The Italian genome reflects the history of Europe and the Mediterranean basin

The Italian Genome reflects the History of Europe and the Mediterranean Basin

Giovanni Fiorito,[1,2] Cornelia Di Gaetano,[1,2] Simonetta Guarrera,[1,2] Fabio Rosa,[2] Marcus W. Feldman,[3] Alberto Piazza,[1,2] Giuseppe Matullo.[1,2]

[1]Department of Medical Sciences, University of Torino, Turin, Italy.

[2]HuGeF Human Genetics Foundation, Turin, Italy.

[3]Department of Biology, Stanford University, Stanford, CA, USA.

Short title: Genetic history of Italy

Corresponding author: Giuseppe Matullo

e-mail: giuseppe.matullo@unito.it

Address: Via Nizza 52, 10026 Turin, Italy.

Tel: +39 011 6709542

Abstract

Recent scientific literature has highlighted the relevance of population genetic studies both for disease association-mapping in admixed populations and for understanding the history of human migrations. Deeper insight into the history of the Italian population is critical for understanding the peopling of Europe. Because of its crucial position at the centre of the Mediterranean basin, the Italian peninsula has experienced a complex history of colonization and migration whose genetic signatures are still present in contemporary Italians. In this study we investigated genomic variation in the Italian population using 2.5 million single nucleotide polymorphisms (SNPs) in a sample of more than 300 unrelated Italian subjects with well-defined geographical origins. We combined several analytical approaches to interpret genome-wide data on 1,272 individuals from European, Middle Eastern, and North African populations. We detected three major ancestral components contributing different proportions across the Italian peninsula, and signatures of continuous gene flow within Italy, which have produced remarkable genetic variability among contemporary Italians. In addition, we have extracted novel details about the Italian population's ancestry, identifying the genetic signatures of major historical events in Europe and the Mediterranean basin from the Neolithic (e.g., peopling of Sardinia) to recent times (e.g., 'barbarian invasion' of Northern and Central Italy). These results are valuable for further genetic, epidemiological and forensic studies in Italy and in Europe.

Keywords: Population genetics, ancestry, admixture, Identity by Descent, Italian population.

Introduction

The Italian population has a greater degree of internal genomic variability than other European countries.[1] This reflects geographic isolation within Italy, due to its mountainous topography, and to historical events that triggered demographic changes. An example of the latter is the long history of the Roman Empire (27 BCE–476 CE), whose decline was accompanied by a considerable reduction in population size and a series of subsequent migratory waves across Italy. Due to its crucial position at the centre of the Mediterranean basin, the Italian peninsula has experienced a complex history of colonization and migration and the genetic signatures of these human origins are still present in contemporary Italians.[2-4] Deeper insight into the history of the Italian population is also critical for understanding the peopling of Europe.

The genetic structure of Italy, whose unity of people and culture is quite recent, was initially analysed using classical genetic markers by Piazza *et al.*[5] and Cavalli-Sforza *et al.*[6] Recently, using genome-wide data O'Dushlaine *et al.* defined fine-scale genetic differences among people from different rural villages in Northern and Central Italy.[7] Geographical patterns of Y chromosomal and mtDNA diversity in Italy, mainly determined by the combined action of drift and founder effects, have been described.[8,9] A correlation between genetic and geographic structure in Europe has also been found[10,11] with a detectable distinction between Southern Italians and other Europeans.

In addition to evolutionary history, the identification of genetic substructure in apparently homogeneous populations can improve association-mapping in admixed populations.[13] Moreover, it has recently been shown that variation in susceptibility to certain diseases, and response to drugs or therapies, is related to different proportions of non-European ancestry in admixed populations.[12]

The aim of this study was to investigate fine-scale Italian population genetic substructure. We used both the large SNP dataset that we collected from a well characterized Italian sample, and the most recent haplotype-based population genetic algorithms, such as fineSTRUCTURE,[14] which are able to provide finer resolution of the genetic structure of populations, as was shown for the UK population by Leslie *et al*.[15] Specifically, our aims were to test the feasibility of identifying differences at the micro-regional level within Italy, to compare and quantify the contributions of populations from Europe and the Mediterranean basin to the genetic composition of the Italians, and to explore the historical events that led to the observed high genomic variability within Italy. Several analytical approaches were combined in order to obtain a complete portrait of 'the Italian genome', and to test the robustness of our results across different methodologies. We also traced the major historical events that led to the complex genomic mosaic observed within Italy, including demographic changes and waves of migration across Europe and the Mediterranean basin.

