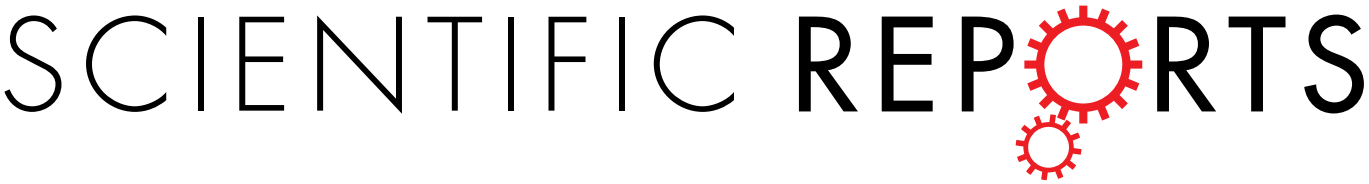

OPEN

Received: 04 August 2016
Accepted: 22 December 2016
Published: 01 February 2017

Overcoming the dichotomy between open and isolated populations using genomic data from a large European dataset

Paolo Anagnostou[1,2,*], Valentina Dominici[1,*], Cinzia Battaggia[1], Luca Pagani[3,4], Miguel Vilar[5,6], R. Spencer Wells[7], Davide Pettener[4], Stefania Sarno[4], Alessio Boattini[4], Paolo Francalacci[8], Vincenza Colonna[9], Giuseppe Vona[10], Carla Calò[10], Giovanni Destro Bisol[1,2,*] & Sergio Tofanelli[11,*]

Human populations are often dichotomized into "isolated" and "open" categories using cultural and/or geographical barriers to gene flow as differential criteria. Although widespread, the use of these alternative categories could obscure further heterogeneity due to inter-population differences in effective size, growth rate, and timing or amount of gene flow. We compared intra and inter-population variation measures combining novel and literature data relative to 87,818 autosomal SNPs in 14 open populations and 10 geographic and/or linguistic European isolates. Patterns of intra-population diversity were found to vary considerably more among isolates, probably due to differential levels of drift and inbreeding. The relatively large effective size estimated for some population isolates challenges the generalized view that they originate from small founding groups. Principal component scores based on measures of intra-population variation of isolated and open populations were found to be distributed along a continuum, with an area of intersection between the two groups. Patterns of inter-population diversity were even closer, as we were able to detect some differences between population groups only for a few multidimensional scaling dimensions. Therefore, different lines of evidence suggest that dichotomizing human populations into open and isolated groups fails to capture the actual relations among their genomic features.

Human groups that have been subject to geographical and/or socio-cultural barriers (e.g. linguistic, social or religious) to inward gene flow during their evolutionary history are commonly referred to as isolated or closed populations (hereafter "isolates/isolated"). In current genetic literature, they are often opposed to open or outbred populations - exempt from known limitations to admixture - since a higher level of inbreeding and drift and a lower efficiency of recombination in redistributing variation across individuals are to be expected under isolation[1–3]. Although common practice, the use of the terms *open* and *isolated* to indicate two discrete dichotomous categories could obscure the existence of further heterogeneity. In fact, when applying such a distinction to genetics based on environmental and socio-cultural factors, we implicitly assume that it is not confounded

[1]Dipartimento di Biologia Ambientale, Sapienza Università di Roma, Piazzale Aldo Moro 5, Rome, 00185, Italy. [2]Istituto Italiano di Antropologia, Piazzale Aldo Moro 5, Rome, 00185, Italy. [3]Estonian Biocentre, Riia 23b, 51010, Tartu, Estonia. [4]Dipartimento di Scienze Biologiche, Geologiche ed Ambientali, Università di Bologna, Via Selmi 3, Bologna, 40126, Italy. [5]Department of Anthropology, University of Pennsylvania, 3260 South St, Philadelphia, Pennsylvania, United States of America. [6]National Geographic Society, 1145 17th Street NW, Washington DC 20036, United States of America. [7]Department of Integrative Biology, University of Texas at Austin, Austin, TX 78712, USA. [8]Dipartimento di Scienze della Natura del Territorio, Università di Sassari, Via Piandanna 4, Sassari, 07100, Italy. [9]Institute of Genetics and Biophysics "A. Buzzati-Traverso", National Research Council (CNR), Via Pietro Castellino, 111, Naples,80131, Italy. [10]Dipartimento di Scienze della Vita e dell'Ambiente, Università di Cagliari, SS 554, km 4.500, Monserrato, Cagliari,09042, Italy. [11]Dipartimento di Biologia, Università di Pisa, Via Ghini 13, Pisa, 56126, Italy. *These authors contributed equally to this work. Correspondence and requests for materials should be addressed to P.A. (email: paolo.anagnostou@uniroma1.it) or G.D.B. (email: destrobisol@uniroma1.it)

by inter-population differences in effective size, growth rate and timing or extent of gene flow reduction. Unfortunately, current knowledge on human isolates cannot help us understand whether this is consistent with the patterning of genetic diversity.

A number of studies has investigated the effects of isolation in human populations by exploiting the high sensitivity to drift of unilinear markers of mtDNA and the non-recombining portion of the Y chromosome[4–9]. However, the lack of recombination limits the power of these genetic systems in the detection of signatures of genetic isolation in different historical and demographic conditions.

With the introduction of SNP microarrays, which enable the simultaneous analysis of hundreds of thousands of loci distributed across the genome, it is now possible to investigate the genetic structure of human populations on a fine scale[10]. Using autosomal variation, we can detect signatures of isolation which are not revealed by unilinear markers, such as the increase in the number and size of stretches of consecutive homozygous genotypes, shared chromosomal segments identical by descent (IBD) and Linkage Disequilibrium (LD)[2,3,11–14]. Investigations published so far have provided accurate genetic characterizations of a number of human genetic isolates, with a prevalent focus on one or few populations and their potential use in gene-disease association studies[15–19]. Relations between genomic differences and demographic or historical factors and their implications for the gene mapping of Mendelian or complex traits have also been studied[1–3], while LD patterns have been compared in isolates distributed worldwide[14]. More recently, other studies have simultaneously investigated multiple isolates, mostly focusing on populations with shared historical and demographic features[17,18,20]. However, to the best of our knowledge, no study has systematically explored the structure of genomic diversity in isolated populations comparing them with a comprehensive set of open populations. The European continent provides optimal conditions for these investigations. There is, in fact, broad convergence regarding the notion that European genomic diversity has been shaped primarily by geography, with the isolation-by-distance model being well supported even at long latitudinal distances[21–23]. Therefore, the comparative study of open and isolated populations may be performed in wider transects with less confounding factors than in other continental areas.

Here we present a study of 24 European populations, nine of which were newly genotyped using the GenoChip 2.0 array[24]. We compare the distribution of intra- and inter-population measures of variation in isolated and open populations in order to understand to what extent the discrete open and isolated dichotomous categories correspond to the way in which their genomic diversity is structured.

Results

In this study, we analysed a dataset including ten linguistically and/or geographically isolated populations, which differ substantially in their census sizes, as well as 14 open populations (see Fig. 1 and Table 1). The isolate group is composed of four German-speaking islands of the Eastern Italian Alps (Sauris, Sappada, Timau and Cimbrians), four Sardinian populations (a sample from North Sardinia, Benetutti, Sulcis Iglesiente and Carloforte), and two well known European groups (Basques and Orcadians). The occurrence of geographical and/or linguistic isolation for all the above populations is supported by historical sources. Furthermore, it should be also noted that members of our linguistic isolates speak a language, and not just a dialect, which is different from that of neighbouring populations[25]. Furthermore, our geographic isolates are settled either at an altitude greater than 610 m above sea level (ie. the lower limit of a mountain range[26]), or in an island where human mobility to and from the mainland is limited due to physical distance and/or adverse sea and weather conditions (see supplementary text S1 for further details).

