.

Table 6 Comparison of the PCA-based and STRUCTURE methods to assign origins to the unknown samples using panel 4 (391 markers)

Origin	*Spain*	*France*	*Belgium*	*United Kingdom*	*Norway*	*Sweden*	*Romania*	*Germany*	*Hungary*	*Slovakia*	*Czech*	*Poland*	*Russia*
PCA approach													
Spain	0.67	0.16	0.11	0.00	0.00	0.00	0.06	0.00	0.00	0.00	0.00	0.00	0.00
France	0.28	0.27	0.26	0.10	0.02	0.00	0.05	0.01	0.00	0.00	0.00	0.00	0.00
Belgium	0.06	0.14	0.47	0.22	0.04	0.02	0.02	0.04	0.00	0.00	0.00	0.00	0.00
United Kingdom	0.00	0.07	0.13	0.55	0.18	0.02	0.00	0.03	0.00	0.00	0.01	0.00	0.00
Norway	0.00	0.00	0.05	0.26	0.43	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Sweden	0.00	0.00	0.00	0.09	0.45	0.32	0.00	0.05	0.00	0.00	0.09	0.00	0.00
Romania	0.00	0.00	0.00	0.00	0.00	0.00	0.82	0.00	0.18	0.00	0.00	0.00	0.00
Germany	0.03	0.01	0.08	0.05	0.10	0.16	0.05	0.17	0.09	0.09	0.17	0.00	0.00
Hungary	0.00	0.02	0.00	0.00	0.00	0.04	0.12	0.08	0.21	0.37	0.09	0.06	0.00
Slovakia	0.00	0.00	0.00	0.00	0.00	0.06	0.03	0.09	0.19	0.34	0.12	0.14	0.03
Czech	0.00	0.00	0.00	0.00	0.03	0.05	0.00	0.13	0.14	0.32	0.19	0.09	0.05
Poland	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.02	0.13	0.09	0.32	0.42
Russia	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.02	0.07	0.03	0.26	0.61
Structure													
Spain	0.44	0.21	0.09	0.03	0.02	0.02	0.08	0.02	0.03	0.02	0.02	0.01	0.01
France	0.18	0.21	0.14	0.12	0.06	0.05	0.06	0.06	0.04	0.03	0.03	0.02	0.02
Belgium	0.11	0.17	0.20	0.12	0.06	0.06	0.05	0.09	0.04	0.03	0.04	0.02	0.02
United Kingdom	0.04	0.10	0.13	0.24	0.14	0.10	0.03	0.07	0.03	0.03	0.04	0.02	0.02
Norway	0.02	0.06	0.05	0.16	0.32	0.11	0.02	0.09	0.03	0.03	0.05	0.03	0.04
Sweden	0.01	0.04	0.06	0.14	0.20	0.20	0.02	0.08	0.04	0.05	0.05	0.07	0.05
Romania	0.09	0.08	0.05	0.03	0.02	0.02	0.40	0.04	0.06	0.06	0.08	0.04	0.03
Germany	0.05	0.08	0.07	0.09	0.08	0.09	0.06	0.11	0.08	0.07	0.09	0.07	0.06
Hungary	0.04	0.05	0.04	0.04	0.03	0.03	0.10	0.07	0.18	0.09	0.14	0.10	0.10
Slovakia	0.03	0.03	0.05	0.03	0.04	0.04	0.06	0.09	0.12	0.10	0.12	0.15	0.14
Czech	0.03	0.05	0.04	0.04	0.04	0.04	0.06	0.09	0.11	0.12	0.14	0.13	0.11
Poland	0.02	0.02	0.02	0.02	0.03	0.03	0.03	0.04	0.08	0.08	0.09	0.25	0.29
Russia	0.01	0.01	0.02	0.02	0.02	0.02	0.03	0.03	0.06	0.07	0.09	0.23	0.39

Each horizontal line in the table shows the proportion of test samples originating from a given country that were assigned to each possible target country. Overall, 20% of the samples from each country were treated as test samples; these samples were not used to generate the population model.

