Serveur Académique Lausannois SERVAL serval.unil.ch

Author Manuscript
Faculty of Biology and Medicine Publication

This paper has been peer-reviewed but dos not include the final publisher proof-corrections or journal pagination.

Published in final edited form as:

Title: Genes mirror geography within Europe.
Authors: Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, Indap A, King KS, Bergmann S, Nelson MR, Stephens M, Bustamante CD
Journal: Nature
Year: 2008 Nov 6
Volume: 456
Issue: 7218
Pages: 98-101
DOI: 10.1038/nature07331

In the absence of a copyright statement, users should assume that standard copyright protection applies, unless the article contains an explicit statement to the contrary. In case of doubt, contact the journal publisher to verify the copyright status of an article.

serval
serveur académique lausannois

UNIL | Université de Lausanne
Faculté de biologie
et de médecine

NIH Public Access
Author Manuscript
Nature. Author manuscript; available in PMC 2009 August 31.

Published in final edited form as:
Nature. 2008 November 6; 456(7218): 98–101. doi:10.1038/nature07331.

Genes mirror geography within Europe

John Novembre[1,2], **Toby Johnson**[4,5,6], **Katarzyna Bryc**[7], **Zoltán Kutalik**[4,6], **Adam R. Boyko**[7], **Adam Auton**[7], **Amit Indap**[7], **Karen S. King**[8], **Sven Bergmann**[4,6], **Matthew R. Nelson**[8], **Matthew Stephens**[2,3], and **Carlos D. Bustamante**[7]

[1] Department of Ecology and Evolutionary Biology, Interdepartmental Program in Bioinformatics, University of California–Los Angeles, Los Angeles, California 90095, USA [2] Department of Human Genetics, University of Chicago, Chicago, Illinois 60637, USA [3] Department of Statistics, University of Chicago, Chicago, Illinois 60637, USA [4] Department of Medical Genetics, University of Lausanne, Rue de Bugnon 27 - DGM 328, CH-1005 Lausanne, Switzerland [5] University Institute for Social and Preventative Medecine, Centre Hospitalier Universitaire Vaudois (CHUV), University of Lausanne, Rue de Bugnon 27 - DGM 328, CH-1005 Lausanne, Switzerland [6] Swiss Institute of Bioinformatics, Central Administration, Quartier Sorge - Batiment Genopode, 1015 Lausanne, Switzerland [7] Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, New York 14853, USA [8] GlaxoSmithKline, Research Triangle Park, North Carolina 27709, USA

Abstract

Understanding the genetic structure of human populations is of fundamental interest to medical, forensic and anthropological sciences. Advances in high-throughput genotyping technology have markedly improved our understanding of global patterns of human genetic variation and suggest the potential to use large samples to uncover variation among closely spaced populations[1–5]. Here we characterize genetic variation in a sample of 3,000 European individuals genotyped at over half a million variable DNA sites in the human genome. Despite low average levels of genetic differentiation among Europeans, we find a close correspondence between genetic and geographic distances; indeed, a geographical map of Europe arises naturally as an efficient two-dimensional summary of genetic variation in Europeans. The results emphasize that when mapping the genetic basis of a disease phenotype, spurious associations can arise if genetic structure is not properly accounted for. In addition, the results are relevant to the prospects of genetic ancestry testing[6]; an individual's DNA can be used to infer their geographic origin with surprising accuracy—often to within a few hundred kilometres.

Recent studies suggest that by combining high-throughput genotyping technologies with dense geographic samples one can shed light on unanswered questions regarding human population structure[1–5]. For instance, it is not clear to what extent populations within continental regions exist as discrete genetic clusters versus as a genetic continuum, nor how

Correspondence and requests for materials should be addressed to J.N. (jnovembre@ucla.edu).

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Author Contributions M.R.N. coordinated sample collection and genotyping. K.S.K., A.I., J.N. and A.R.B. performed quality control and prepared genotypic and demographic data for further analyses. C.B., M.S., M.R.N., S.B., J.N., T.J., K.B., Z.K., A.R.B. and A.A. all contributed to the design of analyses. J.N., S.B., T.J., K.B. and Z.K. performed PCA analyses. M.S. and J.N. designed and performed assignment-based analyses. T.J. and J.N. performed genome-wide association simulations. J.N., C.B., M.S., M.R.N. and A.A. wrote the paper. All authors discussed the results and commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

precisely one can assign an individual to a geographic location on the basis of their genetic information alone.