Regarding the selection of open populations, we considered the following three criteria: (i) geographic proximity with the isolated population dataset; (ii) geographic coverage of the European continent; (iii) sample size of at least 15 individuals.

Genomic variation in open and isolated European populations. In order to better understand how genomic diversity is structured in open and isolated populations, we first analysed the distribution of seven intra-population measures. Three of them are based on variation at a single nucleotide: homozygosity, inter-locus variance between all pairs of loci and intra-population pairwise identities by state (IBS pairwise identities). The remaining four are based on haplotype variation: average number and total length of runs of homozygosity (RoHs), average intra-population sharing of blocks identical by descent between individuals (IBD blocks) and average length of blocks in linkage disequilibrium between pairs of individuals (LD blocks). As expected, the median values of these measures were significantly higher in isolates (alpha = 0.05, Mann-Whitney U test), the length of LD blocks being the only exception (Fig. 2). These results were robust to Bonferroni's correction for multiple testing (alpha = 0.007). However, an overlap between the two groups was observed for six out of seven measures. The most evident one was shown by the LD blocks, in which seven isolates fall within the range of open populations. A clear-cut distinction between the two groups was provided by the IBS pairwise identities only. Similarly, variance between populations was higher among isolates for six measures, and the difference were found to be statistically significant for four of them (Levene Test, Fig. 2). The largest one was observed for the IBD blocks, whose standard deviation was 28 times higher in isolated populations. We also investigated the distribution of RoHs in more detail, because their length and number have been shown to vary between open and isolated populations[27–29]. Although isolated populations had a significantly higher number of RoHs, an overlap between the two groups was observed for all size classes (Supplementary Figure S2).

Moving on from groups to single populations, we observed the strongest signals of isolation in Sauris and Sappada (Supplementary Figures S3 and S4). Due to its small sample size (N = 10; see Supplementary Table S1), the evidence for the former population was tested with resampling procedures, obtaining consistent results. These two populations also have the highest proportion of long RoHs (classes 5 and 6), whereas Basques and Benetutti prevail for the small and medium ones (classes 1 to 3) (Supplementary Figure S5). By contrast, the weakest signal of isolation is provided by the Cimbrians, who show the lowest values for all measures. Inter-individual variation

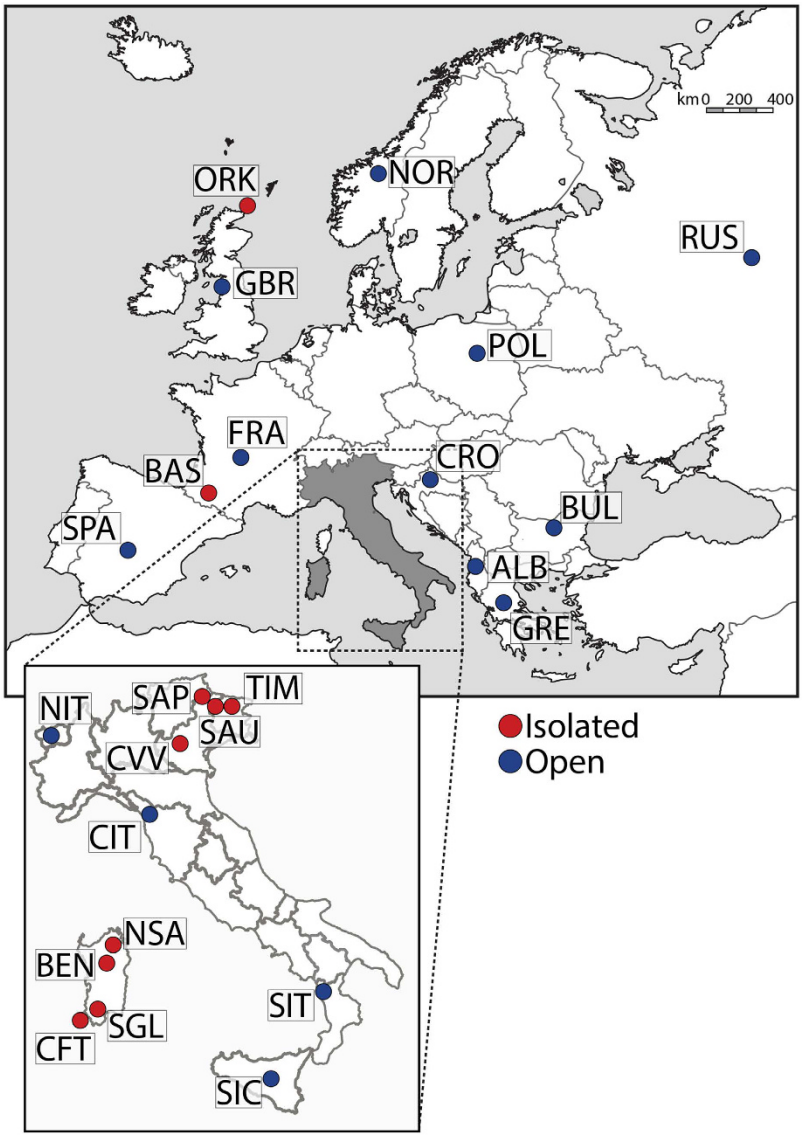

Figure 1. Map showing the geographic location of the 24 populations under study. Labels as in Table 1. Maps available from Wikipedia Common web page (https://commons.wikimedia.org/wiki/File:Blank_political_map_Europe_in_2006_WF.svg?uselang=it#filelinks) were modified using Adobe Photoshop CS6 software.

again reached the highest values in Sauris, Sappada and Timau, whose values were found to be significantly higher in at least 70% of the pairwise comparisons with open populations (see Supplementary Tables S2–S6). A less intense but noticeable signal was observed for North Sardinia, Benetutti and the Basques, which are the only remaining populations with a proportion of significant pairwise comparisons above 50%.

We combined all the measures of intra-population variation using a PCA in order to rank populations according to their degree of isolation (Fig. 3A). All variables heavily load on the first component, which describes 69.7% of the total variance, with the highest contributions by RoHs (total number and length; proportion of medium and large RoHs), IBS identities and Homozygosity (see Fig. 3B). Overall, the first component separates isolates (on the left) from open populations (on the right), with Cimbrians being the only exception. The German-speaking island of Sauris and the Bulgarians are found at the two extremes of the distribution. The second principal component, which describes 17.2% of the total variance, does not set open and isolated populations apart, although the former are more tightly clustered. Sauris (at the upper side), Basques, and Benetutti (at the lower side) are found at the poles of the distribution. Among the factors that contribute most to the positive scores are the average values for the number of very long RoHs (class 6), LD and IBD blocks, whereas the average number of small and intermediate RoHs (class 1 to 3) load on negative scores. When using more relaxed settings for the RoH identification (minimum number of SNPs = 12), no substantial difference was observed for the population distribution of number and total size of RoHs, proportion of the RoHs classes and the resulting PCA plot (Supplementary Table S7).