Materials and Methods

Subjects: Italian subjects were selected from eleven out of the twenty Italian administrative regions according to a list of specific region/geographical area/province typical surnames on the basis of previous works by Zei *et al*.[16,17] Each subject included in the study had a well-defined geographical origin: four grandparents (and parents) born in the same administrative region, assessed through interviews at the time of blood collection. The distribution of the study samples across Italy and the sampling provinces is shown in Figure 1 and Table S1 with sample sizes. Informed consent was provided by all the participants at

the time of enrolment. An internal ethical review board at the Human Genetics Foundation (HuGeF, Comitato Etico HUGEF/15-12-2011) approved this study.

Datasets: Two datasets were used to address the different topics of this study.

The first comprised 1,698,926 SNPs investigated in 300 Italian samples, and was used to compare genetic profiles among subjects sampled from the different Italian regions and macro-areas. Genome-wide genotype data and the macro-area of origin for each sample will be made available via the European Genome-phenome Archive (EGA) repository (https://www.ebi.ac.uk/ega) under accession number EGAS00001001458, and they are already available upon request from the HuGeF repository (see Additional information section for the reference web link). See Supplementary Methods for details about data collection, DNA extraction, genotyping, quality controls, and SNP imputation procedures.

The second dataset comprised 347,131 SNPs assayed on 1,272 samples (Table 1), and was used for the comparison between the Italian and non-Italian populations.

Italian population substructure: the Italian population substructure was investigated with the model-based Bayesian cluster algorithm implemented in the combined software ChromoPainter-fineSTRUCTURE,[14] using a model that takes into account linkage disequilibrium between SNPs ('linked' model) with the default parameters. The results were reported as a heatmap (co-ancestry matrix) of pairwise similarity between subjects, expressed as the number of genomic segments inherited by the same source population. Principal Component Analysis (PCA) was carried out using the same co-ancestry matrix, and a dendrogram based on hierarchical clustering was constructed. The clustering algorithm employs a Bayesian approach, in which the number of donor populations *K* and the expected proportion of chunks from each donor population to each individual are

inferred by maximum likelihood. Assignment to each cluster is then performed using a Markov Chain Monte Carlo (MCMC) method. See Lawson *et al.*[14] for details.

Identical by Descent (IBD) calling: Pairwise shared IBD segments were identified by the fastIBD method implemented in BEAGLE v3.2[18] using default parameters. The post-processing procedure suggested by Ralph and Coop[19] was used to minimize the number of false positive calls. Briefly, the IBD call algorithm was run ten times with ten different random seeds; any segment not overlapping a segment seen in at least one other run was removed; any two segments separated by a gap shorter than at least one of the segments and no more than 5 cM long were merged; any merged segment that did not contain a sub-segment with score below $1x10^{-9}$ was removed.

The statistics W_{AB} defined by Atzmon *et al.*[20] and L_{AB} defined by Botigue *et al.*[21] were computed as a summary of IBD sharing. The first is an index of the total length of the shared IBD blocks averaged over the number of possible pairs of individuals (one from population A and the other from population B); the second is an index of the average length of a segment shared IBD between a pair of individuals normalized over the possible number of pairwise comparisons between populations. A block jackknife procedure was used to compute standard errors and confidence intervals for both estimates.

Isolation by distance and genetic boundary tests: The Mantel test[22] was applied to the correlation between geographical distance (the shortest distance on a roadmap) expressed in kilometres (km) and genetic similarity measured by the W statistic defined above. Statistical significance was evaluated by computing an empirical p-value based on 10,000 permutations. To verify the presence of genetic boundaries across the Italian regions we used Monmonier's algorithm.[23] Statistical significance of the genetic boundaries was

computed as suggested in Manni *et al.*[23] Briefly, the test is based on analysis of resampled bootstrap matrices: a score is associated with all the different edges that constitute barriers and indicates how many times each edge is included in one of the N boundaries computed in order to determine an empirical p-value.

Effective population size: The effective population size (N_e) was inferred for each of the three Italian macro-areas (Northern, Central, and Southern Italy) and each of the eleven Italian regions separately, taking advantage of the relationship between the number and length of shared IBD segments and N_e as described by Palamara and Pe'er.[24] Specifically, we estimated the 'ancestral' N_e from the number of short (less than 2 cM) IBD shared segments only.

Admixture analysis: The unsupervised algorithm implemented in the software ADMIXTURE 1.23[25] was used to infer ancestry proportions for all the Italian, European, Middle Eastern and North African samples. Nine ancestral cluster arrangements (K=2,…,10) were tested, and the cross validation error procedure implemented in the same software was used to define the most realistic contributions of ancestral populations to the currently observed pattern. The procedure was repeated twenty times and the results were averaged. To avoid bias due to differences in sample sizes among populations this analysis was performed with a balanced sample size (twenty individuals per population). Results were further validated using an independent methodology implemented in the software RFMIX.[26] See Supplementary Methods for details.