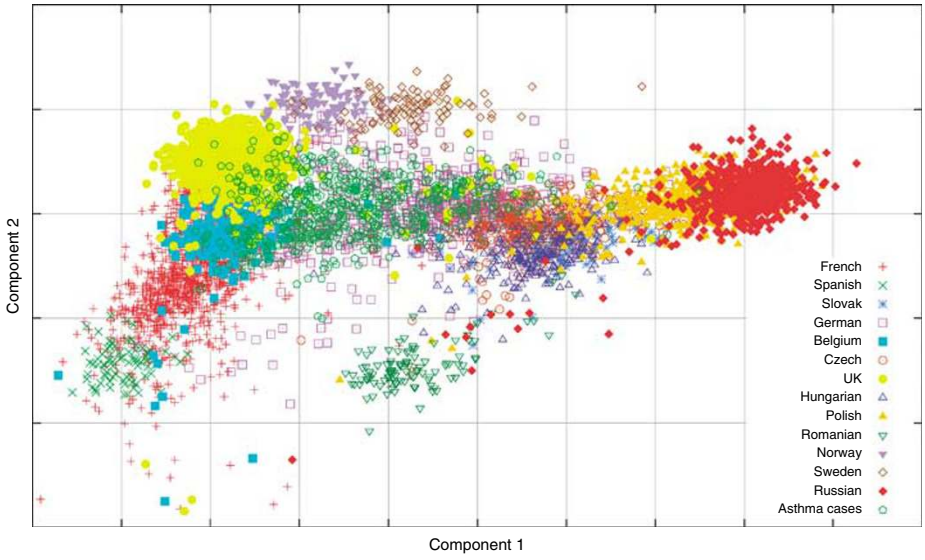

Figure 8 The scatter plot of the first two PCs of the European populations including the (German) Asthma cases using their estimated component weights.

presence of a North–South gradient with which *LCT* is associated has been reported in several European studies using whole genome data, and a strong East–West gradient has also been reported in earlier studies.[7,11] Although it is interesting that the regions most highly associated with the North–South gradient are *LCT*, *HLA* and *HERC2*, it

must be noted that these are just extreme examples of regions being associated with the gradient, and removing these regions had no discernible effect on the plot of the first two PCs.

It should be noted that explaining the observed gradients in terms of human population history is problematic. What is clear from the results is that there is a strong correlation between genetic and physical proximity, but there are several possible reasons why this should be the case.[32] Making inferences about population history from the data presented here is also difficult as there has been an ascertainment bias in the selection of SNPs, notably toward common SNPs, which can bias any conclusions drawn from the data.[33] However, the main conclusion of this study – that there is a strong correlation between genetic and physical proximity that can be used to 'map' unknown samples – is not dependent on the frequency spectrum of allele frequencies as shown by our tests with the low allele frequency and high allele frequency marker panels.

In order that the approaches described here for identifying sample origins or matching controls to cases be useful, it would be necessary that both the reference set of samples and the test set are genotyped for the set of markers used for the original PCA. We have shown that the two gradients are detected with much smaller number markers (391 instead of 129 673 used for the original analysis). However, reducing the number of SNPs does reduce the resolution to detect population structure, and many fewer significant components are found. In addition, the panel with 391 markers did not perform as well at predicting sample origin (Table 6); the reason for which is clear when the increased overlap between countries is seen in Figure 5.

The PCA-based approach, however, appeared to work better than the Bayesian clustering approach implemented in STRUCTURE when applied to the same panel of 391 markers (Table 6). In addition, the STRUCTURE analysis with panel 4 (391 markers) required 3 days of computing time (3 GHz Athlon processor) as opposed to several hours for the PCA approach, and this drops to minutes if the method for approximating the component weights for new samples is used. This makes the PCA approach more practical as a routine part of QC or statistical analysis workflows.

A point to note is that the panel made exclusively of markers with MAF $<= 0.05$ performed slightly worse than an equivalently sized panel of exclusively very common markers (MAF ≥ 0.485), measured by the correlation between the first two components of the reduced marker sets and those from the full marker set, but the PCA with the low-frequency marker panel detected more significant components (6 *vs* 2 for the common marker panel).

Although the question of the minimum number of markers required to detect the first n components is an interesting one,[6,34] we are more concerned with samples genotyped for GWA studies, and these are typically typed, at least initially, on one of the standard panels such as the Illumina HumanHap 300 or Affymetrix Mapping 500k marker sets. The PCA that we performed with a marker set formed from the intersection between the Illumina HumanHap 300 and Affymetrix Mapping 500k panels is therefore particularly interesting as it shows that a common panel is almost equally good at detecting the first few components as the full 129 673 SNP panel; therefore, at least for purposes of identifying sample origins, samples typed on either platform can be included in the same analyses.