Population	Label	N	Current Census	Time Since Isolation (In Years Before Present)	Isolation Factor	Reference
North Eastern Italian isolates						
Cimbrians	CVV	33	13,455	~1000	G/L	Present Study
Sappada	SAP	24	1,307	~1000	G/L	Present Study
Sauris	SAU	10	429	~800	G/L	Present Study
Timau	TIM	24	500	800–1000	G/L	Present Study
Sardinian isolates						
Benetutti	BEN	25	1,971	~5000	G/L	Present Study
Carloforte	CFT	25	6,301	268	G/L	Present Study
North Sardinia	NSA	25	96,448	3900–2900	G/L	Present Study
Sulcis Iglesiente	SGL	23	128,540	2800	G/L	Present Study
European isolates						
Orkney	ORK	15	21,349	~1300	G	64
French Basques	BAS	24	~650,000[65]	5500–3500	L	64
South Europe						
Albania (Gheg)	ALB	24	2,831,741	—	—	Sarno *et al.* in preparation
Croatia	CRO	20	4,284,889	—	—	59
Greece	GRE	20	10,815,197	—	—	66
Spain	SPA	34	46,815,916	—	—	66
East Europe						
Bulgaria	BUL	31	7,202,198	—	—	66
Poland	POL	32	38,511,824	—	—	66
Russia	RUS	25	144,192,450	—	—	64
North Europe						
Norway	NOR	18	5,214,890	—	—	66
British isles	GBR	16	63,181,775	—	—	66
West Europe						
France	FRA	28	67,264,000	—	—	64
Italy						
North Italy (Aosta)	NIT	22	34,619	—	—	Present Study
Central Italy (Piana di Lucca)	CIT	25	394,318	—	—	Tofanelli *et al.* unpub. data
South Italy	SIT	18	14,184,916	—	—	66
Sicily	SIC	20	5,077,487	—	—	66

Table 1. Details on populations under study. N stands for sample size, G for Geographic and G/L for geo/linguistic. References for census size and time since isolation can be found in the Supplementary Text S1. Census sizes were obtained from the National population and housing census – 2011 (ALB, BEN, CIT, CFT, CRO, CVV, GBR, GRE, NIT, NSA, ORK, POL, SAP, SAU, SGL, SIC, SIT, SPA, TIM) – 2014 (BUL) – 2015 (RUS, NOR) – 2016 (FRA).

Effective population size in open and isolated European populations. Among isolated populations, the estimated values of effective population size (Ne) range between 209 (Sauris, 208–210; 95% confidence interval) and 3739 (Basque, 3607–3880; 95% confidence interval; Fig. 4, Supplementary Table S8). Regarding open populations, the values range between 2386 (Albania, 2342–2452; 95% confidence interval) and 8267 (Poland, 7850–8732; 95% confidence interval). Interestingly, seven open (Albania, Croatia, Greece, Aosta, Sicily, South Italy and Norway) and five isolated populations (Basques, Benetutti, Carloforte, North Sardinia, and Cimbrians) fall into the range of the alternative group. When repeating the estimates in the four populations for which SNP data at a higher resolution were available (HGDP panel; 647,789 SNPs) using an IBD sharing based method[30], we obtained values that were different in absolute terms but highly correlated with those produced by GenoChip 2.0 (see Supplementary Table S9).

Figure 5 displays a range of possible combinations of N_e and time since isolation (coloured areas) able to produce the inbreeding coefficients observed in four isolated populations (see Materials and Methods), with an indication of time since their foundation as suggested from historical and genetic sources (Supplementary Text S1). At any given N_e, the values for time since isolation relative to Basques and North Sardinians were found to be lower than in Sappada and Sauris. The ratio ranged from approximately two to three times lower when compared to Sappada and four to eight times higher when compared to Sauris (Supplementary Table 10).

Population isolates in the European genomic background. Having described variation within populations, we next concentrated on their genetic relationships. We first explored the distance matrix based on inter-individual pairwise IBS distances. When sorting populations according to their average genetic distance, the

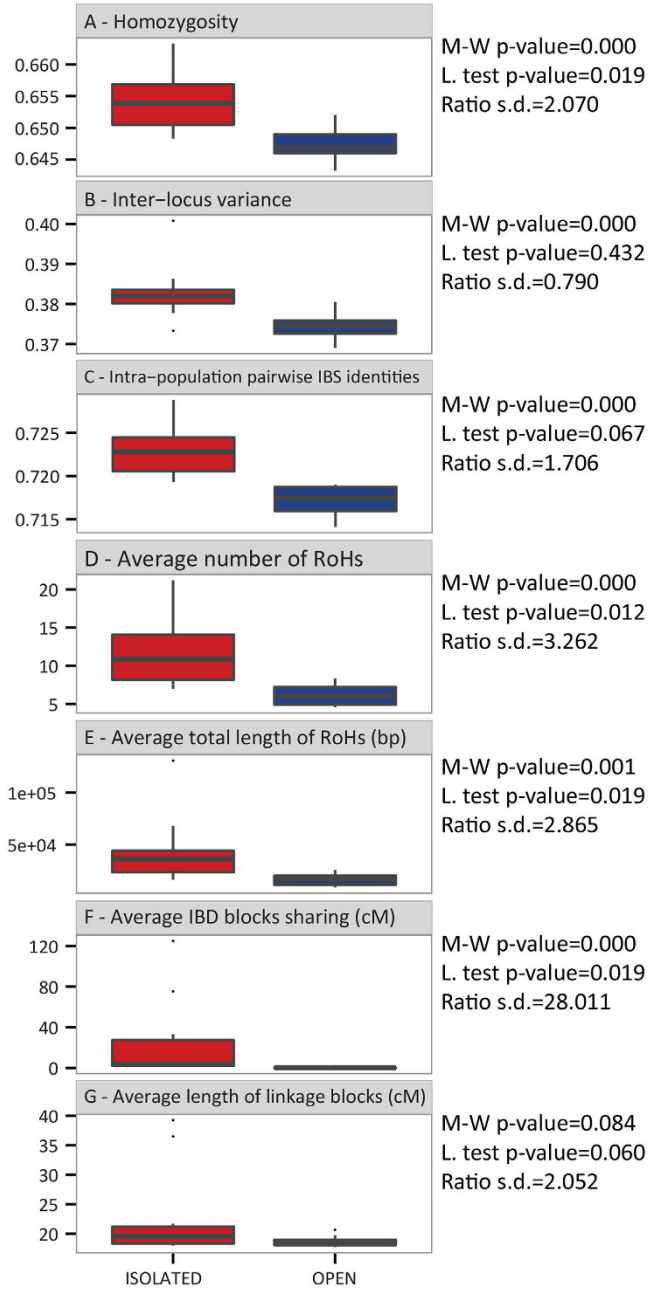

Figure 2. Boxplots of (**A**) Average homozygosity over loci (**B**) Inter-locus variance (**C**) Average number of RoHs (**D**) Average total length of RoHs (**E**) Average intra-population pairwise IBS (**F**) Average population IBD blocks sharing (**G**) Average length of linkage blocks. M-W stands for Mann-Whitney U test, L. for Levene test and s.d. for standard deviation (the tests of the last two statistics were performed excluding the outlier values).

highest value was found for Russians followed by Sauris, South Italy and Sicily (Fig. 6A). At the opposite end of the spectrum, the lowest values were observed for French, Basques, Lessinia Cimbrians and Aosta. Interestingly, the lowest pairwise genetic distances were recorded for the three mainland Sardinian populations, whereas high levels of differentiation were found among the German-speaking islands. Overall, we observed a significant correlation between the genetic and geographic distances considering the dataset both with ($R2 = 0.209$; p-value < 0.05) and without ($R2 = 0.203$; p-value < 0.05) the isolated populations (supplementary Figure S6). Even when applying a correction for the isolation-by-distance effect, patterns of open and isolated populations were found to be very similar (Fig. 6B).

