Current Biology 18, 1241–1248, August 26, 2008 ©2008 Elsevier Ltd All rights reserved DOI 10.1016/j.cub.2008.07.049

Report

Correlation between Genetic and Geographic Structure in Europe

Oscar Lao,[1,22] Timothy T. Lu,[2,22] Michael Nothnagel,[2] Olaf Junge,[2] Sandra Freitag-Wolf,[2] Amke Caliebe,[2] Miroslava Balascakova,[3] Jaume Bertranpetit,[4] Laurence A. Bindoff,[5] David Comas,[4] Gunilla Holmlund,[6] Anastasia Kouvatsi,[7] Milan Macek,[3] Isabelle Mollet,[8] Walther Parson,[9] Jukka Palo,[10] Rafal Ploski,[11] Antti Sajantila,[10] Adriano Tagliabracci,[12] Ulrik Gether,[13] Thomas Werge,[14] Fernando Rivadeneira,[15,16] Albert Hofman,[16] André G. Uitterlinden,[15,16] Christian Gieger,[17,18] Heinz-Erich Wichmann,[17,18] Andreas Rüther,[19] Stefan Schreiber,[19] Christian Becker,[20] Peter Nürnberg,[20] Matthew R. Nelson,[21] Michael Krawczak,[2,23] and Manfred Kayser[1,23,*]

[1]Department of Forensic Molecular Biology
Erasmus University Medical Center Rotterdam
3000 CA Rotterdam
The Netherlands

[2]Institut für Medizinische Informatik und Statistik
Christian-Albrechts University Kiel
D-24105 Kiel
Germany

[3]Institute of Biology and Medical Genetics
University Hospital Motol and 2nd School of Medicine
Charles University Prague
CZ 150 06, Prague 5
Czech Republic

[4]Unitat de Biologia Evolutiva
Pompeu Fabra University
08003 Barcelona, Catalonia
Spain

[5]Department of Neurology
Haukeland University Hospital and Institute of Clinical Medicine
University of Bergen
5021 Bergen
Norway

[6]Department of Forensic Genetics and Forensic Toxicology
National Board of Forensic Medicine
SE 581 33 Linkoping
Sweden

[7]Department of Genetics, Development, and Molecular Biology
Aristotle University of Thessaloniki
GR-540 06 Thessaloniki
Greece

[8]Laboratoire d'Empreintes Génétiques
EFS-RA site de Lyon
69007 Lyon
France

[9]Institute of Legal Medicine
Medical University Innsbruck
A-6020 Innsbruck
Austria

[10]Department of Forensic Medicine
University of Helsinki
Helsinki FIN-00014
Finland

[11]Department of Medical Genetics
Medical University Warsaw
02-007 Warsaw
Poland

[12]Istituto di Medicina Legale
University of Ancona
I-60020 Ancona
Italy

[13]Molecular Neuropharmacology Group and Center for Pharmacogenomics Department of Neuroscience and Pharmacology
University of Copenhagen
2200 Copenhagen
Denmark

[14]Research Institute of Biological Psychiatry and Center for Pharmacogenomics
Mental Health Center Sct. Hans
Copenhagen University Hospital
DK-4000 Roskilde
Denmark

[15]Department of Internal Medicine, Genetics Laboratory
Erasmus University Medical Center Rotterdam
3000 CA Rotterdam
The Netherlands

[16]Department of Epidemiology
Erasmus University Medical Center Rotterdam
3000 CA Rotterdam
The Netherlands

[17]Institute of Epidemiology
Helmholtz Zentrum München - German Research Center for Environmental Health
D-85764 Neuherberg
Germany

[18]Institute of Medical Informatics, Biometry and Epidemiology
Ludwig-Maximilians University
D-81377 Munich
Germany

[19]Institut für Medizinische Molekularbiologie
Christian-Albrechts University Kiel
D-24105 Kiel
Germany

[20]Cologne Center for Genomics and Institut für Genetik
University of Cologne
D-50674 Cologne
Germany

[21]Genetics
GlaxoSmithKline
Research Triangle Park, North Carolina 27709

*Correspondence: m.kayser@erasmusmc.nl

[22]These authors contributed equally to this work
[23]These authors contributed equally to this work

Summary

Understanding the genetic structure of the European population is important, not only from a historical perspective, but also for the appropriate design and interpretation of genetic epidemiological studies. Previous population genetic analyses with autosomal markers in Europe either had a wide geographic but narrow genomic coverage [1, 2], or vice versa [3–6]. We therefore investigated Affymetrix GeneChip 500K genotype data from 2,514 individuals belonging to 23 different subpopulations, widely spread over Europe. Although we found only a low level of genetic differentiation between subpopulations, the existing differences were characterized by a strong continent-wide correlation between geographic and genetic distance. Furthermore, mean heterozygosity was larger, and mean linkage disequilibrium smaller, in southern as compared to northern Europe. Both parameters clearly showed a clinal distribution that provided evidence for a spatial continuity of genetic diversity in Europe. Our comprehensive genetic data are thus compatible with expectations based upon European population history, including the hypotheses of a south-north expansion and/or a larger effective population size in southern than in northern Europe. By including the widely used CEPH from Utah (CEU) samples into our analysis, we could show that these individuals represent northern and western Europeans reasonably well, thereby confirming their assumed regional ancestry.

Table 1. European Subpopulation Summary Statistics

Subpopulation	Code	Total No. Samples	Final No. Samples*	Sex Ratio (M:F)
Norway (Førde)	NO	52	52	1.74
Sweden (Uppsala)	SE	50	46	all male
Finland (Helsinki)	FI	47	47	0.74
Ireland	IE	37	35	4.29
UK (London)	UK	197	194	8.85
Denmark (Copenhagen)	DK	60	59	1.22
Netherlands (Rotterdam)	NL	292	280	all female
Germany I (Kiel)	DE1	500	494	1.08
Germany II (Augsburg)	DE2	500	489	1.02
Austria (Tyrol)	AT	50	50	all male
Switzerland (Lausanne)	CH	134	133	0.81
France (Lyon)	FR	50	50	2.13
Portugal	PT	16	16	0.78
Spain I	ES1	83	81	1.02
Spain II (Barcelona)	ES2	48	47	0.71
Italy I	IT1	107	106	1.38
Italy II (Marches)	IT2	50	49	all male
Former Yugoslavia	YU	58	55	1.90
Northern Greece	EL	51	51	1.43
Hungary	HU	17	17	0.54
Romania	RO	12	12	1.00
Poland (Warsaw)	PO	50	49	all male
Czech Republic (Prague)	CZ	53	45	0.96
Total		2,514	2,457	

Total number of samples, final number of samples after data cleaning, and the sex ratio (male:female) of the final sample data set for each subpopulation. * is after stringent quality control.

