PLOS ONE

An Overview of the Genetic Structure within the Italian Population from Genome-Wide Data

Cornelia Di Gaetano[1,2]*, Floriana Voglino[2], Simonetta Guarrera[2], Giovanni Fiorito[2], Fabio Rosa[2], Anna Maria Di Blasio[3], Paola Manzini[4], Irma Dianzani[5], Marta Betti[5], Daniele Cusi[6,7], Francesca Frau[6,7], Cristina Barlassina[6,7], Dario Mirabelli[8], Corrado Magnani[9], Nicola Glorioso[10], Stefano Bonassi[11], Alberto Piazza[1,2], Giuseppe Matullo[1,2]

1 Department of Genetics, Biology and Biochemistry, University of Turin, Turin, Italy, 2 Human Genetics Foundation (HuGeF), Turin, Italy, 3 Istituto Auxologico Italiano, Cusano Milanino, Italy, 4 Banca del Sangue, San Giovanni-Molinette Hospital, Turin, Italy, 5 Department of Health Sciences, University of Eastern Piedmont, Novara, Italy, 6 Department of Medicine, Surgery and Dentistry, University of Milan, Milan, Italy, 7 Genomic and Bioinformatics Unit, Fondazione Filarete, Milan, Italy, 8 Unit of Cancer Epidemiology, University of Turin and CPO-Piemonte, Turin, Italy, 9 Department of Translational Medicine, University of Eastern Piedmont, Novara, Italy, 10 Hypertension and Related Diseases Center, AOU, University of Sassari, Sassari, Italy, 11 Unit of Clinical and Molecular Epidemiology, IRCCS San Raffaele Pisana, Rome, Italy

Abstract

In spite of the common belief of Europe as reasonably homogeneous at genetic level, advances in high-throughput genotyping technology have resolved several gradients which define different geographical areas with good precision. When Northern and Southern European groups were considered separately, there were clear genetic distinctions. Intra-country genetic differences were also evident, especially in Finland and, to a lesser extent, within other European populations. Here, we present the first analysis using the 125,799 genome-wide Single Nucleotide Polymorphisms (SNPs) data of 1,014 Italians with wide geographical coverage. We showed by using Principal Component analysis and model-based individual ancestry analysis, that the current population of Sardinia can be clearly differentiated genetically from mainland Italy and Sicily, and that a certain degree of genetic differentiation is detectable within the current Italian peninsula population. Pair-wise F_{ST} statistics Northern and Southern Italy amounts approximately to 0.001 between, and around 0.002 between Northern Italy and Utah residents with Northern and Western European ancestry (CEU). The Italian population also revealed a fine genetic substructure underscoring by the genomic inflation (Sardinia vs. Northern Italy = 3.040 and Northern Italy vs. CEU = 1.427), warning against confounding effects of hidden relatedness and population substructure in association studies.

Introduction

Genetic gradients are represented by continuous differences in allele frequencies created by events such as gene flow between two different populations, or by a demographic expansion into a scarcely populated environment, leading to a partial admixture with indigenous populations, genetic drift or differential selection [1]. These differences in allele frequencies may generate population stratification, which is an important confounding factor in genetic association studies [2]. The genetic composition of contemporary Europeans has been repeatedly studied using particular sets of markers that recent technologies have unveiled. Genetic differences between populations have been investigated by Menozzi *et al.* using gene frequencies of 38 classical pre-molecular markers [3] and by Ammerman *et al.* who described that the genetic composition of contemporary Europeans may have been shaped by a prehistoric demic diffusion that drove the expansion of agriculture [4].

Several studies have been carried out in the past 20 years focused on Y-chromosomal haplogroups and mtDNA across Europe [5,6,7,8,9,10,11,12,13,14]. Y chromosome markers are more geographically clustered, while the pattern of variability of mtDNA seems to be less spatially structured, although these single locus markers are easily subjected to genetic drift. Advances in high-throughput genotyping technology have provided greater information on differences between populations, and at the same time have shown that genetic gradients exist and correspond well to geographical areas [15,16,17,18,19,20]. However, genetically homogeneous populations do not always coincide with the 'political' definition of a country, but a recent paper has shown that clinal patterns in principal component analysis (PC) probably develop due to a simple isolation-by-distance process [21].

An initial overall representation of the European population structure on a fine-spatial scale was demonstrated by [15]. In this paper the first principal component (PC1) axis aligns with the

North-Northwest/South-Southeast direction, possibly justified by a special role for this geographic axis in the demographic history of Europeans. Finland was found to be the European population within single country genetic differences [20,22,23,24,25]. Differences between regions of the same country have also been shown within the population of the British Isles [26,27,28], as well as in the Swedish population [29,30], in Estonia [16] and Iceland [31]. Using genome-wide SNPs data, Nelis and colleagues have also shown a population structure on a fine–spatial scale in Italy, with a remarkable distinction between Southern Italians and other European populations [16]. In this scenario, using classical genetic markers [32,33,34], there is a certain degree of genetic substructure within Italy, especially between Sardinia and the Italian mainland.

Considering single nucleotide polymorphisms (SNPs) located in the non-recombining region of the Y chromosome (NRY), Italy remains within the range of European Y-chromosome variability, although a non-random distribution of Y markers was observed with more than 70% of Y chromosome diversity distributed along the North-South axis of the Italian peninsula [35]. One Y chromosome lineage, I-M26, is very common (40.9% of the population) in Sardinia [5], and it is also detected in some Western European populations, but with lower frequencies [36], like in the Bearnais (7.7%) or in the Basque, (Spanish and French) (6.0%) populations [6,36,37,38]. The distribution of this lineage in Europe indicates that M26 mutation occurred in a I1b Y chromosome from Western Europe, most likely in a population in Iberia/Southern France before the main initial peopling of Sardinia [39].

Geographical patterns of mtDNA variations within the Italian peninsula showed North-South clines with clear differences between Sardinian and the mainland populations [40].

Using genome-wide SNPs data, Nelis and colleagues have also shown a population structure on a fine–spatial scale in Italy, with a remarkable distinction between Southern Italians and other European populations [16].

We have investigated the genetic structure of the Italian population on a finer scale with respect to previous work by utilizing a greater number of markers and including more individuals in the study, to the best of our knowledge this is the first genome-wide SNP-based study focusing specifically on Italy. We estimate genetic differentiation among Italian samples and between Italian and other populations from the literature: 1) using a model free analysis like PC and a model-based analysis to infer individual ancestry components (ADMIXTURE software), 2) by calculating pair-wise F_{ST} statistics and estimating the identity-by-state (IBS) sharing between and within populations, 3) by estimating the genomic inflation factor in order to assess the effect of population stratification.

Materials and Methods

Study samples and data sets

DNA samples were obtained from 49 unrelated volunteers from four different Italian macro-areas (Tuscany, Sicily, Piedmont and Sardinia). Details of the affiliation of the municipalities within the macro-areas mentioned in this work are described in Figure S2. These individuals were grouped according to their birth place, and were selected to have their parents and four grandparents born in the same region. This small sample set is not a random sample of the modern, admixed population, but rather it should approach the historical population structure.