In order to capture more subtle signals of differentiation, we carried out a multidimensional scaling analysis (MDS)[31]. The first dimension clearly separates the three populations of mainland Sardinia (Benetutti, North Sardinia and Sulcis Iglesiente) from the others, whereas the second dimension distinguishes the two German-speaking islands of Sappada and Sauris (Fig. 7A). The third and fourth dimensions (Fig. 7B) separate the

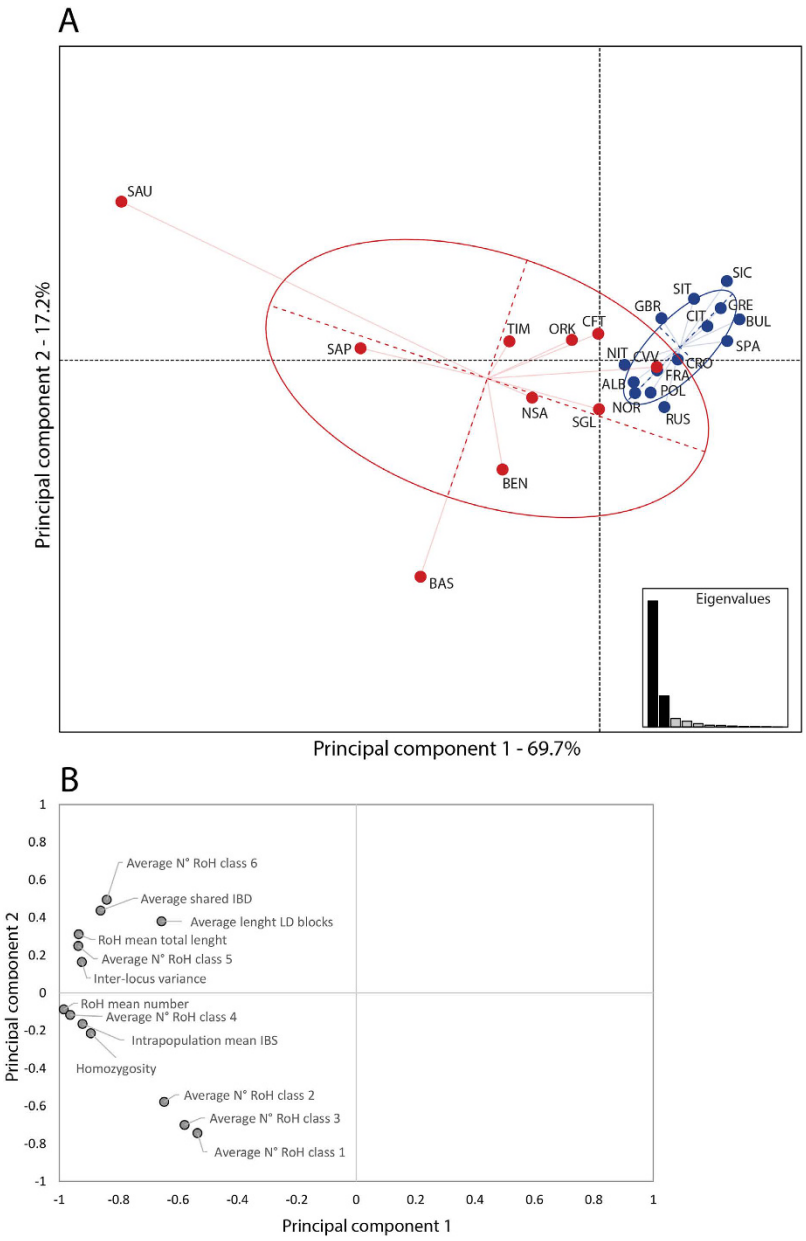

Figure 3. Principal component plots based on intra-population measures. (**A**) Scatter plot of the first two principal components. (**B**) Plot of the factor scores for the first and second principal components. Labels as in Table 1.

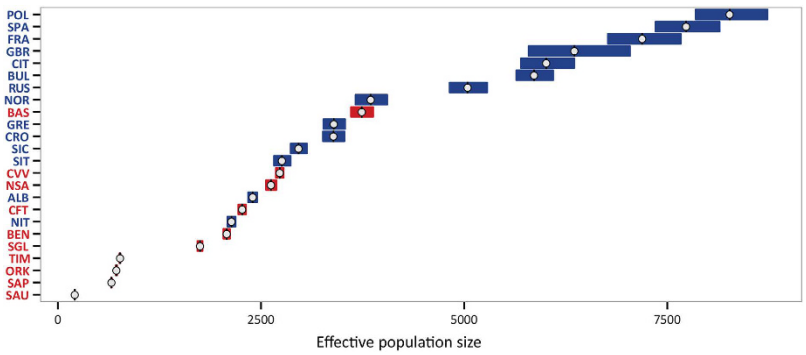

Figure 4. Effective population size estimates based on 68,205 SNPs (16 chromosomes). White circles and bars represent point estimates and 95% confidence interval, respectively. Abbreviations as in Table 1.

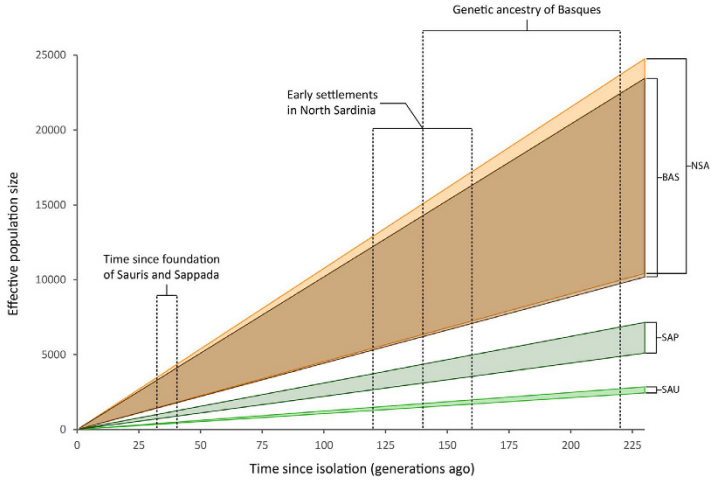

Figure 5. Numbers of generations since isolation (X axis) and corresponding N_e values (Y axis) under a model of constant population size in populations which retain clear signatures of isolation (Basques, North Sardinia, Sappada and Sauris; see also PC1 in Fig. 3). For each population at any given time since isolation, the upper and lower boundaries of Ne were obtained assuming the initial inbreeding coefficients to be equal to the highest and lowest values observed among open populations, respectively. References for time since isolation (indicated by arrows) are reported in Supplementary Text S1.

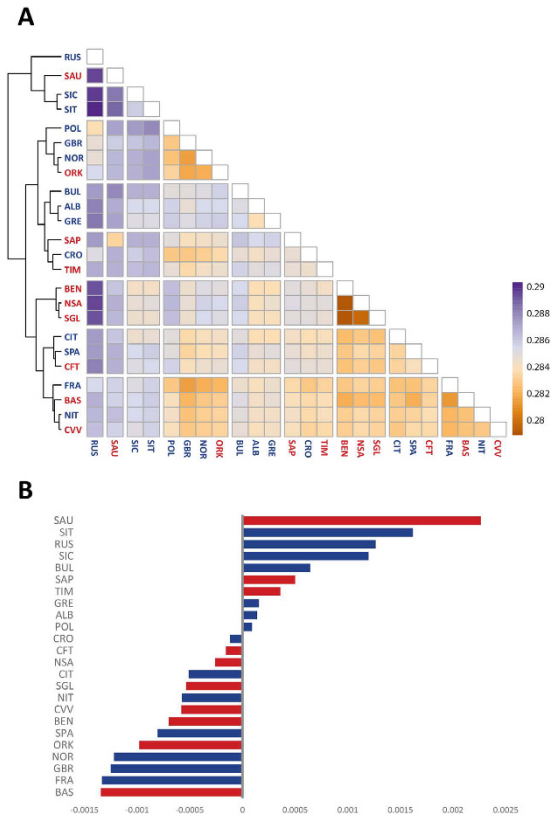

Figure 6. (**A**) Heatmap of pairwise genetic distances (R package Pheatmap). Populations are clustered according to a complete hierarchical approach (**B**) Deviation of the average genetic distances from those predicted by an isolation by distance model in open populations (see Materials and Methods for more details).

Basques (dimension 3) and Timau (dimension 4) respectively. These results mirror the PCA performed directly on SNP data (see Supplementary Figure S7). Furthermore, the best predictive model of ADMIXTURE (four ancestral populations; Supplementary Figure S8) reveals a pattern of population differentiation which is very close to that depicted by MDS.