Results and Discussion

According to current theory, the autosomal gene pool of extant human populations in Europe lacks sharp discontinuities [1, 2], with the exception of known isolates such as the Finns [6, 7]. For classical genetic markers including, for example, erythrocyte antigens, changes in population genetic structure have been observed to follow a predominantly southeast-northwest gradient [1, 2], thereby apparently matching the Pleistocene settlement of Europe, the Neolithic expansion from the Fertile Crescent, and (at least in part) the postglacial resettlement of Europe during the Mesolithic. Such gradient was also observed with particular haplogroups derived from the nonrecombining part of the Y chromosome (NRY), but other NRY data revealed additional population structure in Europe that has been associated with various demographic events in prehistoric, historic, and modern times [8–10]. In contrast, the European mitochondrial DNA pool has been found to be rather homogeneous [11]. Here, we investigated the genetic structure of the European population by using 309,790 single-nucleotide polymorphisms (SNPs) in 2,457 individuals, ascertained at 23 sampling sites (henceforth referred to as "subpopulations") in 20 different European countries. The data emerged from the genotyping of 2,514 European samples with the GeneChip Human Mapping 500K Array, followed by stringent quality control (see Table 1 and Experimental Procedures for details) and represent the largest Europe-wide genetic study to date.

First, we quantified the amount of information that each SNP could potentially provide about an individual's subpopulation affiliation by using the ancestry informativeness index *In* (Figure S1 available online) [12]. The maximum *In* value (0.09) was observed for rs6730157 in the *RAB3GAP1* gene located about 68 kb away from the Lactase (*LCT*) gene. Furthermore, nine of the 20 (45%) most ancestry-informative SNPs, and 17 of the top 100 (Table S1), were from the *LCT* region and previously showed signatures of a selective sweep in CEU (Centre d'Etude du Polymorphism Humain from Utah) samples [13]. The average *In* across markers was 0.0064 (standard deviation: 0.0032), which represents only 0.93% of the maximum possible *In* of 0.69 in our study. (Note that this maximum would be attained if a SNP was fixed for one allele in 12 subpopulations and for the other allele in the remaining 11 subpopulations).

Second, we performed a principal-component analysis (PCA) in which the first two PCs were found to account for 31.6% and 17.3%, respectively, of the total variation, an amount similar to that reported in previous studies [1, 5]. In our study, the first two PCs revealed a SNP-based grouping of European subpopulations that was strongly reminiscent of the geographic map of Europe (Figure 1; Figure S2). The first PC aligned subpopulations according to latitude, with the two Italian subpopulations at one end and the Finnish subpopulation at the other. The second PC tended to separate subpopulations more according to longitude, with the Finnish subpopulation showing the largest values and the Irish and UK subpopulations showing the lowest values. The apparent geographic footing of the two PCs received additional support from an observed statistically significant positive correlation (Pearson $r^2 = 0.632$, two-tailed $p < 10^{-15}$) between the genetic distance (Euclidian distance between the median first two eigenvectors of the PCA) and the geographic (great-circle) distance between the analyzed subpopulations.

Third, we searched for genetic barriers [14] in our dataset by using the same genetic and geographic distance matrices. This analysis identified two statistically significant barriers for the 23 subpopulations. One barrier was observed between the Finnish and all other subpopulations (first PC considering FI against the rest: $r^2 = 0.074$, two-tailed $p < 10^{-15}$; second PC considering FI against the rest: $r^2 = 0.33$, two-tailed $p < 10^{-15}$) and the other one between the two Italian and all other subpopulations (first PC considering IT1 and IT2 against the