To investigate these questions, we surveyed genetic variation in a sample of 3,192 European individuals collected and genotyped as part of the larger Population Reference Sample (POPRES) project[7]. Individuals were genotyped at 500,568 loci using the Affymetrix 500K single nucleotide polymorphism (SNP) chip. When available, we used the country of origin of each individual's grandparents to determine the geographic location that best represents each individual's ancestry, otherwise we used the self-reported country of birth (see Methods and Supplementary Tables 1 and 2). After removing SNPs with low-quality scores, we applied various stringency criteria to avoid sampling individuals from outside of Europe, to create more even sample sizes across Europe, to exclude individuals with grand-parental ancestry from more than location, and to avoid potential complications of SNPs in high linkage disequilibrium (see Methods and Supplementary Table 3). Although our main result holds even when we relax nearly all of these stringency criteria, we focus our analyses on genotype data from 197,146 loci in 1,387 individuals (Supplementary Table 2), for whom we have high confidence of individual origins.

We used principal components analysis (PCA; ref. 8) to produce a two-dimensional visual summary of the observed genetic variation. The resulting figure bears a notable resemblance to a geographic map of Europe (Fig. 1a). Individuals from the same geographic region cluster together and major populations are distinguishable. Geographically adjacent populations typically abut each other, and recognizable geographical features of Europe such as the Iberian peninsula, the Italian peninsula, southeastern Europe, Cyprus and Turkey are apparent. The data reveal structure even among French-, German- and Italian-speaking groups within Switzerland (Fig. 1b), and between Ireland and the United Kingdom (Fig. 1a, IE and GB). Within some countries individuals are strongly differentiated along the principal component (PC) axes, suggesting that in some cases the resolution of the genetic data may exceed that of the available geographic information.

When we quantitatively compare the geographic position of countries with their PC-based genetic positions, we observe few prominent differences between the two (Supplementary Fig. 1), and those that exist can be explained either by small sample sizes (for example, Slovakia (SK)) or by the coarseness of our geographic data (a problem for large countries, for example, Russia (RU)); see Supplementary Information for more detail. Our method also identifies a few individuals who exhibit large differences between their genetic and geographic positions (Supplementary Fig. 2). These individuals may have mis-specified ancestral origins or be recent migrants. In addition, although the sample used here is unlikely to include many members of smaller genetically isolated populations that exist within countries (for example, Basque residing in Spain or France, Orcadians in Scotland, or individuals of Jewish ancestry), in rare cases outlying individuals could reflect membership of such groups. For example, a small set of Italian individuals cluster 'southwest' of the main Italian cluster and one might speculate they are individuals of insular Italian origin (for example, Sardinia or Sicily).

The overall geographic pattern in Fig. 1a fits the theoretical expectation for models in which genetic similarity decays with distance in a two-dimensional habitat, as opposed to expectations for models involving discrete well-differentiated populations. Indeed, in these data genetic correlation between pairs of individuals tends to decay with distance (Fig. 1c). For spatially structured data, theory predicts the top two principal components (PCs 1 and 2) to be correlated with perpendicular geographic axes[9], which is what we observe ($r^2 = 0.71$ for PC1 versus latitude; $r^2 = 0.72$ for PC2 versus longitude; after rotation, $r^2 = 0.77$ for 'north–south' in PC-space versus latitude, and $r^2 = 0.78$ for 'east–west' in PC-space versus

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

longitude). In contrast, when there are *K* discrete populations sampled, one expects discrete clusters to be separated out along $K - 1$ of the top PCs8. In our analysis, neither the first two PCs, nor subsequent PCs, separate clusters as one would expect for a set of discrete, well-differentiated populations (see ref. [8] for examples).

The direction of the PC1 axis and its relative strength may reflect a special role for this geographic axis in the demographic history of Europeans (as first suggested in ref. 10). PC1 aligns north-northwest/south-southeast (NNW/SSE, −16 degrees) and accounts for approximately twice the amount of variation as PC2 (0.30% versus 0.15%, first eigenvalue = 4.09, second eigenvalue = 2.04). However, caution is required because the direction and relative strength of the PC axes are affected by factors such as the spatial distribution of samples (results not shown, also see ref. 9). More robust evidence for the importance of a roughly NNW/SSE axis in Europe is that, in these same data, haplotype diversity decreases from south to north (A.A. *et al.*, submitted). As the fine-scale spatial structure evident in Fig. 1 suggests, European DNA samples can be very informative about the geographical origins of their donors. Using a multiple-regression-based assignment approach, one can place 50% of individuals within 310 km of their reported origin and 90% within 700 km of their origin (Fig. 2 and Supplementary Table 4, results based on populations with $n > 6$). Across all populations, 50% of individuals are placed within 540 km of their reported origin, and 90% of individuals within 840 km (Supplementary Fig. 3 and Supplementary Table 4). These numbers exclude individuals who reported mixed grandparental ancestry, who are typically assigned to locations between those expected from their grandparental origins (results not shown). Note that distances of assignments from reported origin may be reduced if finer-scale information on origin were available for each individual.