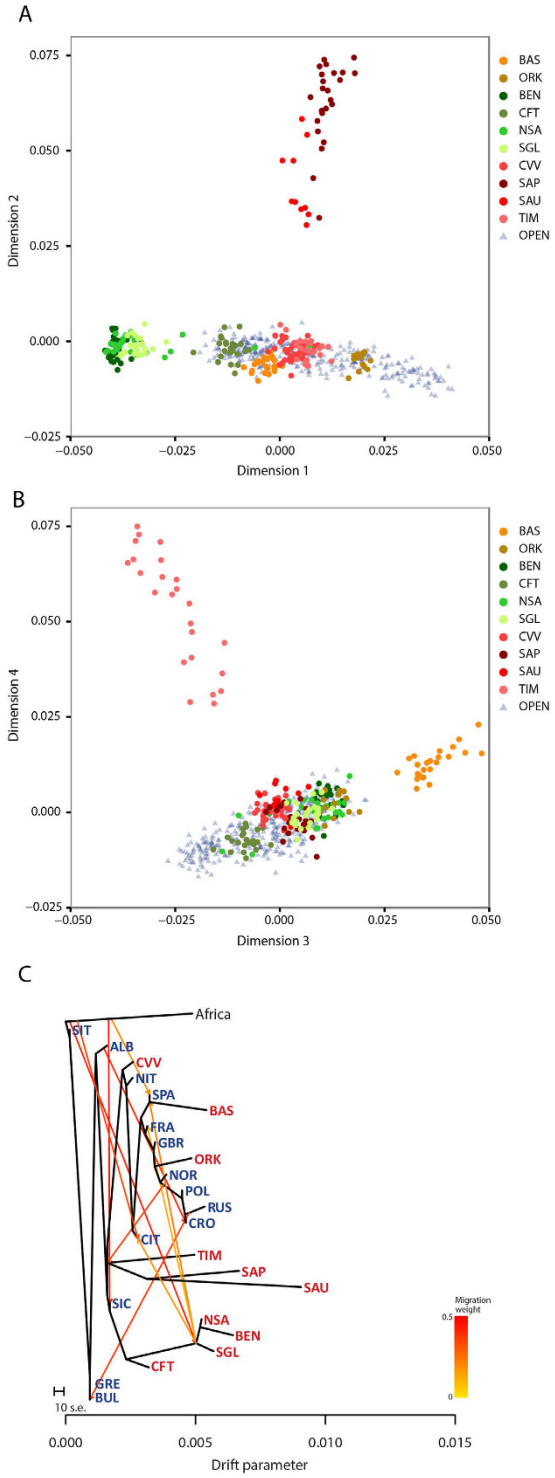

Figure 7. Plot of the first and second dimensions of the Multidimensional scaling analysis (**A**). Plot of the third and fourth dimensions (**B**). Treemix analysis with ten mixture events, with migration arrows coloured according to their weight (**C**).

Finally, we explored patterns of population split and mixture using Treemix. The best fit with the data was obtained for the tree model with the maximum number of migration events tested, i.e. five (f = 0.961). As shown in Fig. 7C, the general structure of the tree largely reflects the well-known relationships among European populations[32]. A high level of genetic drift can be observed for the entire Sardinian branch, and, even more pronounced for single German-speaking islands of Sappada, Sauris and Timau. Interestingly, the Cimbrians from Lessinia are more closely related to the Northern Italians and are located in the tree upstream to all northern and western European populations. Finally, although drifted, Basques and Orcadians cluster on a geographic basis with Spain

and close to British and Norwegians, respectively. Evidence of inward migration events involving our isolated populations was limited to the root of mainland Sardinians (from Africa), Sauris, Sappada and Timau (from a population ancestral to the Polish and Norwegians) and Basques (from the ancestral population of Sardinians).

Discussion

New insights into the genomic diversity of isolated populations. All measures of intra-population genomic variation and the multivariate analysis (as visualized by the PCA plot) highlighted a relative homogeneity among European open populations, a finding which is in accordance with previous investigations[27,29,33]. However, the structure of genomic diversity was found to vary considerably among populations that have been subject to geographic and/or linguistic isolation. While such heterogeneity is consistent with what has been shown in gene-disease association studies[1] and LD patterns[14], our results may help shed more light on its extent and likely causes. We observed a greater dispersion of isolated groups compared to open populations along the first principal component, the variance of scores being 15.8 times higher for the former group. The scores were also found to be highly and significantly correlated with the inbreeding coefficient and drift parameter in the entire population dataset (total $R2 = 0.901$; $p < 0.01$; see Supplementary Table S11 for further details). Although with a lower ratio (6.1), variation among isolates was also higher for the second principal component scores, which was found to be significantly correlated with effective size of isolates ($R2 = 0.620$, p-value $= 0.007$). These results prompt a discussion of three different points.

Firstly, the principal component analysis helped disentangle the effects of the different forces that have shaped the genome of isolates. In fact, the analysis seems to indicate that most of their heterogeneity reflects variable intensities of drift and inbreeding, rather than their size or the time since isolation as suggested by historical sources. Taking the score of the first principal component as a means to rank populations according to their degree of isolation, Sauris, Sappada and Basques were found to be the most isolated, while Orkney, Carloforte and Sulcis Iglesiente were the least, with Timau, Benetutti and North Sardinia in between.

Secondly, the fact we found populations with low scores for both the first and the second principal component - which means a combination of signatures of isolation with a relatively large effective size - calls into question the widespread view that human genetic isolates originate from a small group of founders[1–3,34]. The need for more complex models was earlier recognized by James V. Neel[35], who proposed a categorization of isolates in which he included populations that originated from relatively large groups of individuals. Our study provides evidence that this idea is worth being further developed. In fact, while the estimates of current effective size for Sappada and Sauris seem not to contradict the idea of a small founding group, the values obtained for Basques and North Sardinia overlap with those estimated for a number of open populations. Whether or not this reflects a substantial difference in their founding population size should be considered with caution since our accuracy in estimating changes of effective population size over time[30] was limited by the low SNP density of our genotyping platform. However, it should be noted that our study provides further indirect support to the view that population isolates may largely differ in the size of their founding groups[35]. In fact, when assuming demographic stationarity and equal effective population size among populations, the number of generations needed to reach the observed values of inbreeding coefficients was found to be substantially smaller for Basques and North Sardinians than for Sauris and Sappada (Supplementary Table S10 and Fig. 5). The evident discrepancy between these results and available knowledge about time since foundation of these isolates suggests that one or both assumptions are untenable. Thirdly, our results suggest two quite distinct patterns of local isolation. In the case of the German-speaking islands, signals of heterogeneity among populations seem to prevail. Sappada, Sauris and Timau were found to be clearly different from each other both regarding intra and inter-population diversity. High genetic distances among Sauris, Sappada and Timau have already been observed with unilinear markers[9], a pattern that is probably associated with the occurrence of a form of social behaviour which we termed "local ethnicity". Despite their closely related languages and shared traditions, members of Alpine linguistic islands tend to identify their ancestry with their own village rather than considering themselves as part of the same ethnic group[9]. Such strong territoriality when defining ethnic identities and boundaries may have played a role in marriage strategies, decreasing the genetic exchange among the three linguistic islands. This "isolation among isolates" might have also led to the genomic structure of each of them evolving independently. On the other hand, a much greater homogeneity was observed among mainland Sardinians. The genetic distances among Benetutti, North Sardinia and Sulcis Iglesiente are the lowest in our dataset, and even lower than predicted by their geographic distances (Fig. 6B). This is not surprising because a similarity across the island has already been highlighted in previous studies[36–38] (but see refs 39,40). Therefore, despite the much longer time since isolation compared to German speaking islands, Sardinian populations seem to have maintained a certain homogeneity due to their larger effective population size which, in turn, could have weakened the effects of genetic drift and inbreeding. This could account for their lower variation of intra-population diversity measures, evidenced in the first principal component, compared to that observed among Sappada, Sauris and Timau.