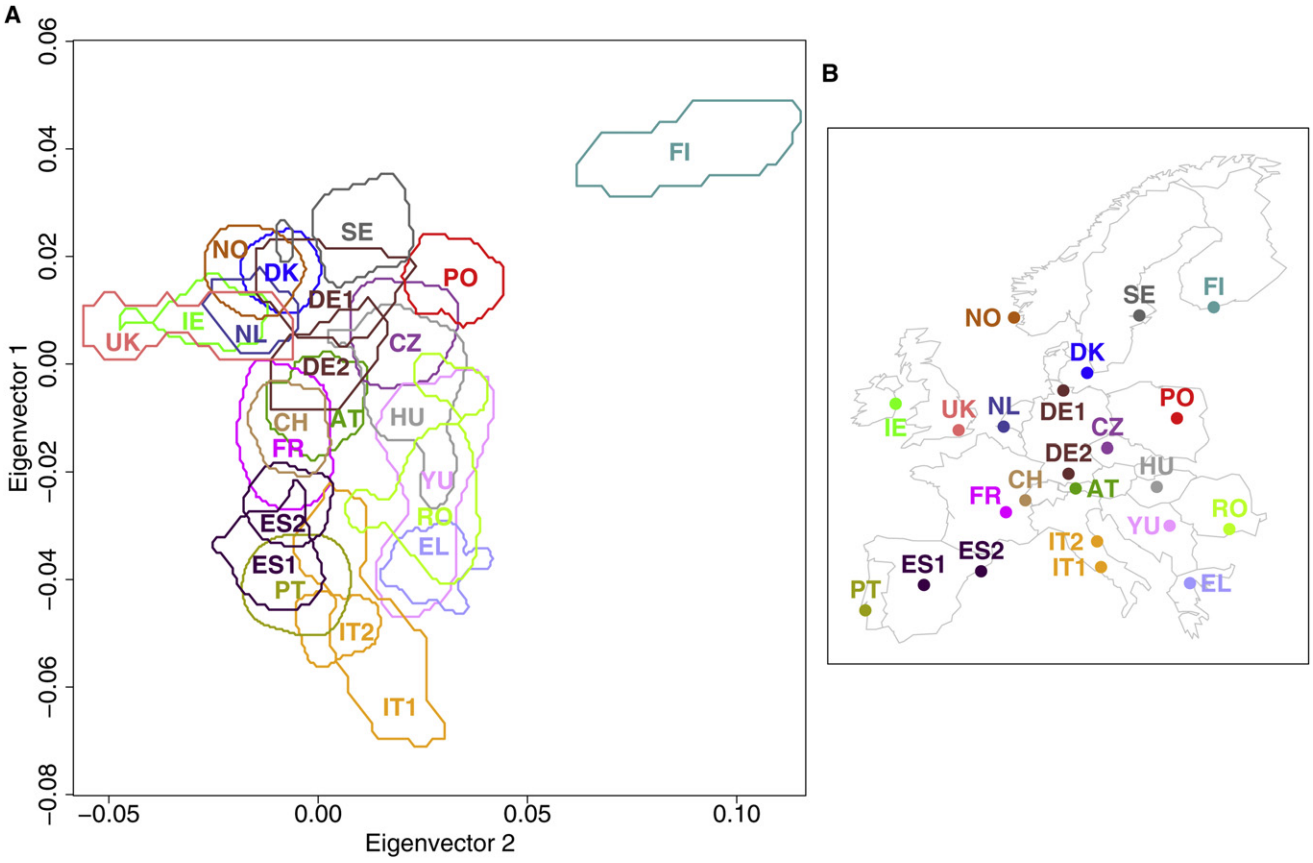

Figure 1. SNP-Based PCA of 2,457 European Individuals from 23 Subpopulations

(A) Kernel density plot of the first two dimensions of a SNP-based PCA using those 309,790 SNPs from the GeneChip Human Mapping 500K Array Set (Affymetrix) that passed quality control.

(B) Geographic distribution of the 23 subpopulations; capitals were used as the respective landmark if location information was either unspecific or lacking (see Table 1 for further sample details).

rest: $r^2 = 0.37$, two-tailed $p < 10^{-15}$; second PC considering IT1 and IT2 against the rest: $r^2 = 0.014$, two-tailed $p = 2.31 \times 10^{-9}$).

Fourth, we studied the geographic distribution of genetic diversity by computing mean heterozygosity and mean linkage disequilibrium (LD) based upon HR^2 [15] between markers at a distance < 10 kb for each subpopulation. Results from both analyses showed that the genetic diversity tended to be larger, and the LD smaller, in southern Europe as compared to northern Europe (Figure 2). Moreover, both analyses supported a genetic gradient of south-north orientation (r^2 adjusted for the number of data points between the mean observed heterozygosity and latitude: 0.76, $p = 3.80 \times 10^{-8}$; adjusted r^2 between HR^2 and latitude: 0.71, two-tailed $p = 4.33 \times 10^{-7}$) but not of west-east orientation (adjusted r^2 between heterozygosity and longitude: 0.03, two-tailed $p = 0.416$; adjusted r^2 between HR^2 and longitude: 0.099, two-tailed $p = 0.078$). Spatial autocorrelation analysis of both variables revealed statistically significant ($p < 0.05$) patterns compatible with a clinal distribution as indicated by the presence of positive and statistically significant autocorrelation values for small pair-wise distances and negative and statistically significant Moran's I values for large distances (see Figure 2). Bearing analysis [16] revealed for the heterozygosity measure the maximal angular correlations ($r = 0.69$) at 87° and the minimal ($r = -0.153$) at 165°, as well as for HR^2 the maximal at 55° ($r = 0.67$) and the minimal ($r = -0.167$) at 160°, thus also suggesting a south-to-north spatial distribution of both variable. These results are compatible with larger effective population sizes in the south than in the north of Europe and/or a population expansion from southern toward northern Europe. Hierarchical analysis of molecular variance (AMOVA) [17] revealed that clustering the individuals according to four geographic groups—north (NO, SE, FI), north-west/central (IE, UK, DK, NL, DE1, DE2, AT, CH, FR), east (HU, RO, PO, CZ), and south (PT, ES1, ES2, IT1, IT2, YU, EL)—explained an average of 0.17% (95% coefficient interval: 0.0% to 0.91%) of the total genetic variance, whereas individual subpopulation affiliation explained 0.25% (95% coefficient interval: 0.0% to 1.25%).

Overall, our study showed that the autosomal gene pool in Europe is comparatively homogeneous but at the same time revealed that the small genetic differentiation that is present between subpopulations is characterized by a significant correlation between genetic and geographic distance. Furthermore, the qualitative nature of these results is in close agreement with expectations based on human migration history in Europe. The major prehistoric waves of human migration in Europe followed south and southeastern to north and northwestern directions [1], including the first Paleolithic settlement of the continent by anatomically modern humans [18], most of the postglacial resettlement during the Mesolithic [19], and the farming-related population expansion during the Neolithic [18, 20]. Thus, both the level and the change in neutral autosomal