F_{st} genetic distance: The genetic distance between all population pairs was assessed by computing the F_{st} (fixation index) measure, as implemented in the R package snpStats.[27] Confidence intervals were obtained by means of a block jackknife procedure.

Time since admixture events: To estimate the time of admixture events, we used the extension of the ROLLOFF method implemented in the software ALDER.[28] For this analysis, the Italian regions were grouped according to the five previously identified clusters (Northern, Central, Southern Italy, Aosta Valley, and Sardinia), and admixture was tested against each of the non-Italian populations included in the study. See Supplementary Methods for details.

Results

Italian population substructure:

The fineSTRUCTURE co-ancestry matrix is shown in Figure 2A. The pie charts on the bottom panel show the overlap between the observed clusters and the regions of origin. The sensitivity of the cluster algorithm in assigning each sample to the correct macro-area was 96.43%, 86.55%, 92.00%, 94.44%, and 100% for Aosta Valley, Northern Italy, Central Italy, Southern Italy, and Sardinia, respectively, whereas the specificity was 98.16%, 96.68%, 94.80%, 99.56%, and 100% respectively. In view of the above results, the Italian regions were divided into five groups for the statistical analyses: Northern, Central, Southern Italy, Aosta Valley, and Sardinia.

The distribution of the Italian individuals from the first two eigenvectors of the PCA is shown in Figure 2, both including and excluding Sardinians (Figures 2B,C). The PCA results provided evidence of large differences between Sardinians and other Italians, and the presence of a genetic gradient across mainland Italy. By performing a genome-wide scan using the first four eigenvectors as independent outcomes regressed against each SNP used as a predictor, we found that the first four PCs strongly correlated with well-known loci under selective pressure, such as the *HLA-A* complex (*hg19*

chr6:g.21,266,925_32,628,428),[29] and the locus related to the lactase persistence phenotype (*hg19 chr2:g.135,907,088_137,013,606*);[30] and loci with a low recombination rate such as the *hg19 chr8:g.8,094,406_11,860,625* region harbouring a known polymorphic inversion in Europeans[31] (see Manhattan plot - Figure S1). In view of the above, we repeated the analysis excluding the above loci, obtaining the same pattern of variability across mainland Italy (Figure S2): the correlation between the first two PCs computed with and without these loci was higher than 0.95. ADMIXTURE analysis was also performed on the 300 Italians: see Supplementary material and Figure S3 for a description of the results.

Shared IBD haplotypes across Italy: we found greater sharing of IBD segments within regions than between regions from both the W and L statistics (Tables S2A-S2B), although in some cases the differences were not statistically significant (data not shown). The Mantel test showed a significant correlation between the geographical distance and the total length of shared IBD segments, both when including Sardinia ($R = -0.483$, $p = 0.0039$), and when excluding Sardinia from the analysis ($R = -0.622$, $p = 0.0027$). Based on Monmonier's algorithm, a genetic barrier was identified between Sardinia and mainland Italy (empirical $p < 0.0001$), whereas there was no evidence of any statistically significant genetic barrier within the peninsula.

A significant southward trend of increasing population size N_e (Figure S4A) was found (p for trend < 0.0001) when the mainland Italians were grouped according to the three main macro-areas: Northern, Central, and Southern. From single region results (Figure S4B), the lowest estimated ancestral N_e values were for Sardinia and the Aosta Valley (less than 5,000), accompanied by a high rate of inbreeding, whereas the effective population sizes were rather homogeneous across the remaining regions.

Comparison with neighbouring populations: We first used PCA (Figure S5) to investigate genetic differences across 35 populations from Europe, the Middle East and North Africa. The projection of the first two eigenvectors reflects well the geographical origins of the subjects included in the analysis (see Supplementary Results for details). We further investigated population structure using ADMIXTURE. Three major components are noticeable in the Italian population, with different proportions among the major Italian macro areas (see Supplementary Results and Figure S6 for a more detailed description).

The distribution of the pairwise F_{st} distances between all population pairs is shown in Table S3. The genetic distance between Southern and Northern Italians (F_{st} = 0.0013) is comparable to that between individuals living in different political units (i.e. Iberians-Romanians F_{st} = 0.0011; British-French F_{st} = 0.0007) and, interestingly, higher than 50% of all the possible pairwise comparisons within Europe (Figure S7).

Finally, when comparing IBD segment sharing between Italians and the other populations, both with total length W and the average length L, we observed that Southern Italians share more IBD with North African and Middle Eastern populations, whereas Northern Italians share more IBD with Europeans, as is predictable by geography (see Supplementary Results and Table S4 for details).