In addition we used existing genetic data relative to control subjects from both published and unpublished genome-wide association studies: a study on malignant pleural mesothelioma (MESO Study, [41]; a study on obesity (GEO-IT, Di Blasio at Auxologic Institute in Milan, unpublished); a study on hypertension (HYPERGENES [42]). For these control samples, we retrieved information on the birth place, thus possibly also including children of first generation migrants who were born in a different place from their parents. All participants signed an informed consent in agreement with the guidelines of the ethical committees of the institutions involved. An internal ethical review board at HuGeF foundation (Comitato Etico HUGEF/15-12-2011) approved the study. An ethical revision of the internal ethical steering board of the HYPERGENES STUDY, GEO-IT study and MESO study approved the entire process for each dataset. The complete dataset after validation was of 1,014 Italian samples.

We also included genotypic data from other populations freely available from the literature and living in the Mediterranean basin, like 57 CEU individuals from the HapMap project (Phase 2; release 23) [43], (Table 1); In addition, 134 individuals coming from the Middle East (Bedouin from Negev, Israel; Druze from Carmel, Israel; Palestinian from Israel) were included; 29 volunteers from Northern Africa (Mozabite from Mzab, Algeria) and 28 subjects from France from the Human Genome Diversity Panel (HGDP-CEPH) [44,45]. Samples from the literature were included in this study to increase the sample size for Italy, in order to compare Italy with other European populations and the Mediterranean basin and finally to estimate the degree of stratification between the CEU sample and Italy. In Table 1 a list of samples and data sets analyzed in our study is reported.

DNA extraction and genotyping

DNA was purified from blood samples by a standard phenol/chloroform extraction method. DNA concentrations were determined by spectrometry (NanoDrop 8000, Thermoscientific). The Illumina HumanOmni 1–QUAD, v1.0 BeadChip Array (Illumina Inc, S. Diego, CA, USA) was used to genotype 1,140,419 SNPs on 49 unrelated volunteers. Genotyping, according to the instructions provided by the manufacturer, was carried out at the Human Genetics Foundation (HuGeF) in Turin. All the information about this data set is reported in Table S1.

Quality Assessment and Control Procedure

Array-based SNP genotypes were subjected to stringent quality control, performed separately for each panel (Table S1). Samples that were too closely related to another sample (identity-by-state distance IBS <0.05) were removed. We also removed samples with genotype missing data >2%. Furthermore, we removed SNPs that had a minor allele frequency (MAF) <0.01 in all combined studies, or that failed Hardy-Weinberg equilibrium (HWE) with P< = 0.001. The average proportion of heterozygous genotypes at X chromosome, as described by Lao et al. 2008 [20], was used to avoid false gender assignments. Data management and quality control were carried out using the PLINK toolset [46]. For the present study we used only autosomal SNPs. The intersection between different data sets was of 163,355 SNPs and after three other steps MAF, HWE test, and *linkage disequilibrium* (LD) pruning, we used 125,799 SNPs in: 1,262 individuals in the European and Mediterranean dataset, 1,099 in the European data set, and 1,014 in the Italian dataset (746 in the Italian dataset excluding Sardinia).

Statistical Data Analyses

Principal Component Analysis. Principal Component Analysis was performed on a set of about 125,799 pruned

Table 1. Datasets and individuals number from each population.

	N-IT	C-IT	S-IT	SAR	CEU	FRE	PAL	BED	DRU	MOZ
HapMap		76			57					
HGDP	13	6		28		28	46	46	42	29
MESO	277	19	58	7						
Our sample	11	22	13	3						
GEO-IT	221	9	21	5						
HYPERGENES				225						
tot	522	130	92	268	57	28	46	46	42	29

Individuals included were filtered for individual call rate >98% and Identity By State (IBS) >0.05. Individuals included belong to Northern Italian (N-IT), Central Italian (C-IT), Southern Italian (S-IT), Sardinian (SAR), CEU HapMap (CEU), French (FRE), Palestinian (PAL), Bedouin (BED), Druze (DRU), and Mozabite (MOZ).
doi:10.1371/journal.pone.0043759.t001

markers. The pruning procedure was used to optimize the analysis of population structure, identifying a set of SNPs with low background LD ($r^2 = 0.5$).

To perform the analysis we used the function *prcomp* of R (package *mva*) [47], where the calculation is carried out by a singular value decomposition of the (centered and scaled) data matrix. We further confirmed PCA clustering by the K-means analysis [48]. The K-means clustering was calculated using the R package stats.

IBS analysis. The identity-by-state (IBS) sharing analysis [49] at both individual and population levels was performed. By using PLINK [46], the genome-wide average proportion of alleles sharing IBS was calculated for all subject pairs among the combined data sets. We further characterized the empirical distributions of IBS sharing within and between populations by using density estimation using the function of R (package *stats*) [47].

ADMIXTURE

The software ADMIXTURE [50] implements a model-based clustering method for estimating ancestry using autosomal genotype data consisting of unlinked markers. We included 126K SNPs resulting after the LD pruning. The analysis uses a cross-validation procedure to validate results, it was run with the number of clusters, K, from 2 to 10; we chose as termination criteria when the log-likelihood change between interactions falls below 0.001 and converged after 100 interactions.

F_{ST} and Mantel test

To estimate F_{ST} pairwise values between populations the Eigensoft program (Patterson et al, 2006) was used. High F_{ST} values implie a high degree of genetic differentiation among populations.

The Mantel test was calculated by using the R package *adegenet* [51]. It was used to test the relationship between the first two PC scores and the latitude and longitude. Finally, we computed the correlation between genetic distance (measured through Identity by Status matrix of PLINK), and geographical distance matrix of individuals (calculated with *dist* function of the R software, taking into account latitude and longitude of birth place).

Genomic control

We calculated the maximum possible inflation factor [52] between pairs of populations using PLINK [46].

Results

Principal component analysis of the Italian population

The eigenvectors for different subsets (HGDP-CEPH data, HapMap CEU and Tuscany data) were calculated in order to project the Italian data sets onto a two dimensional space (Figure 1) using 125,799 autosomal SNPs. The top 100 PCs were generated using the R software; we however focused on the top five, since the eigenvector values remained relatively constant in subsequent PCs, as indicated by the screeplot (Figure S1: top panel a); European dataset).

The position of the Italian population samples suggests that genetic distances between these populations and other European and Middle East populations has a good correlation with geographic distances. At the same time, Sardinia was confirmed to be a genetic "outlier".

Our main goal was to investigate the genetic structure of the Italian population considering four main macro-areas (Northern, Central, Southern Italy and Sardinia). We carried out PC analysis on the Italian samples and plotted the eigenvectors 1 and 2 in Figure 2. Most samples fell within a main cluster which seems to be indicative of Italian peninsula individuals. The first PC divided Italian populations in two clusters, one for Sardinia and the other for the remaining three Italian macro-areas. The Sardinian population is highly dispersed along the first eigenvector.

The second PC divided Italian mainland population into two clusters, with a certain degree of overlapping between Northern and Central Italy, and a separate cluster for Southern Italy, suggesting that genetic variation is generally continuous rather than discrete, at least within Italian mainland. In order to quantify the effect of migration we have done a correlation test [53] between PC1 or PC2 scores and geographical distance, Table S2 (Mantel test 1000 permutations PC1 R = 0.32, p-value = $2.2*e^{-16}$ and PC2 R = 0.49 p-value = $2.2*e^{-16}$). Mantel test was also used to compare genetic distance identities by state (IBS) between individuals and geographic distance: it showed a good correlation with geographic distance between sampled individuals (R = 0.41, p-value = $2.2*e^{-16}$, Table S2). In the same table we also showed a correlation between PC1 and PC2 scores with latitude and longitude.