The results of this study are encouraging for the use of shared control samples in European case–control studies. With or without the origin of the samples, it is possible to locate a set of case samples on a genetic origin map such as Figure 3 on the basis of their genotypes, and use this information to select a suitable set of control samples for the study. Any residual discrepancies between the cases and controls owing to imperfect matching can be corrected for by performing a PCA on the cases and selected control samples and correcting for the diversity detected through the PCs.[22]

The example given of the asthma cases illustrates how this approach could be used in practice. If the cases were of unknown origin, Figure 8 or the classification approach allows a quick identification of suitable control samples. Even if the sample origins were known, in this case it would have been possible to use additional controls from other populations such as the United Kingdom, French or Belgian to augment the power of the study and, potentially, reduce false positives by better matching of controls to the cases. For example, the original German samples for the Asthma study (673 controls and 676 cases) were used to replicate an association signal on chromosome 17.[17] The top SNP from the initial analysis on chromosome 17 was rs7216389, and a χ^2-test of association for this marker with the German Asthma data set gives an uncorrected P-value of 1.9E−7, which improves to 3.1E−9 if an adjustment is made for population structure by regressing on the significant PCs.[22] The analysis of sample origins indicates that Belgium and Czech samples are the closest match to the German samples. Performing the analysis using the Belgium and Czech samples as addition controls improves the corrected and uncorrected P-values to 1.3E−8 and 3.1E−10, respectively. Note that the correction for population structure reduces the GC lambda value from 1.1 to 1.0 in the original analysis and from 1.14 to 1.0 in the analysis with the Belgium and Czech samples. Although in this case sufficient control samples were available from the original study, it can be seen how using the methods described in this paper to identify suitable pre-genotyped controls could significantly increase the power for association studies.

In this study, we described how new genotyped samples can be located on a pre-existing plot of PCs without requiring to perform the PCA for all samples together, and

without needing the individual genotypes of the original data set. This approach has obvious practical advantages over performing a new PCA, but there are potential pitfalls. The main pitfall is applying this technique to a new set of samples that are sufficiently distinct from the training set such that the original PCA is no longer a close approximation to the joint PCA of the old and new samples together. The extreme example discussed in this study was to try and classify non-European HapMap samples using the European samples for the PCA and as a training set for the classification model. The Yoruban and Asian samples were identified as belonging to the countries on the south and east edges, respectively, of the European cluster, and the distance measure clearly indicates that they do not fit well into any of the proposed populations. It cannot be assumed that outliers will always be easily detectable in this way. This drawback could be avoided by adding as much diversity as possible into the initial training set. Using the European samples in addition to all four HapMap populations as a training set does not have an adverse effect in this data set on the ability to distinguish between different European populations, but does allow correct identification of the HapMap African and Asian samples (data not shown). A generally useful resource for localizing and matching samples should, therefore, contain a wide range of genetically different samples so that most new samples can be quickly and successfully mapped.

In conclusion, we have shown that using PCA techniques it is possible to detect fine-level genetic variation in European samples. The genetic and geographic distances between samples are highly correlated, resulting in a striking concordance between the scatter plot of the first two components from a PCA of European samples and a geographic map of sample origins. We have shown how this information can be used to predict the origin of unknown samples in a rapid, precise and robust manner, and that this prediction can be performed without requiring access to the individual genotype data on the original samples of known origin.

The marker panels used in this study, and the summary information on the control samples required to perform the classification of new samples can be obtained on application to the corresponding author.

Acknowledgements

Funding for the genotyping for the eastern European data was provided by INCa, France and the CNG, France. Funding for the genotyping of the German Asthma case and control samples was provided by the GABRIEL European project. The CNG also provided support for genotyping all samples (including those described above) apart from the Wellcome Trust control samples, which were generated by the Wellcome Trust Case-Control Consortium. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. Funding for the project was provided by the Wellcome Trust under award 076113. Johanna Sandling is acknowledged for managing the Uppsala Family Study samples. The longitudinal database of Uppsala Family Study is supported by the Swedish Research Council. The Popgen biobank is supported by the German Ministry of Education and Research (BMBF) through the National Genome Research Network (NGFN). It also received infrastructure support through the DFG excellence cluster 'Inflammation at Interfaces.'