Time since admixture events: We estimated time since admixture events from linkage disequilibrium (LD) decay as a function of genetic distance.

We found evidence of the presence of a mix of Central-Northern European and Middle Eastern-North African ancestries in the Italian individuals (Table S5). The estimated times of admixture range between about 2,050 and 1,300 years ago (y.a.) with an average of about 1,650 y.a. — assuming 29 years per generation[32]— for Northern Italians, and

between about 3,000 and 1,450 y.a. (about 2,100 y.a. on average) for Central Italians. Finally, for the Southern Italian individuals, admixture between European and Northern African-Middle Eastern ancestry was estimated to have occurred about 1,000 y.a. (see Table S5 and Supplementary Results for a complete report of significant results).

Discussion

We evaluated fine-scale genetic differences within the Italian population using a large set of SNPs genotyped in more than 300 Italian individuals and compared them with published genotype data for 1,272 European, Middle Eastern, and North African individuals. Our study focused on only 11 of the twenty Italian regions and, in particular, lacks representation from the Eastern part of Italy. On the other hand, some of the strengths of this study are the sample selection criteria based on typical surnames and the place of birth of the four grandparents, which avoids the inclusion of individuals whose origins are different from their place of birth. In a previous study[1] we provided a first overview of the genetic composition of Italians, selecting individuals based on the place of birth only, but we were not able to discriminate between Northern and Central Italians. We observed that a proportion of individuals born in Northern Italy clustered with Southern Italians. This was explained by the internal migration that occurred during the last two generations, when people from Southern Italy left their place of origin looking for better economic opportunities in the North.

Genetic gradient across mainland Italy: Several of our analyses revealed that the genetic structure of Italians varies to a large extent but that it can be used to assign each mainland Italian to the correct macro-area of origin with very good sensitivity and specificity. The few misclassified individuals could be due to incorrect self-reported origin of mother or

grandparents from the maternal line (sampling according to typical surnames makes the paternal line more likely to be correct), or to different origins of the great-grandparents about whom we have no information. The PCA, IBD and ancestry analyses revealed a genetic gradient across the peninsula that correlates with geography. Since the diversity gradient in Italy remained after excluding loci under selective forces or with low recombination rate, we may speculate that historical events have a complementary role in explaining the great genomic variability within the Italian population. F_{st} genetic distance between Northern and Southern Italians is comparable or even higher than differences observed among individuals living in different countries, further confirming the high genomic variability within Italy. We also replicated the previously described gradient of SNP allelic frequencies on the *hg19 chr2:g.135,907,088_137,013,606* locus,[30] related to the lactase persistence phenotype, which strongly correlates with the second PC, and which in turn reflects the geographical location.

Continuous gene flow or different ancestral populations: We hypothesize two simple historical scenarios leading to the observed genetic variability across Italy: a) continuous ancient gene flow amplified by *isolation by distance* in recent times; b) different ancestral origins of the main Italian macro-areas whose distinguishability has been attenuated by genetic exchange in recent times.

Monmonier's algorithm revealed no evidence of the presence of genetic barriers across the peninsula. Instead, results from the Mantel test provide evidence of a correlation between genetics and geographical distance. The observed higher average length of the segments with shared IBD within regions compared to those shared between regions (Table S2B) suggests recent *isolation by distance* across the wide range of latitude of the Italian

peninsula. Moreover, a North to South gradient of increasing ancestral N_e was inferred for the three main macro-areas (Northern, Central, and Southern), coinciding with increased heterozygosity in Southern Italy. A similar trend was previously described for the rate of inbreeding and genome-wide similarity across Central Europe,[33] and could be interpreted as a signature of the 'Out of Africa' migration during Paleolithic, expansions from *refugia* after the ice age, and of ancient South-to-North migratory waves that occurred at the times of European colonization by Neolithic farmers. The ancestry and IBD analyses provided evidence of admixture in Italy with three major ancestries detected, most represented in Northern Europeans, Southern Europeans, and Middle Eastern respectively (with a small percentage of a North African component found in South Italy and Sardinia), with different prevalence across the peninsula. None of these components is fixed in any population, meaning that there is a poor fit with a strict admixture model, as assumed by the algorithm used, and supporting a process of continuous gene flow in multiple directions (migratory waves to and from Italy). According to previous studies on the Y chromosome and mtDNA,[34,35] the Middle Eastern ancestry in Southern Italians most likely originated at the time of the Greek colonization and, with a smaller percentage, of the subsequent Arabic domination;[7] whereas in Central-Northern Italy it is possibly due to the admixture of the indigenous residents with Middle Eastern populations spreading from the Caucasus to Central Europe.[19,21,28,36] Our results agree with previously published reports describing a possible maritime route of colonization across Europe, including Italy,[37] although we cannot exclude the occurrence of more recent demographic events leading to a similar scenario. Finally, the homogenous ancestral effective population size across Italian regions could be interpreted as reflecting common genetic origins, taking also into account previous

considerations, although the same results might also occur in comparing populations without common origins.