Although a correlation between PCs and geography have been observed within the country (Figure 2), it is difficult, at this stage, to improve the accuracy at a regional level and perhaps a better clustering could be achieved by increasing the number of samples for each region (Figure S2).

The individuals' geographical clustering is mostly attributable to PC1 and PC2 while the subsequent components are influenced by
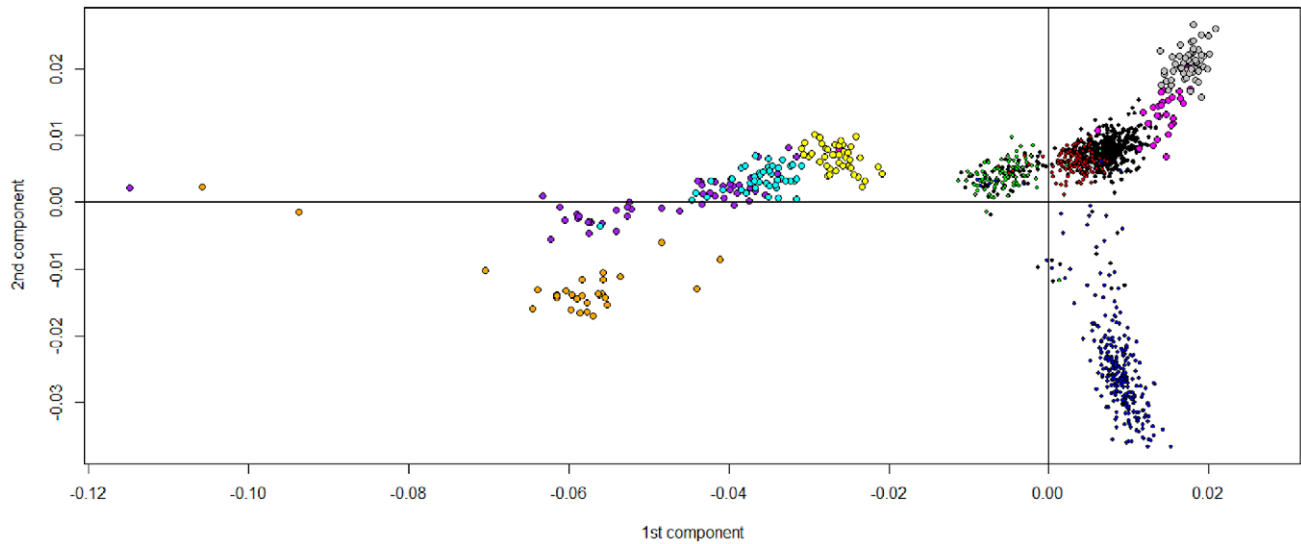
Figure 1. SNP-Based PC of 1,262 individuals from 10 sub-populations. The Italian population plotted onto the first two principal components defined by the European HGDP-CEPH populations and CEU HapMap data. Scatter plot of the first two principal components, obtained using R software (*prcomp*). Analysis based on 125,799 autosomal SNPs. Individuals included belong to Northern Italy (N-IT): black dots, Central Italy (C-IT): red dots, Southern Italy (S-IT): green dots, Sardinian (SAR): blue dots, CEU HapMap (CEU); light blue dots, Beduoin (BED): purple dots, Druze (DRU): yellow dots, Mozabite (MOZ): black triangles, Palestinian (PAL): red triangles, French (FRE): green triangles. The top 100 Eigenvectors and associated Eigenvalues for this plot are given in Supplementary Material: Figure S2 (panel a).
doi:10.1371/journal.pone.0043759.g001

particular regions along the genome. To further confirm this assumption, we calculated PC3 and PC4 on the Italian dataset and noted that they were also not stratified by a population label and at the same time we also observed that by using the K-means, it is possible to differentiate 3 clusters on PC4 (Figure S3). We have then plotted the contribution of each SNP for the first four PCs, against the genomic location, making a Manhattan plot (Figure

S4).We showed that for PC3 and PC4, the top SNPs localize to chromosome 8 between 8.135 and 11.90 Mb.

This genomic region is characterized by a large inversion with an unusual linkage disequilibrium (LD) pattern. Considering the state of this inversion, three different orientations of this region of DNA can be observed, i.e. inverted homozygous, heterozygous, or homozygous non-inverted. Using only the SNPs within this region (163 SNPs),those that contribute mainly to the PC4, we reiterated
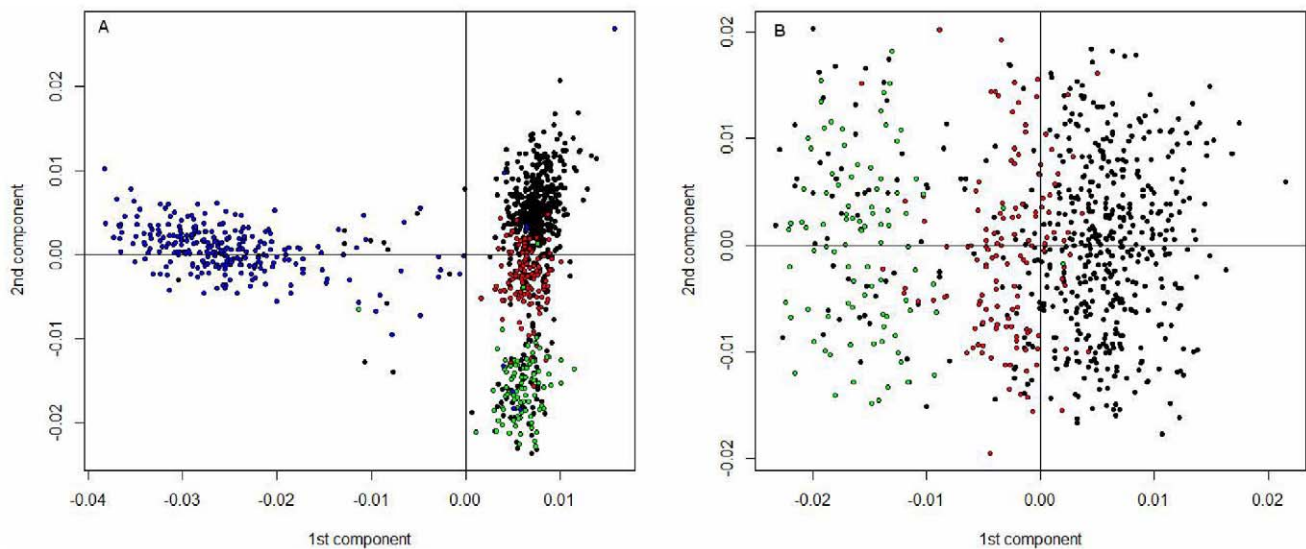

Figure 2. SNP-Based PC of 1,014 individuals from the Italian dataset. A. A Scatter Plot of the Italian population of the first two principal components obtained via R software (*prcomp*). Individuals included belong to Northern Italy : black dots, Central Italy : red dots, Southern Italy : green dots, Sardinian: blue dots. B. Italian population without the Sardinian-projected scatter plot of the first two principal components obtained via the R software (*prcomp*). Both analyses were based on 125,799 autosomal SNPs and 1,014 individuals for the Italian dataset and 746 individuals for the Italian dataset without Sardinia. Top 100 Eigenvectors and associated Eigenvalues for this plot are given in Supplementary Material: Figure S2 (panel b and c, respectively).
doi:10.1371/journal.pone.0043759.g002

the PC Analysis (Figure S5 panel A). We chose 6 HapMap CEU individuals predicted to be homozygous inverted, heterozygous, or homozygous non-inverted on chromosome 8p23.1 from the literature [54].