References

1 Clayton D, Walker N, Smyth D *et al*: Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nat Genet* 2005; **37**: 1243–1246.
2 Freedman M, Reich D, Penney K *et al*: Assessing the impact of population stratification on genetic association studies. *Nat Genet* 2004; **36**: 388–393.
3 Marchini J, Cardon L, Phillips M, Donnelly P: The effects of human population structure on large genetic association studies. *Nat Genet* 2004; **36**: 512–517.
4 Seldin MF, Shigeta R, Villoslada P *et al*: European population substructure: clustering of northern and southern populations. *PLoS Genet* 2006; **2**: e143.
5 Tian C, Plenge RM, Ransom M *et al*: Analysis and application of European genetic substructure using 300 K SNP information. *PLoS Genet* 2008; **4**: e4.
6 Price AL, Butler J, Patterson N *et al*: Discerning the ancestry of European Americans in genetic association studies. *PLoS Genet* 2008; **4**: e236.
7 Bauchet M, McEvoy B, Pearson LN *et al*: Measuring European population stratification with microarray genotype data. *Am J Hum Genet* 2007; **80**: 948–956.
8 Pritchard JK, Stephens M, Donnelly P: Inference of population structure using multilocus genotype data. *Genetics* 2000; **155**: 945–959.
9 Tang H, Coram M, Wang P, Zhu X, Risch N: Reconstructing genetic ancestry blocks in admixed individuals. *Am J Hum Genet* 2006; **79**: 1–12.
10 Patterson N, Price AL, Reich D: Population structure and eigenanalysis. *PLoS Genet* 2006; **2**: e190.
11 Menozzi P, Piazza A, Cavalli-Sforza L: Synthetic maps of human gene frequencies in Europeans. *Science* 1978; **201**: 786–792.
12 Li JZ, Absher DM, Tang H *et al*: Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 2008; **319**: 1100–1104.
13 Consortium IH , Frazer KA, Ballinger DG *et al*: A second generation human haplotype map of over 3.1 million SNPs. *Nature* 2007; **449**: 851–861.
14 Hung R, Mckay J, Gaborieau V *et al*: A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. *Nature* 2008; **452**: 633–637.
15 Consortium WTCC: Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 2007; **447**: 661–678.
16 Consortium WTCC, (TASC) A-A-ASCBurton PR, Clayton DG, Cardon LR *et al*: Association scan of 14,500 nonsynonymous SNPs in four diseases identifies autoimmunity variants. *Nat Genet* 2007; **39**: 1329–1337.
17 Moffatt MF, Kabesch M, Liang L *et al*: Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. *Nature* 2007; **448**: 470–473.
18 Libioulle C, Louis E, Hansoul S *et al*: Novel Crohn disease locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of PTGER4. *PLoS Genet* 2007; **3**: e58.
19 Leon DA, Koupil I, Mann V *et al*: Fetal, developmental, and parental influences on childhood systolic blood pressure in 600 sib pairs: the Uppsala Family study. *Circulation* 2005; **112**: 3478–3485.

20 Krawczak M, Nikolaus S, von Eberstein H, Croucher PJ, El Mokhtari NE, Schreiber S: PopGen: population-based recruitment of patients and controls for the analysis of complex genotype-phenotype relationships. *Community genet* 2006; **9**: 55–61.
21 Bron C, Kerbosch J: Finding all cliques of an undirected graph. *Commun ACM* 1973; **16**: 575–577.
22 Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D: Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 2006; **38**: 904–909.
23 Sturm RA, Duffy DL, Zhao ZZ *et al*: A single SNP in an evolutionary conserved region within intron 86 of the HERC2 gene determines human blue-brown eye color. *Am J Hum Genet* 2008; **82**: 424–431.
24 Kayser M, Liu F, Janssens AC *et al*: Three genome-wide association studies and a linkage analysis identify HERC2 as a human iris color gene. *Am J Hum Genet* 2008; **82**: 411–423.
25 Eiberg H, Troelsen J, Nielsen M *et al*: Blue eye color in humans may be caused by a perfectly associated founder mutation in a regulatory element located within the HERC2 gene inhibiting OCA2 expression. *Hum Genet* 2008; **123**: 177–187.
26 Herva R, de la Chapelle A: A large pericentric inversion of human chromosome 8. *Am J Hum Genet* 1976; **28**: 208–212.
27 Giglio S, Broman KW, Matsumoto N *et al*: Olfactory receptor-gene clusters, genomic-inversion polymorphisms, and common chromosome rearrangements. *Am J Hum Genet* 2001; **68**: 874–883.
28 Devlin B, Roeder K: Genomic control for association studies. *Biometrics* 1999; **55**: 997–1004.
29 Devlin B, Roeder K, Wasserman L: Genomic control, a new approach to genetic-based association studies. *Theor popul biol* 2001; **60**: 155–166.
30 Terwilliger JD, Haghighi F, Hiekkalinna TS, Göring HH: A bias-ed assessment of the use of SNPs in human complex traits. *Curr Opin Genet Dev* 2002; **12**: 726–734.
31 Teare MD, Dunning AM, Durocher F, Rennart G, Easton DF: Sampling distribution of summary linkage disequilibrium measures. *Ann Hum Genet* 2002; **66**: 223–233.
32 Novembre J, Stephens M: Interpreting principal component analyses of spatial population genetic variation. *Nat Genet* 2008; **40**: 646–649.
33 Clark AG, Hubisz MJ, Bustamante CD, Williamson SH, Nielsen R: Ascertainment bias in studies of human genome-wide polymorphism. *Genome Res* 2005; **15**: 1496–1502.
34 Paschou P, Ziv E, Burchard EG *et al*: PCA-correlated SNPs for structure identification in worldwide human populations. *PLoS Genet* 2007; **3**: 1672–1686.