Population structure poses a well-recognized challenge for disease-association studies (for example, refs 11–13). The results obtained here reinforce that the geographic distribution of a sample is important to consider when evaluating genome-wide association studies among Europeans (for example, refs 3–5, 11). A crucial part is also played by spatial variation in phenotype. To examine this, we simulated genome-wide association data for quantitative trait phenotypes with varying degrees of linear latitudinal or longitudinal trends (Supplementary Fig. 4). Even for phenotypes modestly correlated with geography (for example, ≥5% of variance explained by latitude or longitude) the uncorrected *P*-value distribution shows a clear excess of small values, suggesting that population structure correction may be important even in seemingly closely related populations such as Europeans. Note that many factors, including sample size and distribution of sampling locations, will influence the effects of stratification on *P*-value distributions, and so these results should be considered only as illustrative of the settings in which stratification could become a problem in European samples.

In all our simulations, use of a PC-based correction[12,14] adequately controlled for *P*-value inflation (Supplementary Fig. 4). The success of PCA-based correction is not unexpected here, because the PCs are excellent predictors of latitude and longitude, and we used only linear functions of latitude and longitude to determine the means of our simulated phenotypes. For real phenotypes, higher order functions of PC1 and PC2 and/or additional PCs might be necessary to correct for more complex spatial variation in phenotype. We speculate that at the geographic scale of many association studies carried out so far, many phenotypes are relatively uncorrelated with geography, and that this may explain why in many cases PC-based correction has had little impact in practice[3,13]. For phenotypes that are more strongly spatially structured within a sample (for example, height[11,15,16]), spurious associations due to population stratification should be more of a concern.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Although broad correlations between PCs and geography have been observed previously[3–5,17,18] only the large number of loci and dense geographic sampling of individuals used here reveal the clear map-like structure to European genetic variation. Because at any one SNP the average level of differentiation across Europe is small (average $F_{ST} = 0.004$ between geographic regions; F_{ST} is a measure of differentiation between populations that takes values of 0 when there is no differentiation and one when there is maximal differentiation19), it is the combined information across many loci and many individuals that reveals fine-scale population structure in this sample.

An important consideration in interpreting our analyses is that, as a result of ascertainment bias20,21, current SNP genotyping platforms under-represent variation at low-frequency alleles. Low-frequency alleles tend to be the result of a recent mutation and are expected to geographically cluster around the location at which the mutation first arose; hence, they can be highly informative about the fine-scale population structure (for example, ref. 22). In addition, the PCA-based methods used here are based on genotypic patterns of variation and do not take advantage of signatures of population structure that are contained in patterns of haplotype variation1,23–25. Soon-to-be-available whole-genome re-sequencing will give us access to informative low-frequency alleles, and further statistical method development will allow us to leverage patterns of haplotype variation. The prospect of these developments suggests the geographic resolution presented here is only a lower bound on the performance possible in the near future. Thus, our results provide an important insight: the power to detect subtle population structure, and in turn the promise of genetic ancestry tests, may be more substantial than previously imagined.

METHODS SUMMARY

The sample of European individuals used here was assembled and genotyped as part of the larger POPRES project[7]. Genotyping was carried out using the Affymetrix GeneChip Human Mapping 500K Array Set. No significant differentiation was observed between individuals collected and/or genotyped at different times (analysis of variance, ANOVA, $P > 0.05$).

PCA was carried out using the smartpca program8,12. Before running PCA, we removed SNPs that showed evidence of high pairwise linkage disequilibrium as well as unique genomic regions (such as large polymorphic inversions) that might obscure genome-wide patterns of population structure. In addition, an initial PCA run was used to remove extreme genetic outliers.

When comparing the PC results to geography, we assigned each individual a location—typically the geographic centre of their corresponding population (Supplementary Table 3). The rotation of axes used in Fig. 1 is 16 degrees counterclockwise and was determined by finding the angle that maximizes the summed correlation of the median PC1 and PC2 values with the latitude and longitude of each country.

The new assignment method used here is based on independent linear models for latitude and longitude where each is predicted jointly by PC1 and PC2, including quadratic terms and an interaction term. To assess performance, we used leave-one-out cross-validation and adjusted for unequal sample sizes (for example, we weigh each population equally when computing the mean prediction accuracy).

For the genome-wide association simulations, we simulated each individual's phenotype as having a mean determined by his or her geographic position and then simulated Gaussian distributed residual variation to obtain a phenotype with a fixed proportion of variance explained by geographic position. To perform the association test with PC-based correction,

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

we used multiple linear regression with PC1 and PC2 as covariates, as implemented in the program eigenstrat[8,12].

METHODS

Sample collection and genotyping

The samples were assembled and genotyped as part of the larger POPRES project currently consisting of ~6,000 individuals from worldwide populations7. The subsample of European individuals used here is derived from two independent collections: the London Life Sciences Population (LOLIPOP) study[26], which consists mainly of European individuals sampled in London, and (2) the CoLaus study27, which represents a broad set of European individuals sampled from Lausanne, Switzerland. The combined sample contains individuals with origins from across Europe (Supplementary Table 2), although origins from eastern Europe are generally less well represented (for example, Finland, Latvia, Ukraine, Slovakia and Slovenia) and some countries are not sampled at all (for example, Belarus, Estonia, Lithuania and Moldova).