Our study supports the notion that genetic variability across Italy is likely to represent continuous gene flow leading to differences in the proportion of ancestry from different sources, along with genetic exchange among neighbouring populations (e.g. Northern Italian with European countries, Southern Italian with Middle Eastern and North African ones). Previous studies, analysing uniparental markers, found Y chromosome genetic discontinuity across Italy. This contrasts with a general lack of structure for mitochondrial DNA,[2,4] and with a higher homogeneity for maternal than paternal genetic contributions, suggesting different demographic and historical dynamics for females and males in Italy.

Sardinia: Among the eleven Italian regions investigated, Sardinia deserves a separate discussion. We replicated previously reported results showing the large genetic differences between Sardinians and mainland Italians;[29] the occurrence of a genetic barrier between Sardinia and the rest of Italy; the previously described similarity between the ancestries of the Tyrolean iceman and Sardinians;[36] and the very large allelic frequency differences between the islanders and the mainland Italians for the SNPs located on chromosome 6, in the *HLA-A* complex locus involved in the immune response,[38] which may at least partly explain the increased prevalence in the island of immune and autoimmune diseases, such as type 1 diabetes and multiple sclerosis.

The Aosta Valley region: The Aosta Valley is a small region at the North-Western Italian border with France and Switzerland. Its inhabitants showed interesting genetic characteristics. In the PCA (Figures 2B-2C, and Figure S5), Aostans do not cluster with subjects from the other Northern Italian regions, not even after the inclusion of non-Italian

populations in the analysis. Moreover, IBD analysis revealed a high level of inbreeding, comparable to the rate observed in Sardinia. However, the estimated proportions of ancestry are comparable to those in Piedmont, Liguria, Lombardy, and Emilia Romagna. Hence, our results suggest that the observed differences are not due to the effect of long genetic isolation, as is the case for Sardinians, but are the results of recent isolation of the Valley. These results are consistent with the low number of different surnames, mostly of French origin, compared to the number of families, as expected in genetic isolates.[39]

Time since admixture: The overall procedure to estimate time since admixture with the ALDER software is strongly conservative and is based on the assumption of the simplified model of admixture we used – i.e., a single pulse from discrete sources. Our previously discussed results favoured the hypothesis of continuous gene flow across Italy with admixture events that have likely occurred multiple times, and it should be noted that the method used is designed to emphasize the most recent admixture event.[28] Our estimated admixture dates agree with recent literature and several historical events. For example, our results support the hypothesis of an admixture event that occurred about 3,000 y.a. involving populations coming from the Caucasus, the Middle East, and populations that lived in Central Italy (Tuscany and Latium), as previously reported from analyses of mitochondrial and Y chromosome DNA,[40] and genome-wide data.[41] This was interpreted as possible evidence of the Middle Eastern (Anatolian) origin of the Etruscans (Herodotus' theory). Admixture events introducing Northern-Central European ancestry into Italy were estimated to have occurred during the so called 'Migration Period' after the Roman Empire collapsed (476 CE), with the consequent decline in population. After that, the 'Barbarian invasions' took place, with migratory waves from Northern-Central Europe to Northern-

Central Italy. It may be speculated that the estimated Northern-Central European ancestry in contemporary Italians is also the effect of subsequent Italian population growth, as previously reported by studies on mitochondrial and Y chromosome DNA from a genetic isolate in Northern Italy,[42] suggesting that Germanics (Lombards in particular) settled in Northern Italy during the 'Migration Period' and may have contributed to the foundation of some communities in Northern Italy. Finally, admixture events involving the Southern Italian population were inferred to have occurred about 1,000 y.a., coinciding with The Norman conquest of Southern Italy that spanned most of the 11th and 12th centuries and involved many battles and independent conquerors. A much more detailed analysis of geographically well-distributed samples from Southern Italy is required to validate our findings, while a large genetic contribution to the island of Sicily from Greece has previously been estimated.[43]

Conclusions and future directions: To our knowledge this is the first study to investigate genetic variability within the Italian population using a very large number of SNPs and subjects with well-defined geographical origin. We used these data to make inferences on population sub-structure and admixture events in Italy. To achieve a more complete picture of Italian genetic history and composition, future work should include Italian regions not covered in this study (Eastern regions such as Apulia and Veneto) and should compare Italians with other European populations (Greeks in particular). Our study could be useful for further genetic, epidemiological and forensic studies in Italy, as it may provide a set of valuable healthy controls for genome-wide association studies, and may be useful for identifying ancestral informative markers (AIMs). It might also help to explain the North-South prevalence gradient reported in Italy for several types of tumours.[44]

Supplementary data description: Supplementary data include nine figures and five tables. Supplementary information is available at EJHG's website.