Using these individuals as a reference we roughly estimated the frequency of these genotypes by the K-means clustering (Figure S5 panel B). The estimated frequency of homozygous-inverted was 21%; 31.2% for homozygous non-inverted; and 47.9% for heterozygous individuals, on the entire Italian dataset. However, it must be emphasized that the PC analysis does not calculate frequency, rather it shows stratification of the genotype inversions. The real percentage of the three genotypes can only be obtained experimentally, as in Deng et al. 2008 [55] and more recently in Salm et al.2012 [56].

Model-based ancestry analysis

We used the ADMIXTURE software [50] calculating the ancestry fractions for each individual included in the analysis (Fig. 3). This software returns a cross-validation error value for each number of ancestral populations assumed for the analysis. The number of markers needed to resolve populations is generally inversely proportional to the genetic distance between the populations. By using 126K autosomal SNPs, we obtained at K = 4 the lowest cross-validation error. The HapMap CEU individuals showed an average Northern Europe (NE) ancestry (light green) of 83%. A similar pattern is observed in French, Northern Italian and Central Italian populations with a NE ancestry of 70%, 56% and 52% respectively (Figure 3). According to the PCA plot, also in the ADMIXTURE analysis there are relatively small differences in ancestry between Northern Italians and Central Italians while Southern Italians showed a lower average admixture NE proportion (43,6%) than Northern and Central Italy, and a higher Middle East ancestry (light blue) of 28%. The Sardinian samples display a pattern of crimson common to the others European populations but at a higher frequency (70.4%). The HGDP-CEPH Bedouins population showed a strong population substructure and apparently consisted of two different subpopulations on the basis of the percentage of the Middle East ancestry (light blue) and NE ancestry (light green). One of the clans was more similar to Palestinians. The HGDP-CEPH Mozabite population have an admixture proportion from Northern Africa (purple) of 73.2%. Figure S6 shows the results from K = 2 to K = 9. A plot of the distribution of cross-validation error estimate is shown in Figure S7.

IBS analysis

Distributions of IBS sharing between and within population are shown in Figure 4 for Northern Europe (CEU HapMap and French from HGDP-CEPH), for Middle Eastern populations, for Northern Africa populations (Mozabite) and Northern, Central Southern Italy and Sardinia. This analysis used data from 125,799 autosomal SNPs.

Density estimates for empirical distributions of genome-wide mean proportions of alleles shared between Italian and Northern European populations (Figure 4, top panel a) show that the median IBS sharing was higher for Northern Italy and lower for Sardinia and Southern Italy. Conversely, the median IBS sharing between Middle Eastern populations was higher for Southern Italy than for Northern Italy (Figure 4, center panel b). It was also possible to appreciate a lower IBS sharing between Mozabite populations (Figure 4, panel c).

In the Figure 4 bottom panel (D) the mean IBS sharing between pairs of individuals within each Italian population was described: the mean IBS was highest for Sardinians (0.72 with a standard

error $10*e^{-6}$, range 0.706–0.729), followed by Northern and Central Italy (both 0.717 with a standard error $4*e^{-6}$ and $4*e^{-5}$ and a range of 0.710–0.723 and 0.708–0.723, respectively) and then Southern Italy (0.715 with a standard error $2.6*e^{-5}$ a range of 0.707–0.722). The distribution mode was similar and higher for the populations of the Italian Peninsula, and lower for the Sardinians.

These results are consistent with the model-based ancestry analysis and with the position of these populations in the eigenvector PC space.

F_{st} analysis

We quantified genetic differentiation between the Italian population and the subset of HGDP-CEPH populations from Northern Europe (French), the Middle East (Druze, Palestinians and Bedouin), Northern Africa (Mozabite,) and from HapMap CEU, by calculating the pair-wise F_{ST} statistics. Estimates are given in Table 2. Notably, the genetic distance between Sardinia and each of the mainland Italian populations ($F_{ST} = 0.004$) was slightly lower than for many other European population pairs separated by larger geographical distances, for example, Southern Italians and CEU ($F_{ST} = 0.005$). In general, F_{ST} was lower between population pairs in closer geographical proximity like Southern Italians vs Central Italians ($F_{ST} = 0.001$). Southern Italians showed a genetic affinity with Middle East populations, such as Palestinian and Druze; and Northern Italian populations were genetically closer to the French and CEU populations. However, it should be stressed that the estimates are less accurate for pairs involving a population with a small sample size.

Population stratification analysis

Population stratification refers to a situation in which subgroups of individuals within the population of interest are, on average, more closely related to each other than to other individuals of the wider population. These allele frequency differences, can bias testing results and lead to artifact associations in case control studies. The genomic control inflation factor (λ_{GC}) was calculated to evaluate the possible impact of population stratification inside the four defined Italian subpopulations, and other populations from the literature. Results are shown in Table 2. The highest lambda value within Italy was 3.040, between Sardinia and Northern Italy. Moreover, other combinations also show substantial inflation, such as Southern vs Northern Italy ($\lambda_{GC} = 1.247$), and warns against the confounding effects of hidden relatedness and population substructure in association studies. As far as European and Middle Eastern or North African populations are concerned, results showed that a random Northern Italian population was well matched to the French population ($\lambda_{GC} = 1.102$), but this does not correspond to the comparison between Northern Italians and CEU from HapMap ($\lambda_{GC} = 1.427$).

Discussion

In this study, a genome-wide analysis of population structure within the Italian population from autosomal SNP data is presented for the first time. Using data from Human Genome Diversity Panel (HGDP-CEPH) [44,45,57] and from the HapMap Projects [43], we performed an ancestry analysis and PC analysis (Figure 1). We projected the Italian population onto a "map" defined by the first two PC, based on the European subset of the HGDP-CEPH dataset and some HapMap populations. The relative position of the samples reflected their geographic location: the close correlation between PC and geography, was previously reported by several authors [15,16,20,27,58,59]. When compared
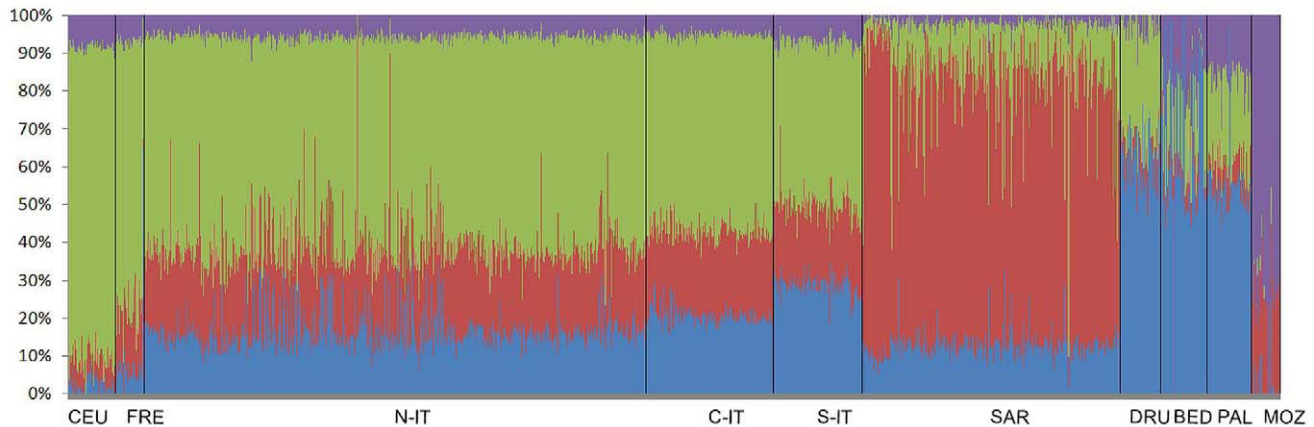
Figure 3. Identity-by-state (IBS) sharing between and within populations. Density estimates for empirical distributions of genome-wide mean proportions of alleles sharing identity-by-state between subjects from different population or within the same populations, are shown for A) Northern Europe (CEU and French), B) Middle East (Bedouin, Palestinian and Druze), C) Northern Africa (Mozambite), D) within Italian populations. Color code as in Figures 1 and 2.
doi:10.1371/journal.pone.0043759.g003