Genotyping was carried out using the Affymetrix GeneChip Human Mapping 500K Array Set according to published protocol. We observe no significant differentiation in the PCA between individuals collected and/or genotyped at different times (ANOVA, $P > 0.05$). A thorough description of the collections, data processing methods and public data release is presented in ref. [7].

To prepare the sample analysed here, we used the demographic data available for each individual to create a 'geographic origin' that represents a single location from which the individual's very recent ancestry is derived. Where possible, we based the geographic origin on the observed country data for grandparents. We used a 'strict consensus' approach: if all observed grandparents originated from a single country, we used that country as the origin. If an individual's observed grandparents originated from different countries, we excluded the individual. Where grandparental data were unavailable, we used the individual's country of birth.

We excluded individuals whose putative geographic origin was from outside of Europe (for example, Europeans from USA, China, Mozambique, Ivory Coast, and so on), individuals who were putatively related (using the same approach as in ref. 7), and individuals found to be outliers in a preliminary PCA run (for more detail, see the section on PCA below). Because of the large number of Swiss individuals available and the availability of language information for most of these individuals, for some analyses, we divided Swiss individuals into three ancestry labels (Swiss-French, Swiss-German and Swiss-Italian) on the basis of their reported primary language. Finally, we chose to include only a random sample of 200 individuals from the United Kingdom and 125 Swiss-French to obtain more even sample sizes across Europe. Supplementary Table 2 provides more detail on how the sample numbers changed with each step in the sample preparation, and Supplementary Table 1 summarizes the number of grandparents observed for the 1,387 individuals used in the final sample.

Geographic locations associated with each country were assigned using the central point of the geographic area of the country (Supplementary Table 3). Three exceptions are the Russian Federation, Sweden and Norway, where the geographic locations were assigned to the location of the capitals of these countries (because central points were assumed to not be as reflective of the probable origins of these individuals). Within Switzerland, we represent the Swiss-French with the geographical coordinates of Geneva, the Swiss-German with

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

those of Zurich, and Swiss-Italian with those of Lugano. Distances between points are always calculated as great circle distances.

For estimating F_{ST}[19] and for assessing the performance of assignment, we combined individuals into geographic groupings with larger and more comparable sample sizes than the original ancestral origins. These groupings do not reflect discrete structure in the data, rather the practical need to create geographical groupings with reasonable sample sizes. The strategy was to create a 3 × 3 grid of regions across Europe, with a tenth region for far southeastern Europe (Supplementary Table 3).

Principal components analysis

To conduct PCA, we used the smartpca software[8,12]. In a preliminary phase of the study, we ran smartpca using default settings and five outlier detection iterations, which resulted in the identification and exclusion of 34 individuals that were greater than six standard deviations from the mean PC position on at least one of the top ten eigenvectors. For our final run, we use the default settings without any outlier removal.

To avoid artefacts due to patterns of linkage disequilibrium[3], we filtered autosomal SNPs using two approaches simultaneously. First, before running PCA we used the PLINK[28] software to exclude SNPs with pairwise genotypic r^2 greater than 80% within sliding windows of 50 SNPs (with a 5-SNP increment between windows). Second, we took an iterative approach by running an initial PCA and removing chromosomal regions that showed evidence of reflecting regions of exceptional long-range linkage disequilibrium rather than genome-wide patterns of structure. These regions are detectable by plotting the correlation between individual PC scores and genotypes against the genome and identifying sharp, concentrated peaks in correlation (alternatively, we could have plotted the magnitude of elements of the SNP-based eigenvectors from the PCA, but here we used the correlation-based approach because much of this work was done before the release of recent versions of smartpca that provide the SNP eigenvectors). SNPs falling within a 4 megabase region of a peak were excluded from the final PCA. Initially, peaks were defined by taking the top 0.01% of SNPs correlating with a PC for each of the top 6 PCs of the preliminary analysis. In this initial analysis PCs 1 and 2 did not appear to be artefacts of long-range linkage disequilibrium, but we still removed regions around the top PC-correlated SNPs. This approach is conservative (in the sense that we potentially remove more SNPs than necessary and hence might hinder ourselves from detecting subtle patterns). The procedure removed SNPs in regions such as the lactase region (2q21), the MHC region and the inversion regions 8p23 and 17q21.31, amongst others. The final number of SNPs used for PCA was 197,146 SNPs. The patterns of structure observed in PCs 1 and 2 were robust to further removal of chromosomal regions correlated with the PCs, suggesting the observed patterns are representative of genome-wide differentiation.