Continuity rather than dichotomy. Our analysis highlights a continuous pattern of genomic variation among populations that has been categorized as open and isolated. Looking at the first principal component, it is possible to identify a denser area, which corresponds to the high homogeneity of open populations, and another sparser zone, which reflects the greater diversity among isolates. In the contact zone, we found Cimbrians, who cluster along the first principal component with the open populations, while Carloforte and Sulcis Iglesiente are borderline. The behaviour of these three populations, all of which have been subject to both geographic and linguistic barriers in the course of their recent history, does not mean that their genomic structure is only marginally different from that of open populations. Rather, it points to the lack of any clear discontinuities between the two groups when multiple indicators of isolation are used simultaneously. In fact, these three populations show signatures of past inbreeding which were undetected in the open ones. This is effectively evidenced by their long upper

tails of pairwise intra-population IBS and IBD block sharing (Supplementary Figure S4) and, in a more irregular fashion, by other measures of intra-population variation. A breakdown of the cultural barrier might account for the behavior of Cimbrians. In fact, only a limited number of individuals is today able to use the Cimbrian language[41], a situation in contrast with the persistence of the original linguistic features in other German speaking communities[42]. This form of cultural assimilation, which started in the middle of the 16th century[43], probably increased the permeability of Cimbrians to gene flow from neighbouring populations. Carloforte is the most recent isolate of our dataset, with the founding event dating back to 1738 AD. The small time since isolation and the genetic introgression associated with migratory waves from Tunisia, Liguria and Campania have presumably limited the effect of inbreeding and drift on the genome[44,45]. The attenuation of isolation signatures for Sulcis Iglesiente compared to other Sardinian populations may be explained by the more exogamous behaviour of the villages in this area, a likely consequence of their location in coastal plains close to the Mediterranean Sea[46].

At inter-population level, the picture obtained by using the genetic distance matrix directly does not discriminate between open and isolated populations. Previous studies revealed that diversity among European populations complies with an isolation by distance model on a continental[21,22] and local scale[47,48]. We were able to find the same pattern over a wide continental range, regardless of the presence of isolates in the dataset, implying that they do not depart significantly from what is to be expected under isolation by distance. The only way to pinpoint a difference for some isolates was by considering specific MDS dimensions, which highlight a more pronounced scattering among individuals from Sauris, Sappada Timau and the Basques. Interestingly, these are also the populations in which we noticed the highest levels of inter-individual variation.

The overall picture provided by our study contrasts with previous observations based on unilinear markers. Using a dataset including all the isolates studied here, with the Orkney Islands being the only exception, we showed that most of them behave as outliers with both mtDNA (hypervariable region 1) and Y chromosome markers (six microsatellite loci)[49]. This discrepancy between genetic systems may be explained by the smaller effective size and the higher mutation rate of unilinear markers. The former feature makes variation of mtDNA and Y chromosome more prone to the effects of genetic drift, while the latter means they can be hit by mutations even after relatively recent population splits.

Conclusions

Through our study, we gain new insights into the genomic diversity of European populations that have been subject to linguistic and geographic barriers to gene flow. We were able to shed more light on their heterogeneity, challenge the generalized view of isolates as units that originated from small founding groups, and reveal that genomic patterns of intra-population variation in open and isolated populations are distributed along a continuum. We believe that there are two possible avenues to follow up these first results. Firstly, a comparison of the structure of open and isolated populations using whole genome sequences would provide a complete representation of their genomic diversity. Secondly, extending comparisons to geographical contexts other than Europe will help us understand to what extent the observed patterns may be appropriate to isolates in other continental or regional scenarios. Waiting for further investigations, we hope this first study can reach its own target: in making us more aware of the value of human population isolates to understand how the interplay of environmental, socio-cultural and demographic factors, has shaped human genomic diversity.

Materials and Methods

Dataset. We assembled the genome-wide SNP chip data of 561 healthy unrelated adult individuals from 24 European populations. New genotype data were obtained for 211 subjects from three areas: (i) Sardinia (Benetutti, Carloforte, North Sardinia, Sulcis Iglesiente); (ii) German-speaking linguistic islands of the eastern Alps (Sappada, Sauris, Timau and Cimbrians from Lessinia); (iii) the Aosta province, in the Val d'Aosta region (north-western Italy). Only individuals with grandparents born in the same geographic area of sampling were enrolled in the study. Informed consent was obtained for all subjects. All methods were carried out in accordance with Italian Law (Decreto Legislativo della Repubblica Italiana, n° 196/2003). All experimental protocols were approved by the Bioethic Committee of the Azienda Ospedaliera Universitaria Pisana (Pisa, Italy. Prot N. 12702).

Genotyping, quality control and validation. All samples were genotyped using the Geno 2.0 DNA Ancestry Kit (www.genographic.com) SNP microarray known as the GenoChip[24] at the Gene-by-Gene laboratory (Family Tree DNA) in Houston, Texas. The autosomal AIMs (Ancestry Informative Markers) implemented in the GenoChip array provide an adequate coverage of the genetic diversity of European populations[24], and include rare variants occurring in small sized population samples. Furthermore, the geographic homogeneity of the typed populations minimized confounders that could potentially have originated from ascertainment bias when performing cross-population comparisons. The newly genotyped samples were merged with the reference data and then filtered according to the standard genotype quality control metrics using PLINK[50]. Only the SNPs with a genotyping success rate >90% were included, giving a total of 87,818 autosomal SNPs after the addition of the literature data. Only the individuals with a genotyping success rate >92% were used. Relatedness to the 3rd generation (Identity by Descent, IBD > 0.185) was tested with PLINK, and from the detected relative pairs, only one sample was randomly chosen for the subsequent analysis. We tested the power of the set of selected SNPs in detecting signals of isolation comparing them with those contained in the HGDP panel (647,789 SNPs) (see Supplementary Text S2).

Intra-population analyses. Runs of homozygosity (RoHs) were estimated using PLINK v1.9 (–homozyg option) (https://www.cog-genomics.org/plink2) under default settings (sliding window of 5 Mb, minimum of 50 SNPs, one heterozygous genotype and five missing calls allowed). Each SNP was considered to be part of

a homozygous segment when the proportion of overlapping homozygous windows was above 5%. RoHs were defined as stretches of at least 0.5 Mb with at least 25 homozygous SNPs[17]. We performed an unsupervised Gaussian fitting of the length distribution using *Mclust* from the R package mclust V3[51] and identified six different classes based on RoH'length (Supplementary Figure S1).

Intra-population sharing of IBD blocks between individuals were identified by the refined IBD algorithm implemented in Beagle v.4.1[52] adopting default parameters. Thereafter, we used the pairwise length of IBD sharing to calculate the statistic W_{int}[53]. This index represents the total length of the shared IBD blocks averaged over the number of possible pairs of individuals.

IBS values were estimated using PLINK v1.9 (–distance ibs option). By default, this option produces a lower-triangular tab-delimited text file with pairwise IBS between all individuals in the dataset. From this matrix, we extracted the values calculated between pairs of individuals belonging to the same population in order to obtain the intra-population pairwise IBS.

Blocks in linkage disequilibrium were calculated by using –blocks option (default settings) in PLINK v1.9. We used the –hardy option in PLINK v1.9 to obtain the average observed heterozygosity (het) per population and the inter-locus variance between all pairs of individuals, calculated as the square root of the standard deviation. Homozygosity was calculated as hom = (1 − het).

The Levene test for the equality of variances was performed with the R software package Car[54].

Principal component analysis[55] was performed with the R software package Ade4[56] using the above-described intra-population measures as variables.

Multiple regression analysis was performed with R software using the scores of the first and second principal component as dependent variables and the inbreeding coefficient, drift parameter (inferred by Treemix, considering the value from the nearest tree node, see below), the effective population size point estimates and the time since isolation as independent variables. The inbreeding coefficient was calculated as the proportion of the autosomal genome in runs of homozygosity, excluding the centromeres[27].