to other European populations, Sardinia was confirmed to be a genetic "outlier", whereas the Northern Italian population was genetically close to the French population, and the Southern Italians had some similarities with other Mediterranean populations such as those from Middle East. Unfortunately, lack of data from other relevant reference populations from the South-East Europe, e.g. from the Balkan peninsula, made it impossible to fully analyze the extent of the Eastern contribution in Italian populations. We also only made the PC analysis on Italian datasets with and without Sardinia (Figure 2). The first two PCs identified a good correlation with geographical distance and discriminate at least three of the four macro-areas within the Italian peninsula: Northern and Central, Sardinia and Southern Italy. Both analyses (Figure 1 and Figure 2, panel a) confirm the differentiation of Sardinia. In the PC analysis there is an appreciable degree of overlap between individuals born in Northern Italy but with a Southern ancestry (Figure 2, panel a), which could be explained by internal migration occurred during the last two generations, where people from Southern Italy have left their place of origin to look for better economic opportunities in other Italian regions.

A finer view of the Italian substructure, can be seen in Figure S2 where the hidden population structure within the Italian dataset is appreciable. Subjects are labeled by municipality, or in the case of the Sardinian samples, by the main linguistic area. In this figure we can appreciate the lack of clustering at the municipality level, also within Sardinia. Individuals seem to cluster within the main macro-area, but the geographic patterning is less obvious for the municipality (or in the case of Sardinia, linguistic) division, and in our opinion this pattern indicates no substructure within regions among municipalities, while the structuring between regions can be easily detected. It is also possible appreciate a certain genetic homogeneity within Sardinia.

The genetic structure observed in our dataset is expected to be mainly a consequence of demographic processes such as internal migration within and between the macro-areas. Indeed, Italy remains characterized by a strong migratory movements of the population within its territory [60] that was particularly significant from 1959–1970, but still present nowadays. Some authors [15] have calculated by using a multiple –regression –based assignment approach, that it was possible to locate more than 70% of Italian individuals within 400 km of their reported area of origin and

more than 90% within 800 km of their origin. Our PC analysis (Figure S2) seems to confirm these observations. It must also be emphasized that the results of this work were obtained using common SNPs and a more efficient geographic clustering could be reached using low –frequency alleles or haplotypes.

The overall F_{st} distribution fits with the PC analysis of the first two component. The F_{st} among the Italian macro-areas is moderate ($F_{st} \leq 0.001$) when considering the Italian Peninsula, but is more pronounced between Sardinia and the other macro-areas ($F_{st} = 0.004$) (Table 2). This is in agreement with observations by other authors [15,16] who reported that the average level of differentiation across Europe at each SNP is minimal (average $F_{st} = 0.004$ between different countries). A certain degree of genetic homogeneity shown by the F_{st} analysis, and by the partial overlapping of the distribution of the pairwise IBS within each of the Italian subpopulations, can be possibly explained by serial historical events and shared ancestry. The F_{st} values presented here are lower than those published by Nelis et al. 2009 using 270K SNPs; they reported that the Southern Italian population sample showed an F_{st} value of 0.005 compared to the Northern Italy sample. The highest value of this pairwise F_{st} matrix was found between the Finns from Kuusamo and Southern Italy ($F_{st} = 0.023$).

ADMIXTURE analysis confirms that there was no clear separation between Northern and Central Italy, at least as considered as macro-areas. Additional comparison of the distribution of pair-wise identity-by-state within each of the four populations and *ADMIXTURE* analysis clarified that this is not an artifact of the PC analysis. However, the PC and *ADMIXTURE* analysis results could be due to the sparse geographical coverage of our samples, especially for the Central and Northern macro-areas. In fact, many of the individuals (N = 413) in the North Italian sample analyzed in this study were from Piedmont- a North West Italian region- that has historically been affected by intense migration. At the same time, many individuals in the Central Italy macro-area (113 samples) are settled in Tuscany, an administrative region which is at the border with northern regions.

An intriguing result of the *ADMIXTURE* analysis was the proportion of ancestry in Sardinia, an ancestry shared with all the European and Northern African populations included in this analysis but with the highest level in Sardinia (Figure 3 crimson colour).
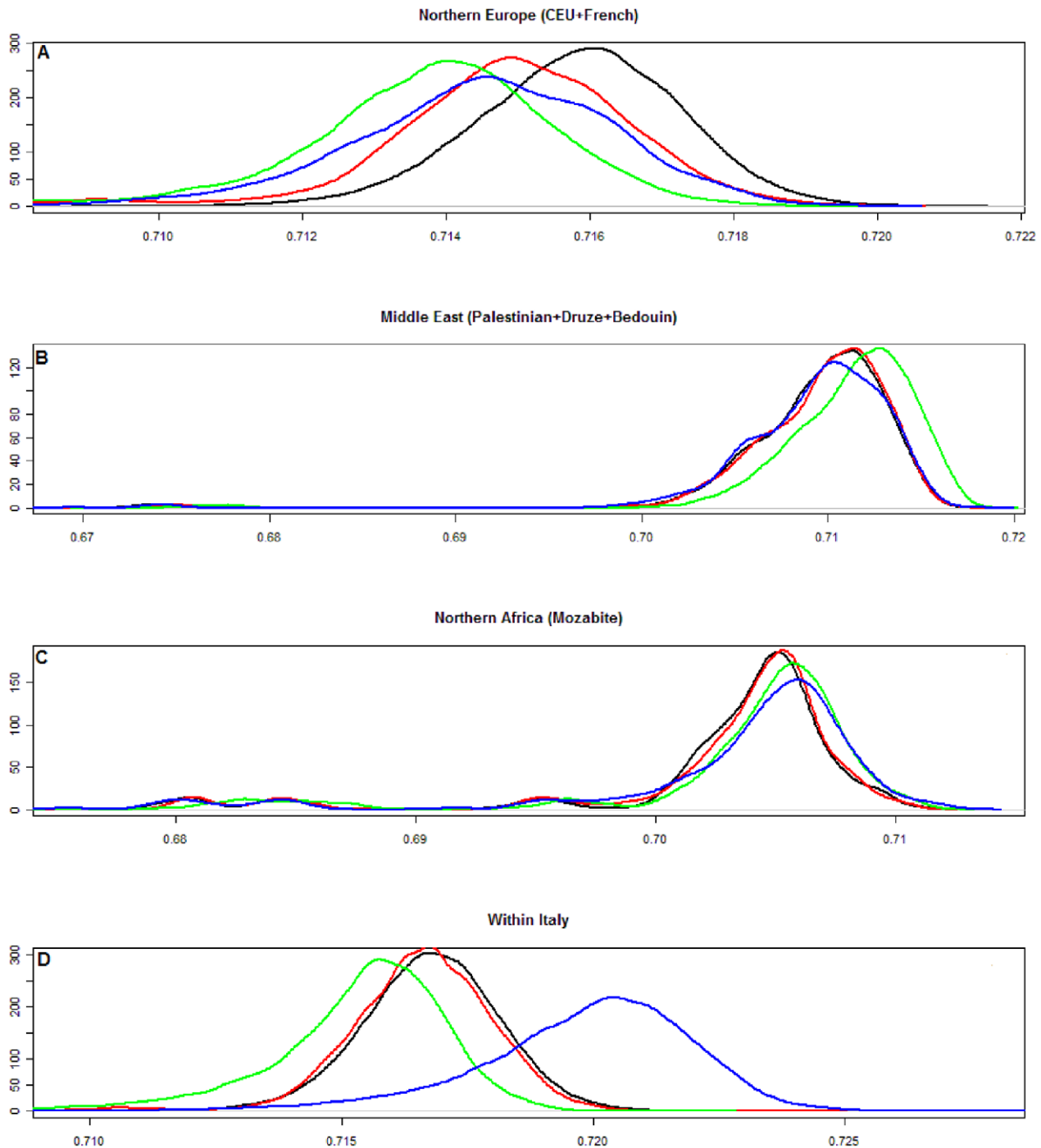
Figure 4. Clustering of the European, Northern African and Middle Eastern individuals by the Structure software. Model-based ancestry analysis based on a subset of HGDP-CEPH and HapMap CEU data using the merged data of 126K autosomal SNPs. Ancestry for each individual was inferred using ADMIXTURE [50] at K = 4. Abbreviations as in Figure 1.
doi:10.1371/journal.pone.0043759.g004