Acknowledgments: The authors thank AVIS (Italian Association of Voluntary Blood Donors and specifically Dr. Domenico Giupponi), all the volunteers participating in this study, and Prof. Luca Cavalli-Sforza for his precious contribution to the study design from the early stages.

Funding. *Compagnia di San Paolo*, HuGeF, and the Stanford Center for Computational, Evolutionary and Human Genomics supported this study; the authors declare no conflict of interest.

Additional information. The list of all the genetics and epigenetics datasets accessible upon request at the HuGeF repository, and information to request data download authorization are available at the following web link: http://www.hugef-torino.org/site/index.php?id=286&t=articolo_secondo_livello&m=extra.

References

1. Di Gaetano C, Voglino F, Guarrera S *et al*: An overview of the genetic structure within the Italian population from genome-wide data. *PLoS One* 2012; **7:** e43759.

2. Boattini A, Martinez-Cruz B, Sarno S *et al*: Uniparental markers in Italy reveal a sex-biased genetic structure and different historical strata. *PLoS One* 2013; **8:** e65441.

3. Brisighelli F, Alvarez-Iglesias V, Fondevila M *et al*: Uniparental markers of contemporary Italian population reveals details on its pre-Roman heritage. *PLoS One* 2012; **7:** e50794.

4. Sarno S, Boattini A, Carta M *et al*: An ancient Mediterranean melting pot: investigating the uniparental genetic structure and population history of sicily and southern Italy. *PLoS One* 2014; **9:** e96074.

5. Piazza A, Cappello N, Olivetti E, Rendine S: A genetic history of Italy. *Ann Hum Genet* 1988; **52:** 203-213.

6. Cavalli-Sforza LL, Menozzi P, Piazza A: *The History and Geography of Human Genes*. Princeton: Princeton Universty Press, 1994.

7. O'Dushlaine C, McQuillan R, Weale ME *et al*: Genes predict village of origin in rural Europe. *Eur J Hum Genet* 2010; **18:** 1269-1270.

8. Di Giacomo F, Luca F, Anagnou N *et al*: Clinal patterns of human Y chromosomal diversity in continental Italy and Greece are dominated by drift and founder effects. *Mol Phylogenet Evol* 2003; **28:** 387-395.

9. Barbujani G, Bertorelle G, Capitani G, Scozzari R: Geographical structuring in the mtDNA of Italians. *Proc Natl Acad Sci U S A* 1995; **92:** 9171-9175.

10. Lao O, Lu TT, Nothnagel M *et al*: Correlation between genetic and geographic structure in Europe. *Curr Biol* 2008; **18:** 1241-1248.

11. Novembre J, Johnson T, Bryc K *et al*: Genes mirror geography within Europe. *Nature* 2008; **456:** 98-101.

12. Seedat YK, Brewster LM: What role does African ancestry play in how hypertensive patients respond to certain antihypertensive drug therapy? *Expert Opin Pharmacother* 2014; **15:** 159-161.

13. Shriner D, Adeyemo A, Ramos E, Chen G, Rotimi CN: Mapping of disease-associated variants in admixed populations. *Genome Biol* 2011; **12:** 223.

14. Lawson DJ, Hellenthal G, Myers S, Falush D: Inference of population structure using dense haplotype data. *PLoS Genet* 2012; **8:** e1002453.

15. Leslie S, Winney B, Hellenthal G *et al*: The fine-scale genetic structure of the British population. *Nature* 2015; **519:** 309-314.

16. Zei G, Barbujani G, Lisa A *et al*: Barriers to gene flow estimated by surname distribution in Italy. *Ann Hum Genet* 1993; **57:** 123-140.

17. Boattini A, Lisa A, Fiorani O, Zei G, Pettener D, Manni F: General method to unravel ancient population structures through surnames, final validation on Italian data. *Hum Biol* 2012; **84:** 235-270.