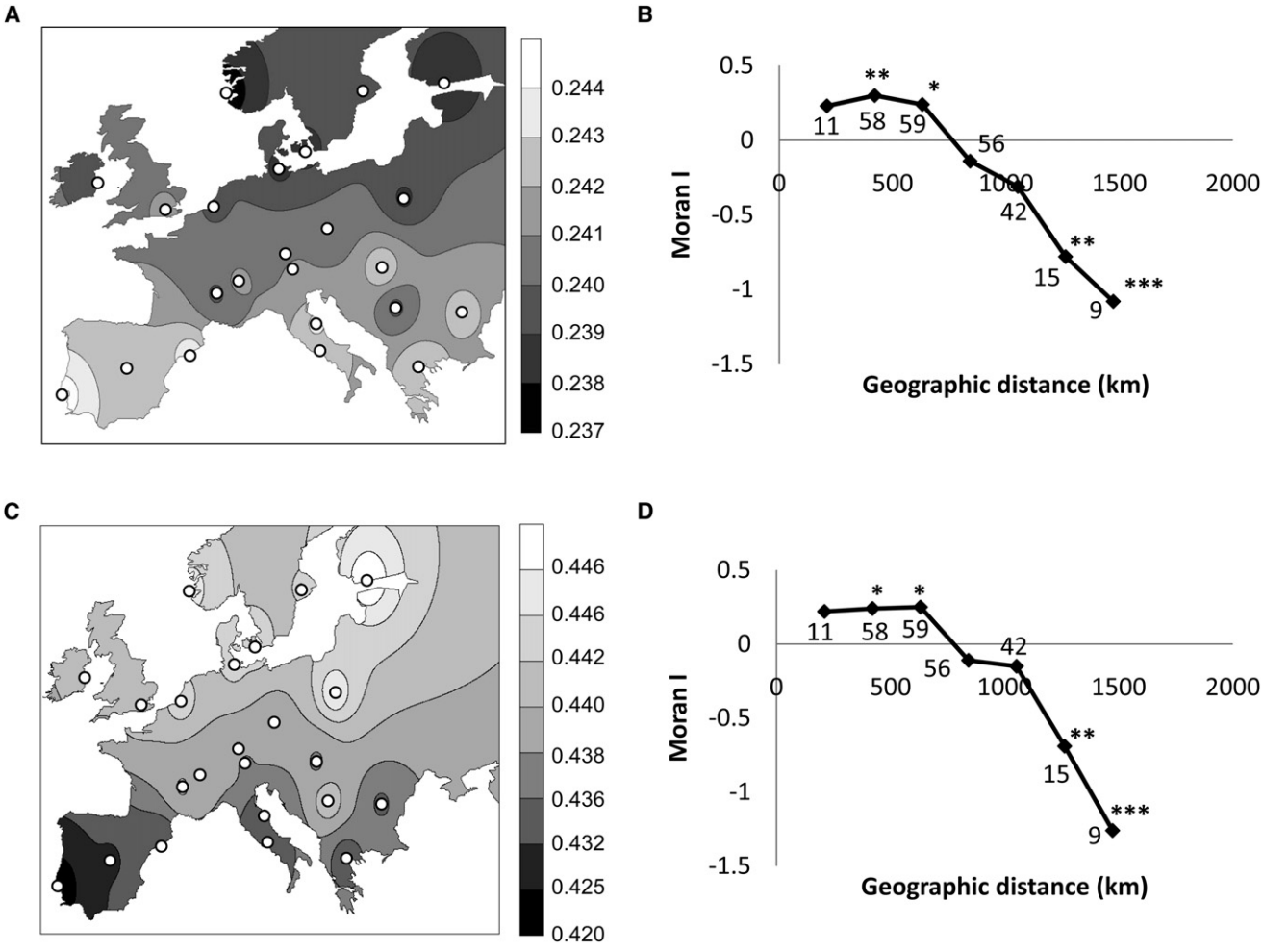

Figure 2. Geographic Distribution of Two Measures of Genetic Diversity across the European Population

(A and B) Isoline map (A) of Europe based on the mean observed heterozygosity in each of 23 European subpopulations with (B) corresponding spatial autocorrelogram.

(C and D) Isoline map (C) of Europe based on the mean observed linkage disequilibrium based on HR^2 in each of 23 European subpopulations with (D) corresponding spatial autocorrelogram. Both spatial autocorrelograms showed statistically significant departures from randomness ($p < 0.05$). For each distance class, the number of subpopulation pairs included and the statistical significance (*, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$) are provided.

variation in Europe can be expected to roughly follow southern-to-northern gradients as we observed, with the possible exception of population isolates as observed for the Finns. On the other hand, migration events in more recent (i.e., historic) times are presumed to have had a more homogenizing effect upon the previously established genetic landscape, as a result of their sporadic nature and haphazard geographic orientation [2]. This implies that genetic differences between extant European subpopulations can be expected to be small indeed. The genetic landscape described by the ~300,000 autosomal SNPs analyzed here closely resembles that previously obtained with 128 alleles from 49 classical markers (see Table 1.3.1 in [1]). This similarity is highlighted by a significant correlation ($r = 0.516$; two-tailed Mantel test $p = 0.0042$, performed with 10,000 Monte Carlo permutations) between the pair-wise F_{ST} values [21] computed for the 19 European subpopulations that overlapped between the two datasets (Danish, Dutch, Yugoslavian, Hungarian, Irish, Italian, Portuguese, Spanish, Swiss, English, German, Austrian, Finnish, French, Greek, Norwegian, Polish, Swedish, and Czechoslovakian). This notwithstanding, a stronger correlation between F_{ST} and great-circle geographic distances was observed for the subpopulations when the SNPs from our study were used ($r = 0.661$; two-tailed Mantel test $p = 0.00010$, performed with 10,000 Monte Carlo permutations) as compared to the classical markers ($r = 0.503$, two-tailed Mantel test $p = 0.00020$, performed with 10,000 Monte Carlo permutations).

Previous studies based on genome-wide SNP diversity reported differences between individuals of southern and northern/central European ancestry [3, 5, 6] and, to a lesser extent, between those of eastern and western European ancestry [3], which were not confirmed in our study. They mostly relied on the analysis of European Americans whose geographic assignment was determined from self-reported family records. Although genetic studies using European Americans can reveal important information about the genetic structure of the European ancestry of European Americans, caution must be exercised when drawing conclusions about the current genetic structure of Europe from European Americans because (1) European migrants may not have been representative of their country of origin, (2) the temporal difference introduced by sampling second- or third-generation descendants means

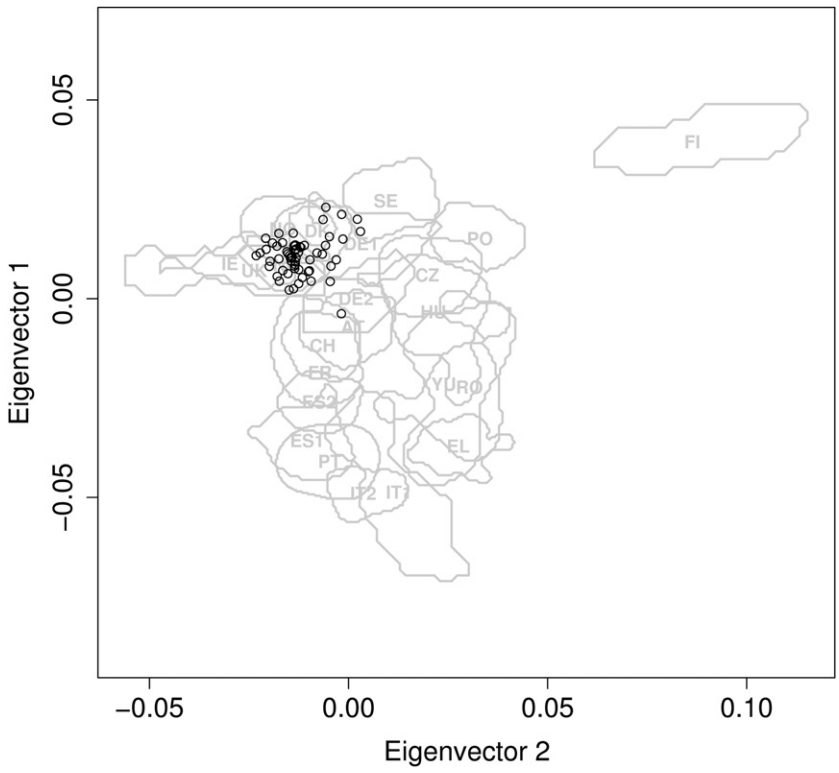

Figure 3. Position of CEPH-CEU Samples in a SNP-Based PCA Kernel-Density Plot of 23 European Subpopulations