The inter-individual genetic correlations used in Fig. 1c were the same as those used for the PCA analysis and were obtained using the formula of ref. [8] as computed by smartpca.

The angle used to create the rotated PC1–PC2 coordinate system that is used in Fig. 1 was obtained by maximizing θ in the objective function:

$$f(\theta)=\mathrm{Cor}(g(\theta, \mathbf{v}_1, \mathbf{v}_2), \mathbf{Long})+\mathrm{Cor}(h(\theta, \mathbf{v}_1, \mathbf{v}_2), \mathbf{Lat})$$

where $g(\theta, \mathbf{v}_1, \mathbf{v}_2)$ and $h(\theta, \mathbf{v}_1, \mathbf{v}_2)$ are functions that return coordinates of $\mathbf{v}_1$ (the PC1 eigenvector) and $\mathbf{v}_2$ (the PC2 eigenvector) after rotation about the point (0,0) in PC1–PC2 space by the angle θ. **Lat** and **Long** are vectors of the latitude and longitude of each

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

individual, and Cor(·, ·) is the correlation function. The resulting optimal value of θ was found to be −16 degrees.

Spatial assignment

We assigned each individual to a specific geographic location by fitting independent linear models for latitude and longitude as predicted jointly by PC1 and PC2. We used the rotated PC1 and PC2 scores because these more strongly correlate with latitude and longitude (see main text). Specifically, we use the linear models:

$$\mathbf{x}=\beta_{x1}\mathbf{u}_1+\beta_{x2}\mathbf{u}_2+\beta_{x11}\mathbf{u}^2{}_1+\beta_{x22}\mathbf{u}^2{}_2+\beta_{x12}\mathbf{u}_1\mathbf{u}_2+\boldsymbol{\varepsilon}$$

$$\mathbf{y}=\beta_{y1}\mathbf{u}_1+\beta_{y2}\mathbf{u}_2+\beta_{y11}\mathbf{u}^2{}_1+\beta_{y22}\mathbf{u}^2{}_2+\beta_{y12}\mathbf{u}_1\mathbf{u}_2+\boldsymbol{\varepsilon}$$

where **x** and **y** are vectors containing the longitude and latitude, respectively, of each individual, $\mathbf{u}_1$ and $\mathbf{u}_2$ are vectors containing the rotated PC1 and PC2 scores, respectively, for each individual (that is, $\mathbf{u}_1 = g(\theta, \mathbf{v}_1, \mathbf{v}_2)$, $\mathbf{u}_2 = h(\theta, \mathbf{v}_1, \mathbf{v}_2)$, where $\theta = -16$ degrees), β coefficients are regression coefficients, and ε represents residual error.

To perform assignment, we first estimated the β coefficients by means of least-squares regression with a training set of individuals with known locations and then used the estimated coefficients of the linear model to predict the latitude and longitude of a test individual on the basis of their PC1 and PC2 values (we call this a 'continuous assignment'). We also made a 'discrete assignment' by assigning individuals to the country for which the centre-point is closest to the latitude and longitude predicted by the continuous assignment method. In practice, the two methods produce roughly similar results (Supplementary Table 4). As a reference point for evaluating performance, the Supplementary Table also reports statistics for how a method would perform if all individuals were assigned to a central location within Europe (here taken to be Austria).

Simulation of genome-wide association study for a spatially structured quantitative trait

We simulated two types of traits: one with a latitudinal trend in the mean and the other with a longitudinal trend. For each type of trait, we simulated a range of different degrees to which the geographical axis (latitude or longitude) contributed to the overall variance in the trait. Specifically, we let $\mathbf{x}'$ and $\mathbf{y}'$ be normalized latitudinal and longitudinal variables, respectively (that is, $\mathbf{x}' = (\mathbf{x} - \bar{\mathbf{x}})/\sigma_x$ and $\mathbf{y}' = (\mathbf{y} - \bar{\mathbf{y}})\ \sigma_y$, where **x** is a vector of each individual's longitude, **y** is likewise for latitude, $\bar{t}$ is the mean value of the elements of **t**, and σ_t is their standard deviation). We then simulated two phenotypes with the mean determined by $\mathbf{x}'$ or $\mathbf{y}'$: $\boldsymbol{\varphi}_x = \mathbf{x}' + \boldsymbol{\varepsilon}_x$ and $\boldsymbol{\varphi}_y = \mathbf{y}' + \boldsymbol{\varepsilon}_y$, where **ε** is a vector of random normal deviates with mean 0 and variance s^2. We let s^2 take values of (1, 4, 19, 99), so that the resulting variance in the traits are approximately (2, 5, 20, 100), and the proportion of variance explained is approximately (50, 20, 5, 1) per cent.