Inter-population analyses. Maximum likelihood estimation of individual ancestries was performed using ADMIXTURE v1.23[57] under default values (the block relaxation algorithm, a termination criterion set to stop when the log-likelihood increases by less than $\varepsilon = 10^{-4}$ between iterations and the quasi-Newton convergence acceleration method with q = 3 secant conditions). We applied unsupervised clustering analysis to the whole sample set, exploring the hypothesis of K = 1 to 10 clusters. We assessed cross-validation errors for each value of K using the ADMIXTURE's Cross Validation procedure.

MDS analysis was performed using PLINK v1.9 (–distance-matrix option). The information carried by each dimension was assessed by calculating the ratio of their respective eigenvalues compared to the sum of all eigenvalues.

Genetic structure and gene flow were investigated using TreeMix v1.1[58]. We set the position of the root (-root option) using a North African population (Egyptians[59]). To account for the fact that nearby SNPs are not independent, we grouped them together in windows of 500 SNPs using the -k flag. We ran Treemix with an increasing number of migration events, with 0 <= m <= 10. Runs with m comprised between 5 and 10 yielded comparable tree topologies and, since a higher number of migrations only partly improved the overall goodness of fit we chose to display m = 5 following a parsimonious approach.

The geographic distance matrix was calculated using the Geographic Distance Matrix Generator (http://biodiversityinformatics.amnh.org/open_source/gdmg). The geographical coordinates of the sampling areas were downloaded from the http://maps.cga.harvard.edu/gpf/. When the exact locations of sampling were unknown, we used the coordinates of the centroid of the nation as reported in http://gothos.info/resources/. In order to control for the effects of geographical proximity, we calculated the deviation of any observed genetic distance from the one predicted by the regression line obtained for geographic and genetic distances of open populations.

Estimate of the effective population size. Effective population size for all populations was estimated using the LD method[60] implemented in NeEstimator 2.0[61]. The LDNe algorithm estimates effective population size from the extent of linkage disequilibrium in the sample. Pairwise LD was calculated between 68,205 autosomal SNPs from 16 randomly chosen chromosomes. Other random combinations gave results that were very close to those reported in Supplementary Table S2. We decided against using the entire dataset because with too many loci, the method could not compute confidence intervals. We used a threshold of 0.05 as the lowest allele frequency, which gives the least biased results[62]. We reported estimated N_e with 95% (parametric) confidence intervals.

The number of generations since isolation and the relative values of effective population size under a model of demographic stationarity were calculated from the formula[63],

$$\Delta F = 1 - (1 - 1/2Ne)^t$$

where ΔF stands for the difference between the inbreeding coefficient estimated for each population and its hypothetical value at the time of population split. The latter parameter was assumed to range between the highest and lowest inbreeding coefficients observed among open populations.

References

1. Arcos-Burgos, M. & Muenke, M. Genetics of population isolates. *Clin Genet.* **61,** 233–247 (2002).
2. Kristiansson, K., Naukkarinen, J. & Peltonen, L. Isolated populations and complex disease gene identification. *Genome Biol.* **9,** 109 (2008).

3. Hatzikotoulas, K., Gilly, A. & Zeggini, E. Using population isolates in genetic association studies. *Brief Funct Genomics*. **13,** 371–377 (2014).
4. Pichler, I. *et al.* Genetic structure in contemporary south Tyrolean isolated populations revealed by analysis of y-chromosome, mtDNA, and Alu Polymorphisms. *Hum Biol.* **81,** 875–898 (2009).
5. Bosch, E. *et al.* Paternal and maternal lineages in the Balkans show a homogeneous landscape over linguistic barriers, except for the isolated Aromuns. *Ann Hum Genet*. **0,** 060721082338047 (2005).
6. Brandstätter, A. *et al.* Migration rates and genetic structure of two Hungarian ethnic groups in Transylvania, Romania. *Ann Hum Genet*. **71,** 791–803 (2007).
7. Nasidze, I., Quinque, D., Udina, I., Kunizheva, S. & Stoneking, M. The Gagauz, a linguistic enclave, are not a genetic isolate. *Ann Hum Genet*. **71,** 379–389 (2006).
8. Thomas, M. G. *et al.* New genetic evidence supports isolation and drift in the Ladin communities of the south Tyrolean Alps but not an ancient origin in the middle east. *Eur J Hum Genet*. **16,** 124–134 (2007).
9. Capocasa, M. *et al.* Detecting genetic isolation in human populations: A study of European language minorities. *PLoS ONE.* **8,** e56371 (2013).
10. Sheils, O., Finn, S. & O'Leary, J. Nucleic acid microarrays: An overview. *Curr Diagn Pathol*. **9,** 155–158 (2003).
11. de la Chapelle, A. & Wright, F. A. Linkage disequilibrium mapping in isolated populations: The example of Finland revisited. *Proc Natl Acad Sci USA* **95,** 12416–12423 (1998).
12. Jorde, L. B., Watkins, W. S., Kere, J., Nyman, D. & Eriksson, A. W. Gene mapping in isolated populations: New roles for old friends? *Hum Hered.* **50,** 57–65 (2000).
13. Varilo, T. *et al.* Linkage disequilibrium in isolated populations: Finland and a young sub-population of Kuusamo. *Eur J Hum Genet*. **8,** 604–612 (2000).
14. Service, S. *et al.* Magnitude and distribution of linkage disequilibrium in population isolates and implications for genome-wide association studies. *Nat Genet*. **38,** 556–560 (2006).
15. Garagnani, P. *et al.* Isolated populations as treasure troves in genetic epidemiology: The case of the Basques. *Eur J Hum Genet*. **17,** 1490–1494 (2009).
16. Veeramah, K. R. *et al.* Genetic variation in the Sorbs of eastern Germany in the context of broader European genetic diversity. *Eur J Hum Genet*. **19,** 995–1001 (2011).
17. Colonna, V. *et al.* Small effective population size and genetic homogeneity in the Val Borbera isolate. *Eur J Hum Genet*. **21,** 89–94 (2012).
18. Karafet, T. M. *et al.* Extensive genome-wide autozygosity in the population isolates of Daghestan. *Eur J Hum Genet*. **23,** 1405–1412 (2015).
19. Ayub, Q. *et al.* The Kalash genetic isolate: Ancient divergence, drift, and selection. *Am J Hum Genet*. **96,** 775–783 (2015).
20. Esko, T. *et al.* Genetic characterization of northeastern Italian population isolates in the context of broader European genetic diversity. *Eur J Hum Genet*. **21,** 659–665 (2012).
21. Novembre, J. *et al.* Genes mirror geography within Europe. *Nature*. **456,** 274–274 (2008).
22. Lao, O. *et al.* Correlation between genetic and geographic structure in Europe. *Curr Biol.* **18,** 1241–1248 (2008).
23. Jay, F., Sjodin, P., Jakobsson, M. & Blum, M. G. B. Anisotropic isolation by distance: The main orientations of human genetic differentiation. *Mol Biol Evol*. **30,** 513–525 (2012).
24. Elhaik, E. *et al.* The GenoChip: A new tool for genetic anthropology. *Genome Biol Evol*. **5,** 1021–1031 (2013).
25. International, S. *Ethnologue: Languages of the world*. http://www.ethnologue.com (2016)
26. Stevenson, A. & Dictionaries, O. *Oxford dictionary of English*. (Oxford University Press: New York, NY, 2010)
27. McQuillan, R. *et al.* Runs of Homozygosity in European populations. *Am J Hum Genet*. **83,** 359–372 (2008).
28. Kirin, M. *et al.* Genomic runs of Homozygosity record population history and consanguinity. *PLoS ONE.* **5,** e13996 (2010).
29. Pemberton, T. J. *et al.* Genomic patterns of Homozygosity in worldwide human populations. *Am J Hum Genet*. **91,** 275–292 (2012).
30. Palamara, P. F., Lencz, T., Darvasi, A. & Peer, I. Length distributions of identity by descent reveal fine-scale demographic history. *Am J Hum Genet*. **91,** 809–822 (2012).
31. Kruskal, J. B. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*. **29,** 1–27 (1964).
32. Pickrell, J. K. & Pritchard, J. K. Inference of population splits and mixtures from genome-wide Allele frequency data. *PLoS Genet.* **8,** e1002967 (2012).
33. Auton, A. *et al.* Global distribution of genomic diversity underscores rich complex history of continental human populations. *Genome Res.* **19,** 795–803 (2009).
34. Peltonen, L., Palotie, A. & Lange, K. Use of population isolates for mapping complex traits. *Nat Rev Genet.* **1,** 182–190 (2000).
35. Neel, J. K. Minority populations as genetic isolates: the interpretation of inbreeding results In Isolation, migration and health (D. F. Roberts, N. Fujiki, K. Torizuka, Eds) 17–23 (Cambridge University Press, 1992).
36. Di Gaetano, C. *et al.* Sardinians genetic background explained by runs of Homozygosity and Genomic regions under positive selection. *PLoS ONE.* **9,** e91237 (2014).
37. Pardo, L. M. *et al.* Dissecting the genetic make-up of north-east Sardinia using a large set of haploid and autosomal markers. *Eur J Hum Genet*. **20,** 956–964 (2012).
38. Contu, D. *et al.* Y-chromosome based evidence for Pre-Neolithic origin of the genetically homogeneous but diverse Sardinian population: Inference for association scans. *PLoS ONE.* **3,** e1430 (2008).
39. Piras, I. S. *et al.* Genome-wide scan with nearly 700 000 SNPs in two Sardinian sub-populations suggests some regions as candidate targets for positive selection. *Eur J Hum Genet*. **20,** 1155–1161 (2012).
40. Pistis, G. *et al.* High differentiation among Eight villages in a secluded area of Sardinia revealed by genome-wide high density SNPs analysis. *PLoS ONE.* **4,** e4654 (2009).
41. Molinari, G. La comunità linguistica In *Isole di Cultura: saggi sulle minoranze storiche germaniche in Italia* (ed. Prezzi, C.) (Centro Documentazione Luserna, 2004).
42. Toso, F. Le Minoranze *Linguistiche in Italia*. (Il mulino, 2008).
43. Rapelli, G. XIII comuni veronesi. La formazione dell'isola linguistica in *Isole di cultura* (Prezzi, 2004).
44. Ferraro, G. *Da Tabarka a San Pietro. Nasce Carloforte* (Musanti Editrice, 1986).
45. Robledo, R. *et al.* Analysis of a genetic isolate: The case of Carloforte (Italy). *Hum Biol.* **84,** 735–754 (2012).
46. Sanna, E., Iovine, M. C. & Floris G. Evolution of marital structure in 20 Sardinian villages from 1800 to 1974. *Anthropologischer Anzeiger.* **62,** 169–184 (2004).
47. Helgason, A., Yngvadóttir, B., Hrafnkelsson, B., Gulcher, J. & Stefánsson, K. An Icelandic example of the impact of population structure on association studies. *Nat Genet.* (2004).
48. Salmela, E. *et al.* Genome-wide analysis of single nucleotide Polymorphisms Uncovers population structure in northern Europe. *PloS ONE.* **3,** e3519 (2008).
49. Capocasa, M. *et al.* Linguistic, geographic and genetic isolation: a collaborative study of Italian populations. *J Anthropol Sci*. **92,** 201–231 (2014).
50. Purcell, S. *et al.* PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. **81,** 559–575 (2007).