CEU individuals (U.S. Americans of European descent from Utah) are plotted as open circles. For details, see Figure 1 and Table 1.

that allele-frequency estimates inevitably ignored recent population movements (i.e., WWII-related migrations), and (3) self-reported geographic origin is error prone [22]. Our study avoided these potential pitfalls by using large samples of individuals of genuinely European origin, as evidenced by the documentation of their respective place of birth or residence being in one of the named subpopulations, and with comprehensive continent-wide coverage.

It is of general interest to place the CEU samples, widely used in genetic epidemiological and population genetic studies as representing the European population, into the context of our findings. The CEPH-CEU panel comprises U.S. Americans who were collected in Utah in 1980 and who are assumed to have descended from migrants originating from northern and western parts of Europe [23]. The samples were also included in the International HapMap Project and formed the basis of selecting tagging SNPs used in current genome-wide association studies with Illumina SNP arrays. Whereas a previous study [3] confirmed the grouping of the CEPH-CEU samples with other northern and western European subpopulations, our study was capable of providing their most precise positioning on the European genetic map (Figure 3). It turned out that, while the CEPH-CEU panel was indeed largely representative of northwestern and central Europeans, parts of Scandinavia as well as southern and eastern Europe were not well represented by these samples (Figure 3). Estimated inflated false-positive rates for all subpopulations were largest in the Finns, followed by the two Italian subpopulations (see Table S2). This implies that researchers conducting genetic-association studies in at least these regions, using the CEPH-CEU samples as controls, may be at increased risk of false-positive associations. Our confirmation of the regional European origin of the CEPH-CEU samples also indicates that inferring the geographic origin of an unknown person from autosomal DNA markers, which is highly relevant in the forensic context, might now be feasible down to the level of European subregions, at least when a large number of genetic markers and a reference database, such as are applied here, are used.

Conclusions

Our comprehensive SNP genotype data from 23 European subpopulations, providing a dense coverage at both the geographic and genomic level and representing the largest Europe-wide genetic study to date, allowed us to describe the genetic structure of the European population with the highest resolution. Although the amount of differentiation within the European autosomal gene pool was found to be small, the existing genetic differences nevertheless correlated well with geographic distances. Furthermore, mean heterozygosity was larger, and mean linkage disequilibrium smaller, in southern than in northern European subpopulations, and both parameters exhibited a continuous clinal distribution across Europe. Overall, our results were compatible with expectations based on European population history, mainly the prehistoric population expansion from southern to northern Europe and/or a larger effective population size in the south as compared to the north of Europe. Our dataset also allowed placement of the widely used CEPH-CEU samples onto the European genetic landscape, essentially confirming their genetic ancestry in northern and western Europe.

Experimental Procedures

Samples and Genotyping

The GeneChip Human Mapping 500K Array Set (Affymetrix) was used to genotype 500,568 SNPs in 2,514 individuals from 23 different sampling sites (henceforth termed "subpopulations") located in one of 20 different European countries. Genotyping according to the instructions provided by the manufacturer was carried out at one of seven specialized centers: the Cologne Center for Genomics at the University of Cologne (Germany) for DE1, NO, SE, FI, AT, FR, ES2, IT2, EL, PO, and CZ; the Helmholtz Zentrum München - German Research Center for Environmental Health for DE2; the genetics laboratory of the Department of Internal Medicine, Erasmus MC (Netherlands) for NL; and the RH Microarray Centre Rigshospitalet, Copenhagen University Hospital (Denmark) for DK (see Table 1 for abbreviation explanations). Samples from the GlaxoSmithKline-sponsored POPRES project (IE, UK, CH, PT, ES1, IT1, YU, HU, and RO) were genotyped at Expression Analysis (Durham, NC, USA) and at Gene Logic (Gaithersburg, MD, USA) (see Table 1 for abbreviation explanations). Some samples belonged to existing control population studies, with detailed descriptions available elsewhere: KORA [24] for DE2, PopGen [25] for DE1, the Rotterdam Study [26–28] for NL, and POPRES (drawn from the LOLIPOP and CoLaus studies) for IE, UK, CH, PT, ES1, IT1, YU, HU, and RO [29–31]. Samples were drawn randomly from these pools or, in the case of POPRES, were ascertained on the basis of sample-size requirements. European migrants from non-European regions were not included in the initial analysis. For 11 of the subpopulations (NO, SE, FI, AT, FR, ES2, IT2, EL, PO, CZ, and DK), samples were

obtained from healthy unrelated volunteers: Norwegian samples (NO) from blood donors of the Førde region, Swedish samples (SE) from the Uppsala region [32], Finnish samples (FI) from the Helsinki area with parents and grandparents originating from various regions in Finland, Austrian samples (AT) from the Tyrol region with parents originating from Tyrol, French samples (FR) from blood donors of Lyon with parents originating from the Rhône Alpes area, Spanish samples (ES2) from Catalonia of blood donors from rural areas who speak Catalan as their mother tongue and who had regional Catalan ancestry for at least two generations [33], Italian samples (IT2) from blood donors of the upland of the Marches region [34], Greek samples (EL) from the north of the country [35], Polish samples (PO) from the Warsaw region of central Poland [36], Czech samples (CZ) from the central Bohemian region in and around Prague, and Danish samples (DK) from the Danish Blood Donor Corps in the Copenhagen area. In addition, GeneChip Human Mapping 500K Array data from CEPH-CEU samples were retrieved from the Affymetrix website (http://www.affymetrix.com).