18. Browning SR, Browning BL: Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet* 2007; **81:** 1084-1097.

19. Ralph P, Coop G: The geography of recent genetic ancestry across Europe. *PLoS Biol* 2013; **11:** e1001555.

20. Atzmon G, Hao L, Pe'er I *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* 2010; **86:** 850-859.

21. Botigue LR, Henn BM, Gravel S *et al*: Gene flow from North Africa contributes to differential human genetic diversity in southern Europe. *Proc Natl Acad Sci U S A* 2013; **110:** 11791-11796.

22. Mantel N: The detection of disease clustering and a generalized regression approach. *Cancer Res* 1967; **27:** 209-220.

23. Manni F, Guerard E, Heyer E: Geographic patterns of (genetic, morphologic, linguistic) variation: how barriers can be detected by using Monmonier's algorithm. *Hum Biol* 2004; **76:** 173-190.

24. Palamara PF, Pe'er I: Inference of historical migration rates via haplotype sharing. *Bioinformatics* 2013; **29:** i180-188.

25. Falush D, Stephens M, Pritchard JK: Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Mol Ecol Notes* 2007; **7:** 574-578.

26. Maples BK, Gravel S, Kenny EE, Bustamante CD: RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *Am J Hum Genet* 2013; **93:** 278-288.

27. Clayton D: snpStats: SnpMatrix and XSnpMatrix classes and methods. *R package version 1160* 2014.

28. Moorjani P, Patterson N, Hirschhorn JN *et al*: The history of African gene flow into Southern Europeans, Levantines, and Jews. *PLoS Genet* 2011; **7:** e1001373.

29. Di Gaetano C, Fiorito G, Ortu MF *et al*: Sardinians genetic background explained by runs of homozygosity and genomic regions under positive selection. *PLoS One* 2014; **9:** e91237.

30. Curry A: Archaeology: The milk revolution. *Nature* 2013; **500:** 20-22.

31. Ma J, Amos CI: Investigation of inversion polymorphisms in the human genome using principal components analysis. *PLoS One* 2012; **7:** e40224.

32. Fenner JN: Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. *Am J Phys Anthropol* 2005; **128:** 415-423.

33. Abdellaoui A, Hottenga JJ, de Knijff P *et al*: Population structure, migration, and diversifying selection in the Netherlands. *Eur J Hum Genet* 2013; **21:** 1277-1285.

34. Francalacci P, Morelli L, Underhill PA *et al*: Peopling of three Mediterranean islands (Corsica, Sardinia, and Sicily) inferred by Y-chromosome biallelic variability. *Am J Phys Anthropol* 2003; **121:** 270-279.

35. Semino O, Magri C, Benuzzi G *et al*: 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. *Am J Hum Genet* 2004; **74:** 1023-1034.

36. Sikora M, Carpenter ML, Moreno-Estrada A *et al*: Population genomic analysis of ancient and modern genomes yields new insights into the genetic ancestry of the Tyrolean Iceman and the genetic structure of Europe. *PLoS Genet* 2014; **10:** e1004353.

37. Paschou P, Drineas P, Yannaki E *et al*: Maritime route of colonization of Europe. *Proc Natl Acad Sci U S A* 2014; **111:** 9211-9216.

38. Gough SC, Simmonds MJ: The HLA Region and Autoimmune Disease: Associations and Mechanisms of Action. *Curr Genomics* 2007; **8:** 453-465.

39. Berton R: *Anthroponomie Valdôtaine*. Quart - Aoste: Musumeci, 1988.

40. Brisighelli F, Capelli C, Alvarez-Iglesias V *et al*: The Etruscan timeline: a recent Anatolian connection. *Eur J Hum Genet* 2009; **17:** 693-696.

41. Pardo-Seco J, Gomez-Carballa A, Amigo J, Martinon-Torres F, Salas A: A genome-wide study of modern-day Tuscans: revisiting Herodotus's theory on the origin of the Etruscans. *PLoS One* 2014; **9:** e105920.

42. Boattini A, Sarno S, Pedrini P *et al*: Traces of medieval migrations in a socially stratified population from Northern Italy. Evidence from uniparental markers and deep-rooted pedigrees. *Heredity (Edinb)* 2014.

43. Di Gaetano C, Cerutti N, Crobu F *et al*: Differential Greek and northern African migrations to Sicily are supported by genetic evidence from the Y chromosome. *Eur J Hum Genet* 2009; **17:** 91-99.

44. Baili P, De Angelis R, Casella I *et al*: Italian cancer burden by broad geographical area. *Tumori* 2007; **93:** 398-407.