This average admixture proportion is widespread across all over the Sardinia island, with no geographic clustering, underlining an internal genetic homogeneity among the Sardinians. At the same time, this admixture proportion could be the signature of a common ancient genetic background of all the continental European populations but the isolation of the Sardinians have preserved this ancestry. The recent sequencing of the Iceman's genome, argues strongly in favor of the hypothesis that at least

continental Europeans, living 5,300 years ago, were more similar to the current Sardinians [61].

The average admixture proportions for Northern European ancestry within current Sardinian population is 14.3% with some individuals exhibiting very low Northern European ancestry (less than 5% in 36 individuals on 268 accounting the 13% of the sample).

Table 2. F_{st} values and genomic control inflation factor (λ_{GC}) between National areas.

	N-IT	C-IT	S-IT	SAR	CEU	BED	DRU	MOZ	PAL	FRE
N-IT	0	0	0.001	0.004	0.002	0.015	0.011	NA	0.01	0.001
C-IT	1.120	0	0.001	0.004	0.004	0.014	0.01	NA	0.009	0.002
S-IT	1.247	1.113	0	0.004	0.005	0.011	0.008	NA	0.006	0.003
SAR	3.040	2.213	1.878	0	0.009	0.018	0.014	NA	0.013	0.006
CEU	1.427	1.517	1.676	2.562	0	0.022	0.017	NA	0.016	0
BED	3.329	2.831	2.169	3.689	3.177	0	0.013	NA	0.008	0.019
DRU	2.578	2.220	1.790	2.985	2.654	2.073	0	NA	0.009	0.014
MOZ	3.961	3.569	2.980	**4.037**	3.611	2.508	3.019	0	NA	NA
PAL	2.574	2.174	1.703	3.028	2.663	1.677	1.831	2.529	0	0.014
FRE	**1.102**	1.153	1.293	1.590	1.044	2.308	1.998	2.818	1.968	0

F_{st} values above the diagonal; λ_{GC} below. Individuals included belong to Northern Italian (N-IT), Central Italian (C-IT), Southern Italian (S-IT), Sardinian (SAR), CEU HapMap (CEU), Bedouin (BED), Druze (DRU), Mozabite (MOZ), Palestinian (PAL), and French (FRE) populations.
doi:10.1371/journal.pone.0043759.t002

It is known that the major components are influenced by geographic clustering and secondly from areas with strong LD [17], and more precisely PC1 and PC2 are manly influenced by geography [15,17]) and PC3 and PC4 indeed may be influenced by large scale genome structural variation, as the HLA region or 8p23 or 15q24 and 17q21.31 and many others. In order to verify which genomic region mainly contribute to each PC we plotted the contribution of each SNP to the first four PCs (eigenvalues) against the genomic location (Figure S4). The major contribution for PC4 was provided by 163 SNPs located on 8p23. The inversion of 4 Mb on chr 8 (8p23) is perhaps the largest inversion included in our genome. For this reason PC3 and PC4 do not display a geographical clustering, but seem to organize into three groups, also underlined by a K-means analysis (Figure S3 panel A and B). Once this evidence was obtained, we selected just these 163 markers for use in PCA1/PC2 and K-means analysis (Figure S5). We then estimated the percentage of inverted-homozygous or heterozygous or homozygous non-inverted to 8p23.1 using six samples previously typed by HapMap as a reference.

When the combined information across many loci and many individuals is used, for example in the λ_{GC} analysis a higher degree of fine-scale population structure can be revealed. Systematic differences in sampling and genotyping are potential confounders, and may introduce a bias in association studies. The degree of genetic substructure between population pairs has been also measured by inflation factor (λ_{GC}) statistics [52,62]. For example, between Northern and Central Italy $\lambda_{GC} = 1.12$, and in cases from Sardinia and controls from Northern Italy the genomic control inflation factor was 3.040 (Table 2).

Within Italy allele frequency differences warrant caution when matching controls and cases, especially when involving individuals with Sardinian descent.

In conclusion, autosomal GWAS data, confirm that the genetic structure of the Italian population was strongly influenced by of the geographical distance. Moreover all Italian subpopulations show inflation factors among the largest within Europe, second only to the Finns [20]. Our work also described the appealing potential of reconstructing the genetic structure of Italy by using existing collections of samples with genome-wide data, even when a reduced amount of information concerning the ancestral background of the sample donors is available. The geographical resolution presented in this study, which use a reduced sample size, also demonstrates the possibility of detecting subtle population structures using samples where only the birth place is know. Hence, further National and International collaborative initiatives should be developed in order to most effectively exploit existing genomic data. However, a higher level of resolution can only be achieved by increasing the sample size, including subjects with well-defined geographical origins and selected local surnames, and using SNP genotyping platforms containing low-frequency alleles, in order to have a more balanced and complete representation of the Italian regions.