51. Fraley, C. & Raftery, A. E. Model-based clustering, Discriminant analysis, and density estimation. *Journal of the American Statistical Association*. **97,** 611–631 (2002).
52. Browning, B. L. & Browning, S. R. Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics*. **194,** 459–471 (2013).
53. Atzmon, G. *et al.* Abraham's children in the genome era: Major Jewish Diaspora populations comprise distinct genetic clusters with shared middle eastern ancestry. *Am J Hum Genet*. **86,** 850–859 (2010).
54. Fox, J., Weisberg, H. S. & Weisberg, S. *An R companion to applied regression – 2nd edition*. (SAGE Publications, 2011).
55. Hotelling, H. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*. **24,** 417–441 (1933).
56. Dray, S. & Dufour, A.-B. The ade4 package: Implementing the Duality diagram for ecologists. *Journal of Statistical Software*. **22,** (2007).
57. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19,** 1655–1664 (2009).
58. Pickrell, J. K. & Pritchard, J. K. Inference of population splits and mixtures from genome-wide Allele frequency data. *PLoS Genet.* **8,** e1002967 (2012).
59. Behar, D. M. *et al.* The genome-wide structure of the Jewish people. *Nature* **466,** 238–242 (2010).
60. Waples, R. S. & Do, C. Ldne: A program for estimating effective population size from data on linkage disequilibrium. *Mol Ecol Res.* **8,** 753–756 (2008).
61. Do, C. *et al.* NeEstimator v2: Re-implementation of software for the estimation of contemporary effective population size (Ne) from genetic data. *Mol Ecol Res* **14,** 209–214 (2013).
62. Waples, R. S. & Do, C. Linkage disequilibrium estimates of contemporary N e using highly variable genetic markers: A largely untapped resource for applied conservation and evolution. *Evol Appl.* **3,** 244–262 (2010).
63. Harmon, L. J. & Braude, S. Conservation of Small Populations: Effective Population Sizes, Inbreeding, and the 50/500 Rule In *An introduction to methods and models in ecology, evolution, and conservation biology* (eds Braude, S. & Low, B. S.) (Princeton University Press, 2010).
64. Foster, M. W. Human genome diversity project (HGDP). *Encyclopedia of Life Sciences*, doi: 10.1002/9780470015902.a0005173.pub2 (2008).
65. Baztarrika, P. *et al. Fourth sociolinguistic survey 2006: [Basque Autonomous Community, Northern Basque Country, Navarre, Basque Country]* (Vitoria-Gasteiz, 2008).
66. Hellenthal, G. *et al.* A genetic Atlas of human Admixture history. *Science*. **343,** 747–751 (2014).

Acknowledgements

We are greatly indebted to all the blood donors. We would also like to thank Marcella Benedetti (Municipality of Sappada), Nino Pacilè and Lucia Protto (Municipality of Sauris), Vito Massalongo (Giazza), Ottaviano Matiz and Velia Plozner (Timau) for their valuable assistance in the sample collection and for their warm hospitality. This study was supported by a 2013 National Geographic Society Genographic 2.0 grant to ST. The survey in the Eastern Italian Alps was also funded by the Università di Roma "La Sapienza" (ref. C26A13HSHB) and the Istituto Italiano di Antropologia. The study was also supported by the European Research Council ERC-2011-AdG 295733 grant (Langelin) to DP.

Author Contributions

Conceived and designed the study: G.D.B., P.A., S.T., V.D. Provided samples: G.D.B., C.C., S.T., D.P., G.V. Extracted and prepared DNA for genomic analysis: C.B., C.C., S.T. Analyzed the data: P.A., V.D., L.P., S.T., P.F., V.C., S.S., A.B. Contributed reagents/materials/analysis tools: S.T., G.D.B., L.P., M.G.V., R.S.W., Wrote the manuscript: G.D.B., P.A. in collaboration with L.P. and V.C. Read and approved the final manuscript: all authors.

Additional Information

Accession codes: All data are available at the Zenodo Database (https://zenodo.org/) with accession number: 10.5281/zenodo.50114.

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Anagnostou, P. *et al.* Overcoming the dichotomy between open and isolated populations using genomic data from a large European dataset. *Sci. Rep.* **7**, 41614; doi: 10.1038/srep41614 (2017).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.