Title and Legend to Figures

Figure 1. Sampling location barycentre (average sampling place weighted by the number of individuals from each sampling point) and sample size for each of the eleven Italian regions analysed in this study (Latitude and Longitude position in brackets): Aosta Valley (45°70'N - 7°30'E); Basilicata (40°69'N - 16°57'E); Calabria (38°10'N - 15°70'E); Emilia Romagna (44°80'N - 11°60'E); Latium (42°40'N - 12°10'E); Liguria (44°30'N - 8°50'E; Lombardy (45°39'N - 9°85'E); Piedmont (45°19'N - 7°90'E); Sardinia (40°22'N - 9°30'E); Sicily (37°41'N - 13°72'E); Tuscany (43°31'N - 11°35'E).

Figure 2. Heatmap representing the co-ancestry matrix indicating the number of genomic segments inherited from the same ancestral populations for each pair of samples: Dendrogram based on hierarchical clustering (at the top), and pie charts representing the overlap between inferred and self-reported origin of the Italian individuals **(A)**. Principal Component Analysis based on the same co-ancestry matrix including Sardinians **(B)** and excluding Sardinians **(C)**; x- and y-axes were inverted to emphasize similarity to the geographical map of Italy.


Aosta Valley
28
30
35
Lombardy
Piedmont
Emilia Romagna
30
24
Liguria
Tuscany
21
29
Latium
28
Basilicata
31
Sardinia
12
Calabria
32
Sicily

Fig.1

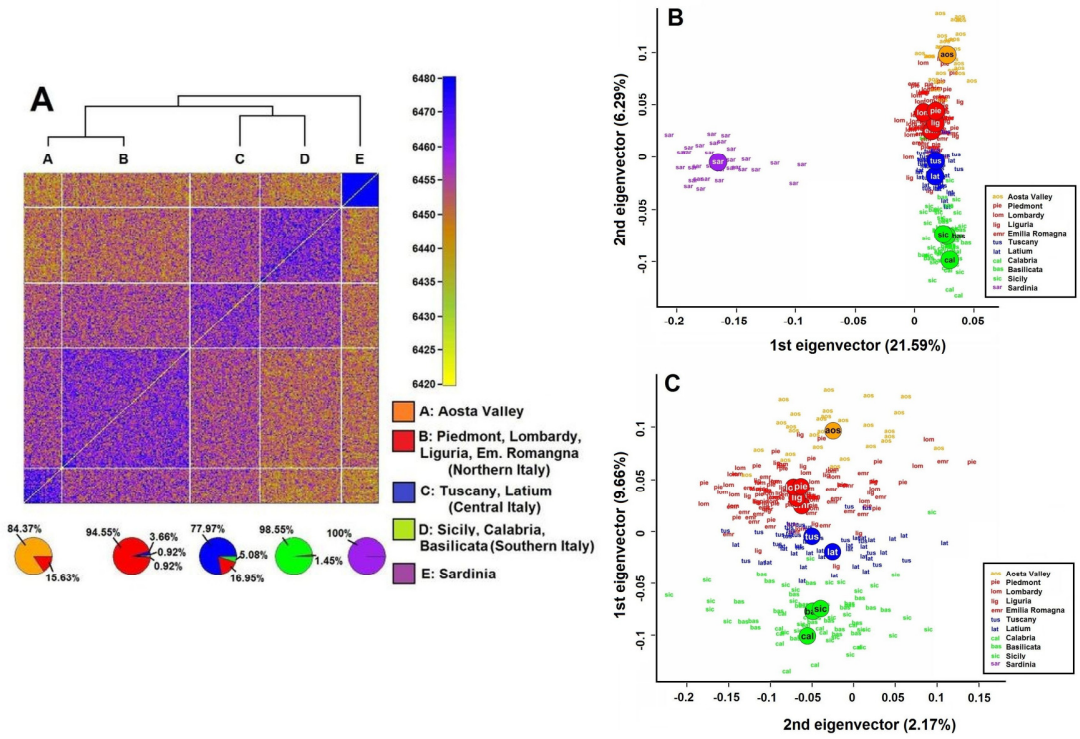
A
A
B
C
D
E
6480
6470
6460
6455
6450
6440
6435
6430
6425
6420
A: Aosta Valley
B: Piedmont, Lombardy, Liguria, Em. Romangna (Northern Italy)
C: Tuscany, Latium (Central Italy)
D: Sicily, Calabria, Basilicata (Southern Italy)
E: Sardinia
84.37%
15.63%
94.55%
3.66%
0.92%
0.92%
77.97%
5.08%
16.95%
98.55%
1.45%
100%
B
2nd eigenvector (6.29%)
1st eigenvector (21.59%)
C
1st eigenvector (9.66%)
2nd eigenvector (2.17%)
Aosta Valley
Piedmont
Lombardy
Liguria
Emilia Romagna
Tuscany
Latium
Calabria
Basilicata
Sicily
Sardinia

Fig.2