Supporting Information

Figure S1 Top 100 PC and associated Eigenvalues in different datasets. Top 100 PC and associated Eigenvalues for European dataset (a), Italian dataset (b) and Italian dataset without Sardinia (c).
(TIF)

Figure S2 Hidden population structure within the Italian dataset. Scatter plot of the first two eigenvectors based on 125,799 autosomal SNPs and 1,012 individuals. Colors represent the four different macro-areas; green- Southern Italy (Apulia, Calabria/Sicily, Campania, Basilicata), red- Central Italy (Tuscany, Lazio, Emilia Romagna and Abruzzo/Marche), black- Northern Italy (Piedmont,Liguria, Aosta Valley and Lombardy), blue- Sardinia (these samples were labeled for the linguistic area). Subjects are symbol- labeled by municipality. Information on municipality was not used for calculations.
(TIF)

Figure S3 Italian population projected scatter plot of the PC3 and PC4. Panel A Analysis based on 125,799 autosomal SNPs and 1,012 individuals. Color code shows different Italian subpopulations; green: Southern Italy, red: Central Italy, black: Northern Italy, blue: Sardinia.Panel B K-mean (K = 3) of PC3/PC4. Different colors shows the three diffent clusters.
(TIF)

Figure S4 Variable contribution of each SNP to the first four PC against the genomic location. Manhattan plot done in the Italian data set, the top SNPs in panel d localize to chromosome 8 inside the *8p23* region.
(TIF)

Figure S5 Individuals predicted to be homozygous inverted or heterozygous or homozygous non-inverted

for the 8p23.1. Panel A Scatter plot of the PC1 and PC2 done using only SNPs located inside the *8p23* region (163 markers). Panel B Individuals predicted to be homozygous inverted or heterozygous or homozygous non-inverted using K mean clustering, K = 3. The frequency of homozygous inverted (light blue triangles) was of 21%, of homozygous non-inverted (orange triangles) was of 31.19%, heterozygous of 47.9%. Individuals from HapMap used to confirm the predictions for the 8p23.1 were respectively NA12815 (homozygous inverted); NA11992 and NA12057 (homozygous non-inverted); NA11993 NA06993 and NA11994 (heterozygous).
(TIF)

Figure S6 Model-based ancestry analysis based on a subset from HGDP-CEPH and HapMap CEU data on 1260 individuals. Ancestry for each individual was inferred with ADMIXTURE [50] from K = 2 to K = 9.
(TIF)

Figure S7 Cross-validation error plot from the *ADMIX-TURE* program. Populations coming from the Italian dataset plus HGDP-CEPH (Palestinian; Druze; Mozambite; Bedouins; French) and some Hapmap populations. K = 1–10.
(TIF)

Table S1 Number of SNPs inside each panel before and after a filter for SNP call rate. SNPs intersection between the six studies was 163,355. After a filtering for minor allele frequency

(MAF>0.01) the number of polymorphisms was reduced to 163,350. A subset of 163,095 SNPs passed *Hardy-Weinberg equilibrium*. A subsequent dataset of 125,799 SNPs after *Linkage Disequilibrium* pruning were used for PC analysis, F_{st}, genomic control estimation and IBS analysis.
(DOC)

Table S2 Correlation between PC's score and genetic/ geographical values. Correlation between PC1 score, PC2 score, PC3 score PC4 score and genetic distance (IBS) and latitude, longitude and geographical distance (great circle distance) within the Italian dataset. All the correlation values were significative (p-value less than 2.2 e^{-16}).
(DOC)

Author Contributions

Conceived and designed the experiments: CDG FV GF AP GM. Performed the experiments: SG MB FF CB. Analyzed the data: CDG FV GF GM FR. Contributed reagents/materials/analysis tools: AMDB PM DC ID DM SB CM NG. Wrote the paper: CDG GM.

References

1. Cavalli-Sforza LL (2007) Human evolution and its relevance for genetic epidemiology. Annu Rev Genomics Hum Genet 8: 1–15.
2. Marchini J, Cardon LR, Phillips MS, Donnelly P (2004) The effects of human population structure on large genetic association studies. Nat Genet 36: 512–517.
3. Menozzi P, Piazza A, Cavalli-Sforza L (1978) Synthetic maps of human gene frequencies in Europeans. Science 201: 786–792.
4. Ammerman AJ, Cavalli-Sforza LL (1984) The Neolithic transition and the genetics of populations in Europe. The Neolithic Transition and the Genetics of Populations in Europe.
5. Semino O, Passarino G, Oefner PJ, Lin AA, Arbuzova S, et al. (2000) The genetic legacy of paleolithic Homo sapiens sapiens in extant europeans: A Y chromosome perspective. Science 290: 1155–1159.
6. Rootsi S, Magri C, Kivisild T, Benuzzi G, Help H, et al. (2004) Phylogeography of Y-chromosome haplogroup I reveals distinct domains of prehistoric gene flow in Europe. American Journal of Human Genetics 75: 128–137.
7. Achilli A, Olivieri A, Pala M, Metspalu E, Fornarino S, et al. (2007) Mitochondrial DNA variation of modern Tuscans supports the Near Eastern origin of Etruscans. American Journal of Human Genetics 80: 759–768.
8. Achilli A, Rengo C, Magri C, Battaglia V, Olivieri A, et al. (2004) The molecular dissection of mtDNA haplogroup H confirms that the Franco-Cantabrian glacial refuge was a major source for the European gene pool. American Journal of Human Genetics 75: 910–918.
9. Malyarchuk BA, Grzybowski T, Derenko MV, Czarny J, Wozniak M, et al. (2002) Mitochondrial DNA variability in Poles and Russians. Ann Hum Genet 66: 261–283.
10. Di Giacomo F, Luca F, Anagnou N, Ciavarella G, Corbo RM, et al. (2003) Clinal patterns of human Y chromosomal diversity in continental Italy and Greece are dominated by drift and founder effects. Mol Phylogenet Evol 28: 387–395.
11. Torroni A, Bandelt HJ, D'Urbano L, Lahermo P, Moral P, et al. (1998) mtDNA analysis reveals a major late Paleolithic population expansion from southwestern to northeastern Europe. Am J Hum Genet 62: 1137–1152.
12. Semino O, Magri C, Benuzzi G, Lin AA, Al-Zahery N, et al. (2004) Origin, Diffusion, and Differentiation of Y-Chromosome Haplogroups E and J: Inferences on the Neolithization of Europe and Later Migratory Events in the Mediterranean Area. American Journal of Human Genetics 74: 1023–1034.
13. Balanovsky O, Rootsi S, Pshenichnov A, Kivisild T, Churnosov M, et al. (2008) Two Sources of the Russian Patrilineal Heritage in Their Eurasian Context. American Journal of Human Genetics 82: 236–250.
14. Loogvali EL, Roostalu U, Malyarchuk BA, Derenko MV, Kivisild T, et al. (2004) Disuniting uniformity: a pied cladistic canvas of mtDNA haplogroup H in Eurasia. Mol Biol Evol 21: 2012–2021.
15. Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, et al. (2008) Genes mirror geography within Europe. Nature 456: 98–101.
16. Nelis M, Esko T, Magi R, Zimprich F, Zimprich A, et al. (2009) Genetic structure of Europeans: a view from the North-East. PLoS One 4: e5472.
17. McEvoy BP, Montgomery GW, McRae AF, Ripatti S, Perola M, et al. (2009) Geographical structure and differential natural selection among North European populations. Genome Research 19: 804–814.
18. Tian C, Plenge RM, Ransom M, Lee A, Villoslada P, et al. (2008) Analysis and application of European genetic substructure using 300 K SNP information. PLoS Genet 4: e4.
19. Tian C, Kosoy R, Nassir R, Lee A, Villoslada P, et al. (2009) European population genetic substructure: further definition of ancestry informative markers for distinguishing among diverse European ethnic groups. Mol Med 15: 371–383.
20. Lao O, Lu TT, Nothnagel M, Junge O, Freitag-Wolf S, et al. (2008) Correlation between Genetic and Geographic Structure in Europe. Current Biology 18: 1241–1248.
21. Novembre J, Stephens M (2008) Interpreting principal component analyses of spatial population genetic variation. Nat Genet 40: 646–649.
22. Jakkula E, Rehnstrom K, Varilo T, Pietilainen OP, Paunio T, et al. (2008) The genome-wide patterns of variation expose significant substructure in a founder population. Am J Hum Genet 83: 787–794.
23. Huyghe JR, Fransen E, Hannula S, Van Laer L, Van Eyken E, et al. (2011) A genome-wide analysis of population structure in the Finnish Saami with implications for genetic association studies. Eur J Hum Genet 19: 347–352.
24. Seldin MF, Shigeta R, Villoslada P, Selmi C, Tuomilehto J, et al. (2006) European population substructure: clustering of northern and southern populations. PLoS Genet 2: e143.
25. Sajantila A, Salem AH, Savolainen P, Bauer K, Gierig C, et al. (1996) Paternal and maternal DNA lineages reveal a bottleneck in the founding of the Finnish population. Proc Natl Acad Sci U S A 93: 12035–12039.
26. Winney B, Boumertit A, Day T, Davison D, Echeta C, et al. (2011) People of the British Isles: preliminary analysis of genotypes and surnames in a UK-control population. Eur J Hum Genet.
27. (Wellcome Trust Case Control Consortium 2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447: 661–678.
28. O'Dushlaine CT, Morris D, Moskvina V, Kirov G, Consortium IS, et al. (2010) Population structure and genome-wide patterns of variation in Ireland and Britain. Eur J Hum Genet 18: 1248–1254.
29. Humphreys K, Grankvist A, Leu M, Hall P, Liu J, et al. (2011) The genetic structure of the Swedish population. PLoS One 6: e22547.
30. Salmela E, Lappalainen T, Liu J, Sistonen P, Andersen PM, et al. (2011) Swedish population substructure revealed by genome-wide single nucleotide polymorphism data. PLoS One 6: e16747.
31. Price AL, Helgason A, Palsson S, Stefansson H, St Clair D, et al. (2009) The impact of divergence time on the nature of population structure: an example from Iceland. PLoS Genet 5: e1000505.