Quality Assessment and Control Procedure

Array-based SNP genotypes were subjected to stringent quality control: First, each individual was required to have a genotype call rate $\geq$ 93%, with the dynamic model (DM) algorithm with a confidence score of 0.26, and a per-individual call rate $\geq$ 95% for all individuals genotyped by the same facility, with the Bayesian robust linear model with Mahalanobis distance classifier (BRLMM) algorithm with a confidence score of 0.5. The call rate was defined here as the proportion of unambiguous genotypes among either all SNPs (per-individual call rate) or all individuals (per-marker call rate), respectively. Markers that were monomorphic (1.4% of the total), that were located on the X chromosome (2.1%), or that had a per-marker call rate $\leq$ 90% in at least one genotyping facility (5.7%) were excluded, as were those showing a significant ($p \leq 0.05$) deviation from Hardy-Weinberg equilibrium (HWE) in at least one subpopulation (31.3%). HWE was tested by means of a χ^2 test, or by Fisher's exact test when the observed or expected number of a given genotype was less than 5. This method was preferred over others that have been shown to be more powerful [37] because the computational requirements of these methods increase exponentially with sample size and were thus too resource intensive for our study. The average proportion of heterozygous genotypes at X chromosomal markers was estimated per individual in order to detect false gender assignments. Male subjects can be expected to show X chromosomal heterozygosity proportions $\leq$ 1%, reflecting the overall genotyping error rate, and female subjects should show proportions near the average heterozygosity (26%) of the analyzed X chromosomal SNPs. Average identity-by-state (IBS) distances were calculated for a given set of markers as the average genetic dissimilarity between pairs of individuals. Analysis of IBS values within subpopulations allowed us to detect two types of outliers: (1) cognate relatives, i.e., individuals that were genetically more similar than expected to another member of the same subpopulation, and (2) "aliens," i.e., individuals that were far less genetically similar than expected to the rest of the subpopulation. Formally, cognate relatives were defined as pairs of individuals having a pair-wise IBS value larger than the so-called "Tukey outlier criterion" when compared with the rest of pairs of individuals of the same subpopulation, i.e., the median IBS plus three times the interquartile range (IQR) in that subpopulation. In this case, the partner with the lower call rate was excluded. Aliens were defined as individuals with at least 60% of their pair-wise IBS values below the median minus three times the IQR. These two criteria led to the exclusion of 56 individuals from further analysis (Table 1). One individual identified as female had an average proportion of heterozygous X chromosomal markers of only 0.6% and was thus excluded from further analysis. In total, quality control left 2,457 individuals (97.6%) and 309,790 markers (62.4%) for inclusion in subsequent analysis. AMOVA [17] was performed to ascertain the magnitude of variation attributable to the respective genotyping center or subpopulation. The mean amount of genetic variance explained among genotyping centers was 0.095% (95% confidence interval: 0% to 0.71%), whereas subpopulation affiliation explained 0.63% of the variance (95% confidence interval: 0% to 2.86%). As expected, the largest amount of genetic variation was explained by differences between individuals (99.72%; 95% confidence interval: 98.61% to 100.00%). Data are available on request from the authors according to the regulations of the participating studies and sample cohorts.

Statistical Data Analyses

The ancestry-informativeness index *In* was estimated for each marker as described eslewhere [12]. Principal-component analysis was performed with the *Eigensoft* program with the default settings [38]. Population-wise kernel densities were computed from the first two PCs with the *adehabitat* R package [39] and subjected to least-squares crossvalidation [40] that used 80% of individuals per subpopulation for training. Pearson correlation coefficients were computed for the genetic distance between the subpopulations (represented by the respective median over all individuals in that subpopulation of the first two eigenvectors) and the great-circle geographic distance. The statistical significance of these correlation coefficients was assessed by means of a Mantel test [41]. Barrier analysis was performed on the basis of the Monmonier's algorithm [14]. Locus-wise AMOVA [17] was conducted after clustering the European subpopulations by genotyping center as well as by the use of four geographic groups. Negative percentages of explained variation were settled to 0. Both mean heterozygosity and mean linkage disequilibrium computed by means of HR^2 [15] were computed with a subsample of ten individuals per population in order to adjust for possible influence of sample size [42]. Spatial autocorrelation and Bearing analyses were performed with the software PASSAGE 1.1 [43]. Isoline maps were performed with the Golden Surfer 8 software [44], with the inverse-distance method used for interpolation points. Isoline levels were defined to include the value of at least one of the 23 populations with intervals of 0.001 in the case of heterozygosity and 0.002 in the case of HR^2. For evaluation of the extent to which the CEPH-CEU samples are representative of the subpopulations used in the present study, marker-wise tests of association (Fisher's exact test) were performed each time with the CEPH-CEU samples as "controls" and a given subpopulation as "cases." The false-positive rate was defined as the percentage of markers yielding a p value < 0.05. If the CEPH-CEU samples were representative of a subpopulation, the false-positive rate would be around 0.05, whereas higher false-positive rates indicate that the CEPH-CEU samples may not be representative of the respective subpopulation.

Supplemental Data

Supplemental Data include two tables and two figures and can be found with this article online at http://www.current-biology.com/cgi/content/full/18/16/1241/DC1/.