To perform the association test with PC-based correction, we used multiple linear regression with PC1 and PC2 as covariates as implemented in the software eigenstrat[12]. The Armitage χ^2 statistic was used to test the strength of the association. We also calculate an inflation statistic, by taking the ratio of the 50% quantile of the observed Armitage χ^2 statistic with that expected under the null χ^2_1 distribution.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank J. Kooner and J. Chambers of the LOLIPOP study and G. Waeber, P. Vollenweider, D. Waterworth, J. S. Beckmann, M. Bochud and V. Mooser of the CoLaus study for providing access to their collections. Financial support was provided by the Giorgi-Cavaglieri Foundation (S.B.), the Swiss National Science Foundation (S.B.), US National Science Foundation Postdoctoral Fellowship in Bioinformatics (J.N.), US National Institutes of Health (M.S., C.D.B.) and GlaxoSmithKline (M.R.N.).

References

1. Jakobsson M, et al. Genotype, haplotype and copy-number variation in worldwide human populations. Nature 2008;451:998–1003. [PubMed: 18288195]
2. Li JZ, et al. Worldwide human relationships inferred from genome-wide patterns of variation. Science 2008;319:1100–1104. [PubMed: 18292342]
3. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 2007;447:661–678. [PubMed: 17554300]
4. Tian C, et al. Analysis and application of European genetic substructure using 300K SNP information. PLoS Genet 2008;4:e4. [PubMed: 18208329]
5. Price AL, et al. Discerning the ancestry of European Americans in genetic association studies. PLoS Genet 2008;4:e236. [PubMed: 18208327]
6. Shriver MD, Kittles RA. Genetic ancestry and the search for personalized genetic histories. Nature Rev Genet 2004;5:611–618. [PubMed: 15266343]
7. Nelson MR, et al. The Population Reference Sample (POPRES): a resource for population, disease, and pharmacological genetics research. Am J Hum Genet. (in the press).
8. Patterson N, Price A, Reich D. Population structure and eigenanalysis. PLoS Genet 2006;2:e190. [PubMed: 17194218]
9. Novembre J, Stephens M. Interpreting principal component analyses of spatial population genetic variation. Nature Genet 2008;40:646–649. [PubMed: 18425127]
10. Menozzi P, Piazza A, Cavalli-Sforza L. Synthetic maps of human gene frequencies in Europeans. Science 1978;201:786–792. [PubMed: 356262]
11. Campbell CD, et al. Demonstrating stratification in a European American population. Nature Genet 2005;37:868–872. [PubMed: 16041375]
12. Price AL, et al. Principal components analysis corrects for stratification in genome-wide association studies. Nature Genet 2006;38:904–909. [PubMed: 16862161]
13. McCarthy MI, et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nature Rev Genet 2008;9:356–369. [PubMed: 18398418]
14. Zhu X, Zhang S, Zhao H, Cooper RS. Association mapping, using a mixture model for complex traits. Genet Epidemiol 2002;23:181–196. [PubMed: 12214310]
15. Weedon MN, et al. Genome-wide association analysis identifies 20 loci that influence adult height. Nature Genet 2008;40:575–583. [PubMed: 18391952]
16. Lettre G, et al. Identification of ten loci associated with height highlights new biological pathways in human growth. Nature Genet 2008;40:584–591. [PubMed: 18391950]
17. Cavalli-Sforza, LL.; Menozzi, P.; Piazza, A. The History and Geography of Human Genes. Vol. 292. Princeton Univ. Press; 1994.
18. Bauchet M, et al. Measuring European population stratification with microarray genotype data. Am J Hum Genet 2007;80:948–956. [PubMed: 17436249]
19. Weir BS, Cockerham CC. Estimating F-statistics for the analysis of population structure. Evolution 1984;38:1358–1370.
20. Eberle MA, Kruglyak L. An analysis of strategies for discovery of single nucleotide polymorphisms. Genet Epidemiol 2000;19:S29–S35. [PubMed: 11055367]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

21. 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. [PubMed: 16251459]
22. Slatkin M. Rare alleles as indicators of gene flow. Evolution 1985;39:53–65.
23. Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 2003;164:1567–1587. [PubMed: 12930761]
24. 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. [PubMed: 16773560]
25. Hellenthal G, Auton A, Falush D. Inferring human colonization history using a copying model. PLoS Genet 2008;4:e1000078. [PubMed: 18497854]
26. Kooner J, et al. Genome-wide scan identifies variation in MLXIPL associated with plasma triglycerides. Nature Genet 2008;40:149–151. [PubMed: 18193046]
27. Firmann M, et al. The CoLaus study: A population-based study to investigate the epidemiology and genetic determinants of cardiovascular risk factors and metabolic syndrome. BMC Cardiovasc Dis 2008;8:6.
28. Purcell S, et al. PLINK: A tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007;81:559–575. [PubMed: 17701901]