32. Cavalli-Sforza LL, Piazza A (1993) Human genomic diversity in Europe: a summary of recent research and prospects for the future. Eur J Hum Genet 1: 3–18.

33. Piazza A, Cappello N, Olivetti E, Rendine S (1988) A genetic history of Italy. Ann Hum Genet 52: 203–213.

34. Cavalli-Sforza LLPA, Menozzi P, editor (1994) History and geography of human genes.

35. Capelli C, Brisighelli F, Scarnicci F, Arredi B, Caglia A, et al. (2007) Y chromosome genetic variation in the Italian peninsula is clinal and supports an admixture model for the Mesolithic-Neolithic encounter. Mol Phylogenet Evol 44: 228–239.

36. Capelli C, Redhead N, Abernethy JK, Gratrix F, Wilson JF, et al. (2003) A Y chromosome census of the British Isles. Curr Biol 13: 979–984.

37. Bosch E, Calafell F, Comas D, Oefner PJ, Underhill PA, et al. (2001) High-resolution analysis of human Y-chromosome variation shows a sharp discontinuity and limited gene flow between northwestern Africa and the Iberian Peninsula. Am J Hum Genet 68: 1019–1029.

38. Maca-Meyer N, Sanchez-Velasco P, Flores C, Larruga JM, Gonzalez AM, et al. (2003) Y chromosome and mitochondrial DNA characterization of Pasiegos, a human isolate from Cantabria (Spain). Ann Hum Genet 67: 329–339.

39. Morelli L, Contu D, Santoni F, Whalen MB, Francalacci P, et al. (2010) A comparison of Y-chromosome variation in Sardinia and Anatolia is more consistent with cultural rather than demic diffusion of agriculture. PLoS One 5: e10419.

40. Barbujani G, Bertorelle G, Capitani G, Scozzari R (1995) Geographical structuring in the mtDNA of Italians. Proc Natl Acad Sci U S A 92: 9171–9175.

41. Betti M, Neri M, Ferrante D, Landi S, Biava A, et al. (2009) Pooled analysis of NAT2 genotypes as risk factors for asbestos-related malignant mesothelioma. Int J Hyg Environ Health 212: 322–329.

42. Salvi E, Kutalik Z, Glorioso N, Benaglio P, Frau F, et al. (2011) Genomewide Association Study Using a High-Density Single Nucleotide Polymorphism Array and Case-Control Design Identifies a Novel Essential Hypertension Susceptibility Locus in the Promoter Region of Endothelial NO Synthase. Hypertension.

43. Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, et al. (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449: 851–861.

44. Cann HM, de Toma C, Cazes L, Legrand MF, Morel V, et al. (2002) A human genome diversity cell line panel. Science 296: 261–262.

45. Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, et al. (2008) Worldwide human relationships inferred from genome-wide patterns of variation. Science 319: 1100–1104.

46. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, et al. (2007) PLINK: A tool set for whole-genome association and population-based linkage analyses. American Journal of Human Genetics 81: 559–575.

47. Development Core Team R (2008) R: *A Language and Environment for Statistical Computing*. In: R Foundation for Statistical Computing, editor. Vienna, Austria

48. Hartigan JAaMAW (1979) Algorithm AS 136: A K-Means Clustering Algorithm. Journal of the Royal Statistical Society Series A: Statistics in Society 28: 100–108.

49. Bowcock AM, Ruiz-Linares A, Tomfohrde J, Minch E, Kidd JR, et al. (1994) High resolution of human evolutionary trees with polymorphic microsatellites. Nature 368: 455–457.

50. Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. Genome Res 19: 1655–1664.

51. Jombart T (2008) adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics 24: 1403–1405.

52. Devlin B, Roeder K (1999) Genomic control for association studies. Biometrics 55: 997–1004.

53. Mantel N (1967) The detection of disease clustering and a generalized regression approach. Cancer Res 27: 209–220.

54. Bosch N, Morell M, Ponsa I, Mercader JM, Armengol L, et al. (2009) Nucleotide, cytogenetic and expression impact of the human chromosome 8p23.1 inversion polymorphism. PLoS One 4: e8269.

55. Deng L, Zhang Y, Kang J, Liu T, Zhao H, et al. (2008) An unusual haplotype structure on human chromosome 8p23 derived from the inversion polymorphism. Hum Mutat 29: 1209–1216.

56. Salm MP, Horswell SD, Hutchison CE, Speedy HE, Yang X, et al. (2012) The origin, global distribution, and functional impact of the human 8p23 inversion polymorphism. Genome Res 22: 1144–1153.

57. Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, et al. (2002) Genetic structure of human populations. Science 298: 2381–2385.

58. Tian C, Plenge RM, Ransom M, Lee A, Villoslada P, et al. (2008) Analysis and application of European genetic substructure using 300 K SNP information. PLoS Genet 4: 0029–0039.

59. Price AL, Butler J, Patterson N, Capelli C, Pascali VL, et al. (2008) Discerning the ancestry of European Americans in genetic association studies. PLoS Genet 4: e236.

60. Bonifazi C, and Heins F. (2000) Long-term trends of internal migration in Italy. International Journal of Population Geography 2: 111–131.

61. Keller A, Graefen A, Ball M, Matzas M, Boisguerin V, et al. (2012) New insights into the Tyrolean Iceman's origin and phenotype as inferred by whole-genome sequencing. Nat Commun 3: 698.

62. Hao K, Li C, Rosenow C, Wong WH (2004) Detect and adjust for population stratification in population-based association study using genomic control markers: an application of Affymetrix Genechip Human Mapping 10K array. Eur J Hum Genet 12: 1001–1006.