Acknowledgments

All volunteers are gratefully acknowledged for sample donation. We thank the following colleagues for their help and support: J. Kooner and J. Chambers of the LOLIPOP study and D. Waterworth, V. Mooser, G. Waeber, and P. Vollenweider of the CoLaus study for providing access to their collections via the GlaxoSmithKline-sponsored Population Reference Sample (POPRES) project; K. King for preparing the POPRES data; M. Simoons, E. Sijbrands, A. van Belkum, J. Laven, J. Lindemans, E. Knipers, and B. Stricker for their financial contribution to the generation of the Rotterdam Study dataset; P. Arp, M. Jhamai, W. van IJken, and R. van Schaik for generating the Rotterdam Study dataset; T. Meitinger, P. Lichtner, G. Eckstein, and all other members of the Helmholtz Zentrum München genotyping staff for generating the KORA Study dataset; H. von Eller-Eberstein for management of the PopGen project; F.C. Nielsen, R. Borup, C. Schjerling, H. Ullum, E. Haastrup, and numerous colleagues at the Copenhagen University Hospital Blood Bank for assistance in making the Danish data available; and S. Brauer for DNA sample management. We would additionally like to thank Affymetrix for making the GeneChip Human Mapping 500K Array genotypes of the CEPH-CEU trios publicly available and the Centre d'Etude du Polymorphisme Humain (CEPH) for the original sample collection. We are grateful to three anonymous reviewers for their comments, which stimulated us to improve the manuscript. This work was supported by the Netherlands Forensic Institute to M.Ka.; Affymetrix to M.Ka. and M.Kr.; the German National Genome Research Network and the German Federal Ministry of Education and Research to H.-E.W., S.S., M.Kr., and P.N. (01GR0416 to P.N.); the Helmholtz Zentrum München - German Research Center for Environmental Health, Neuherberg, and the Munich Center of Health Sciences as part of LMUinnovativ to H.-E.W.; the Netherlands Organization for Scientific Research (NWO 175.010.2005.011) to A.G.U.; the European Commission to A.G.U. (GEFOS; 201865) and A.S. (LD Europe; QLG2-CT-2001-00916); the Czech Ministry of Health (VZFNM 00064203 and IGA NS/9488-3) to M.M.; Helse-Vest, Regional Health Authority Norway to L.A.B., the Swedish National Board of Forensic Medicine (RMVFoU 99:22, 02:20) to G.H.; and the Academy of Finland to A.S. (80578, OMLL) and J.P. (109265 and 111713). None of the funding organizations had any influence on the design, conduct, or conclusions of the study.

Received: May 2, 2008
Revised: July 9, 2008
Accepted: July 10, 2008
Published online: August 7, 2008