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

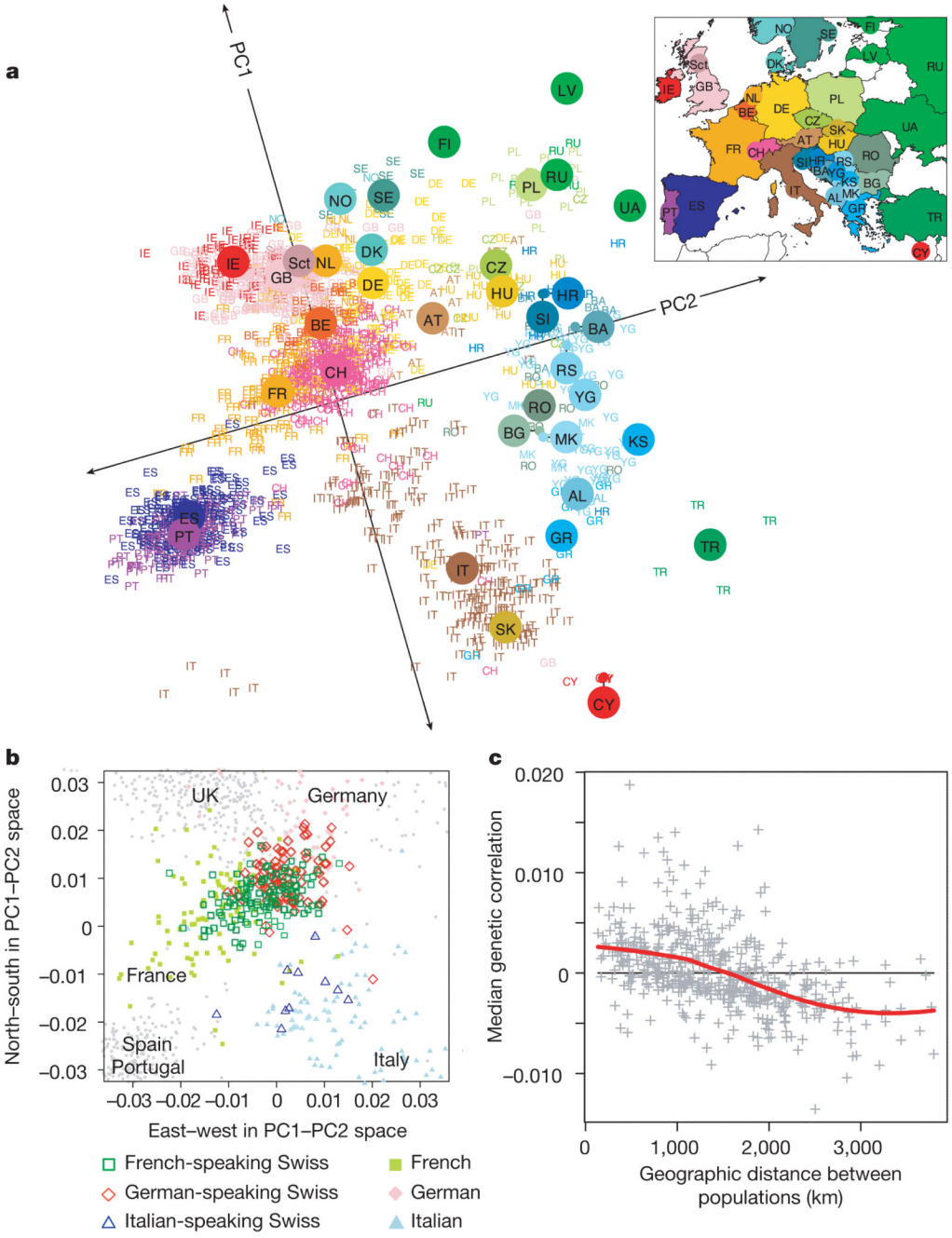

Figure 1. Population structure within Europe

a, A statistical summary of genetic data from 1,387 Europeans based on principal component axis one (PC1) and axis two (PC2). Small coloured labels represent individuals and large coloured points represent median PC1 and PC2 values for each country. The inset map provides a key to the labels. The PC axes are rotated to emphasize the similarity to the geographic map of Europe. AL, Albania; AT, Austria; BA, Bosnia-Herzegovina; BE, Belgium; BG, Bulgaria; CH, Switzerland; CY, Cyprus; CZ, Czech Republic; DE, Germany; DK, Denmark; ES, Spain; FI, Finland; FR, France; GB, United Kingdom; GR, Greece; HR, Croatia; HU, Hungary; IE, Ireland; IT, Italy; KS, Kosovo; LV, Latvia; MK, Macedonia; NO, Norway; NL, Netherlands; PL, Poland; PT, Portugal; RO, Romania; RS, Serbia and

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Montenegro; RU, Russia, Sct, Scotland; SE, Sweden; SI, Slovenia; SK, Slovakia; TR, Turkey; UA, Ukraine; YG, Yugoslavia. **b**, A magnification of the area around Switzerland from **a** showing differentiation within Switzerland by language. **c**, Genetic similarity versus geographic distance. Median genetic correlation between pairs of individuals as a function of geographic distance between their respective populations.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

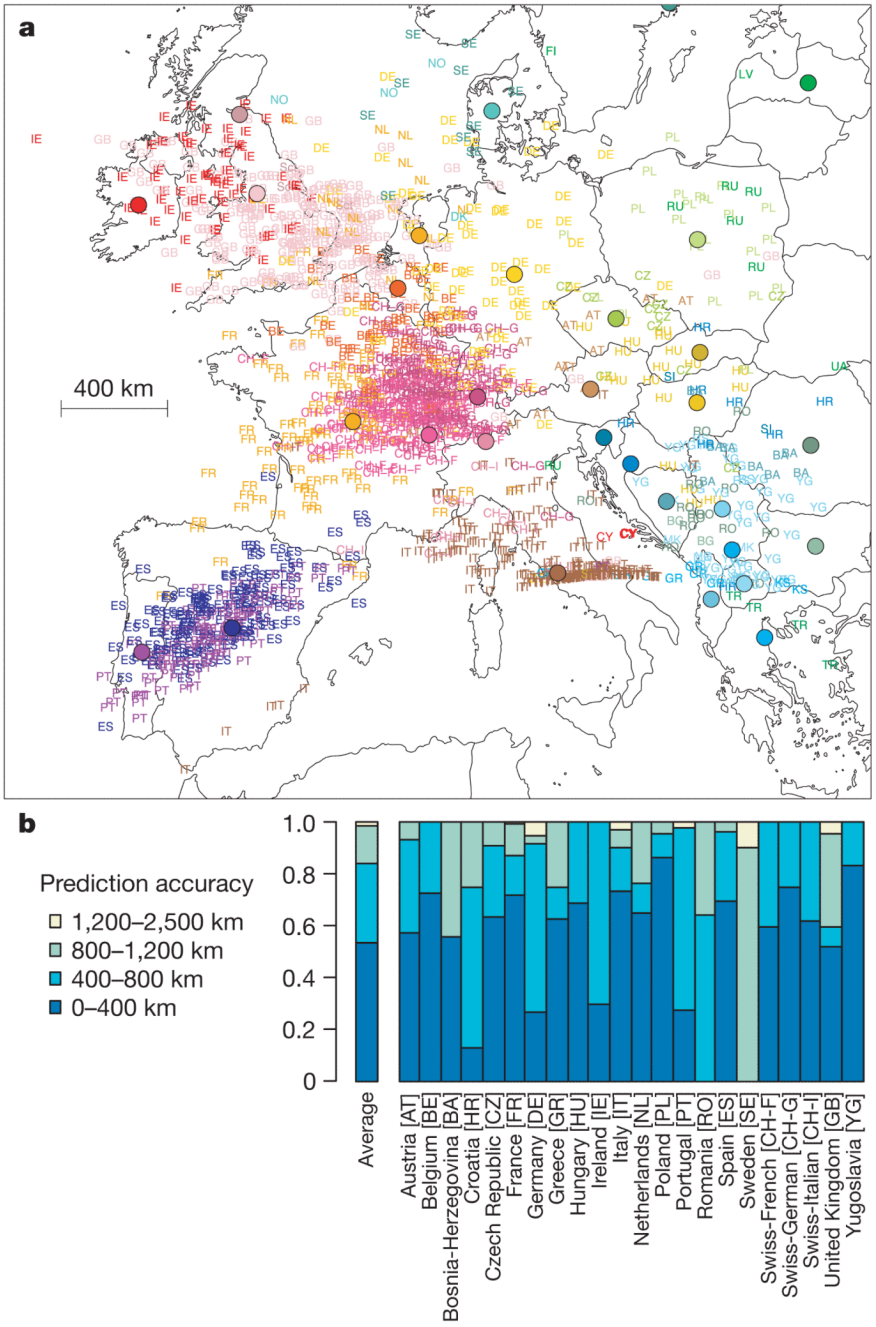

Figure 2. Performance of assignment method

a, Predicted locations for each of 1,387 individuals based on leave-one-out cross validation and the continuous assignment method. Small coloured labels (for definitions, see Fig. 1 legend, except here CH-I, CH-F, and CH-G denote Swiss individuals who speak Italian, French, or German respectively) represent individual assignments. Coloured points denote the locations used to train the assignment method. **b**, Distribution of prediction accuracy by country. Distances are measured between the population assigned by the discrete assignment method and the geographic origin of the individual. The average is taken of the proportions across populations and each population is given equal weight. The panel shows results for

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

populations with greater than six individuals; performance decreases for populations with smaller sample sizes (Supplementary Fig. 3).