References

1. Cavalli-Sforza, L.L., Menozzi, P., and Piazza, A. (1994). The History and Geography of Human Genes (Princeton, NJ: Princeton University Press).
2. Sokal, R.R., Harding, R.M., and Oden, N.L. (1989). Spatial patterns of human gene frequencies in Europe. Am. J. Phys. Anthropol. *80*, 267–294.
3. Bauchet, M., McEvoy, B., Pearson, L.N., Quillen, E.E., Sarkisian, T., Hovhannesyan, K., Deka, R., Bradley, D.G., and Shriver, M.D. (2007). Measuring European population stratification with microarray genotype data. Am. J. Hum. Genet. *80*, 948–956.
4. Price, A.L., Butler, J., Patterson, N., Capelli, C., Pascali, V.L., Scarnicci, F., Ruiz-Linares, A., Groop, L., Saetta, A.A., Korkolopoulou, P., et al. (2008). Discerning the ancestry of European Americans in genetic association studies. PLoS Genet *4*, e236.
5. Tian, C., Plenge, R.M., Ransom, M., Lee, A., Villoslada, P., Selmi, C., Klareskog, L., Pulver, A.E., Qi, L., Gregersen, P.K., et al. (2008). Analysis and application of European genetic substructure using 300 K SNP information. PLoS Genet *4*, e4.
6. Seldin, M.F., Shigeta, R., Villoslada, P., Selmi, C., Tuomilehto, J., Silva, G., Belmont, J.W., Klareskog, L., and Gregersen, P.K. (2006). European population substructure: Clustering of northern and southern populations. PLoS Genet *2*, e143.
7. Sajantila, A., Salem, A.H., Savolainen, P., Bauer, K., Gierig, C., and Paabo, S. (1996). Paternal and maternal DNA lineages reveal a bottleneck in the founding of the Finnish population. Proc. Natl. Acad. Sci. USA *93*, 12035–12039.
8. Roewer, L., Croucher, P.J., Willuweit, S., Lu, T.T., Kayser, M., Lessig, R., de Knijff, P., Jobling, M.A., Tyler-Smith, C., and Krawczak, M. (2005). Signature of recent historical events in the European Y-chromosomal STR haplotype distribution. Hum. Genet. *116*, 279–291.
9. Rosser, Z.H., Zerjal, T., Hurles, M.E., Adojaan, M., Alavantic, D., Amorim, A., Amos, W., Armenteros, M., Arroyo, E., Barbujani, G., et al. (2000). Y-chromosomal diversity in Europe is clinal and influenced primarily by geography, rather than by language. Am. J. Hum. Genet. *67*, 1526–1543.
10. Kayser, M., Lao, O., Anslinger, K., Augustin, C., Bargel, G., Edelmann, J., Elias, S., Heinrich, M., Henke, J., Henke, L., et al. (2005). Significant genetic differentiation between Poland and Germany follows present-day political borders, as revealed by Y-chromosome analysis. Hum. Genet. *117*, 428–443.
11. Simoni, L., Calafell, F., Pettener, D., Bertranpetit, J., and Barbujani, G.V. (2000). Geographic patterns of mtDNA diversity in Europe. Am. J. Hum. Genet. *66*, 262–278.
12. Rosenberg, N.A., Li, L.M., Ward, R., and Pritchard, J.K. (2003). Informativeness of genetic markers for inference of ancestry. Am. J. Hum. Genet. *73*, 1402–1422.
13. Voight, B.F., Kudaravalli, S., Wen, X., and Pritchard, J.K. (2006). A map of recent positive selection in the human genome. PLoS Biol. *4*, e72.
14. Manni, F.C., Guérard, E., and Heyer, G.E. (2004). Geographic patterns of (genetic, morphologic, linguistic) variation: How barriers can be detected by "Monmonier's algorithm.". Hum. Biol. *76*, 173–190.
15. Sabatti, C., and Risch, N. (2002). Homozygosity and linkage disequilibrium. Genetics *160*, 1707–1719.
16. Falsetti, A.B., and Sokal, R.R. (1993). Genetic structure of human populations in the British Isles. Ann. Hum. Biol. *20*, 215–229.
17. Excoffier, L., Smouse, P.E., and Quattro, J.M.V. (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: Application to human mitochondrial DNA restriction data. Genetics *131*, 479–491.
18. Belle, E.M., Landry, P.A., and Barbujani, G. (2006). Origins and evolution of the Europeans' genome: Evidence from multiple microsatellite loci. Proc Biol Sci *273*, 1595–1602.
19. Torroni, A., Bandelt, H.J., Macaulay, V., Richards, M., Cruciani, F., Rengo, C., Martinez-Cabrera, V., Villems, R., Kivisild, T., Metspalu, E., et al. (2001). A signal, from human mtDNA, of postglacial recolonization in Europe. Am. J. Hum. Genet. *69*, 844–852.
20. Chikhi, L., Nichols, R.A., Barbujani, G., and Beaumont, M.A.V. (2002). Y genetic data support the Neolithic demic diffusion model. Proc. Natl. Acad. Sci. USA *99*, 11008–11013.
21. Weir, B.S., and Cockerham, C.C. (1984). Estimating F-statistics for the analysis of population structure. Evolution Int. J. Org. Evolution *38*, 1358–1370.
22. Burnett, M.S., Strain, K.J., Lesnick, T.G., de Andrade, M., Rocca, W.A., and Maraganore, D.M. (2006). Reliability of self-reported ancestry among siblings: Implications for genetic association studies. Am. J. Epidemiol. *163*, 486–492.
23. Dausset, J., Cann, H., Cohen, D., Lathrop, M., Lalouel, J.M., and White, R. (1990). Centre d'etude du polymorphisme humain (CEPH): Collaborative genetic mapping of the human genome. Genomics *6*, 575–577.
24. Lowel, H., Doring, A., Schneider, A., Heier, M., Thorand, B., Meisinger, C., and Group, M.K.S. (2005). The MONICA Augsburg surveys–basis for prospective cohort studies. Gesundheitswesen *67* (*Suppl 1*), S13–S18.
25. Krawczak, M., Nikolaus, S., von Eberstein, H., Croucher, P.J., El Mokhtari, N.E., and Schreiber, S. (2006). PopGen: Population-based recruitment of patients and controls for the analysis of complex genotype-phenotype relationships. Community Genet. *9*, 55–61.
26. Hofman, A., Breteler, M.M., van Duijn, C.M., Krestin, G.P., Pols, H.A., Stricker, B.H., Tiemeier, H., Uitterlinden, A.G., Vingerling, J.R., and Witteman, J.C. (2007). The Rotterdam Study: Objectives and design update. Eur. J. Epidemiol. *22*, 819–829.
27. Hofman, A., Grobbee, D.E., de Jong, P.T., and van den Ouweland, F.A. (1991). Determinants of disease and disability in the elderly: The Rotterdam Elderly Study. Eur. J. Epidemiol. *7*, 403–422.
28. Kayser, M., Liu, F., Janssens, A.C., Rivadeneira, F., Lao, O., van Duijn, K., Vermeulen, M., Arp, P., Jhamai, M.M., van Ijcken, W.F., et al. (2008). Three genome-wide association studies and a linkage analysis identify HERC2 as a human iris color gene. Am. J. Hum. Genet. *82*, 411–423.
29. Kooner, J.S., Chambers, J.C., Aguilar-Salinas, C.A., Hinds, D.A., Hyde, C.L., Warnes, G.R., Gomez Perez, F.J., Frazer, K.A., Elliott, P., Scott, J., et al. (2008). Genome-wide scan identifies variation in MLXIPL associated with plasma triglycerides. Nat. Genet. *40*, 149–151.
30. Nelson, M.R., Bacanu, S.A., Mosteller, M., Li, L., Bowman, C.E., Roses, A.D., Lai, E.H., and Ehm, M.G. (2008). Genome-wide approaches to identify pharmacogenetic contributions to adverse drug reactions. Pharmacogenomics J., in press. Published online February 26, 2008. 10.1038/tpj.2008.4.
31. Sandhu, M.S., Waterworth, D.M., Debenham, S.L., Wheeler, E., Papadakis, K., Zhao, J.H., Song, K., Yuan, X., Johnson, T., Ashford, S., et al. (2008). LDL-cholesterol concentrations: A genome-wide association study. Lancet *371*, 483–491.
32. Karlsson, A.O., Wallerstrom, T., Gotherstrom, A., and Holmlund, G. (2006). Y-chromosome diversity in Sweden - a long-time perspective. Eur. J. Hum. Genet. *14*, 963–970.
33. Plaza, S., Calafell, F., Helal, A., Bouzerna, N., Lefranc, G., Bertranpetit, J., and Comas, D. (2003). Joining the pillars of Hercules: mtDNA sequences show multidirectional gene flow in the western Mediterranean. Ann. Hum. Genet. *67*, 312–328.
34. Onofri, V., Alessandrini, F., Turchi, C., Fraternale, B., Buscemi, L., Pesaresi, M., and Tagliabracci, A. (2007). Y-chromosome genetic structure in sub-Apennine populations of Central Italy by SNP and STR analysis. Int. J. Legal Med. *121*, 234–237.
35. Kondopoulou, H., Loftus, R., Kouvatsi, A., and Triantaphyllidis, C. (1999). Genetic studies in 5 Greek population samples using 12 highly polymorphic DNA loci. Hum. Biol. *71*, 27–42.
36. Ploski, R., Wozniak, M., Pawlowski, R., Monies, D.M., Branicki, W., Kupiec, T., Kloosterman, A., Dobosz, T., Bosch, E., Nowak, M., et al. (2002). Homogeneity and distinctiveness of Polish paternal lineages revealed by Y chromosome microsatellite haplotype analysis. Hum. Genet. *110*, 592–600.
37. Schaid, D.J., Batzler, A.J., Jenkins, G.D., and Hildebrandt, M.A. (2006). Exact tests of Hardy-Weinberg equilibrium and homogeneity of disequilibrium across strata. Am. J. Hum. Genet. *79*, 1071–1080.
38. Patterson, N., Price, A.L., and Reich, D. (2006). Population structure and eigenanalysis. PLoS Genet *2*, e190.
39. Calenge, C. (2006). The package "adehabitat" for the R software: A tool for the analysis of space and habitat use by animals. Ecol. Modell. *197*, 516–519.

40. Silverman, B.W. (1986). Density Estimation for Statistics and Data Analysis (Boca Raton, Florida: Chapman & Hall / CRC Press).
41. Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. Cancer Res. *27*, 209–220.
42. Jakobsson, M., Scholz, S.W., Scheet, P., Gibbs, J.R., VanLiere, J.M., Fung, H.C., Szpiech, Z.A., Degnan, J.H., Wang, K., Guerreiro, R., et al. (2008). Genotype, haplotype and copy-number variation in worldwide human populations. Nature *451*, 998–1003.
43. Rosenberg, M.S. (2001). PASSAGE: Pattern Analysis, Spatial Statistics, and Geographic Exegesis. 1.1 Edition, A.S.U. Department of Biology, ed. (Tempe, AZ.).
44. Golden Software. (2007). Surfer Version 8.08.3267. Colorado, USA.