A genome-wide portrait of Italy

A. Raveane, F. Montinaro, H. Lancioni, A. Mulas, V. Grugni, I. Cardinali, M. Zoledziewska, S. Aneli, A. Baali, S. Barlera, et al.

A genome-wide portrait of Italy

A. Raveane[1,2], F. Montinaro[2], H. Lancioni[3], A. Mulas[4], V. Grugni[1], I. Cardinali[3], M. Zoledziewska[4], S. Aneli[5,6], A. Baali[7], S. Barlera[8], G. Boncoraglio[9], F. Brisighelli[10], AM. Di Blasio[11], M. Cherkaoui[7], C. Di Gaetano[5,6], JM. Dugoujon[12], S. Guerrera[5,6], T. Kivisild[13] , M. Melhaoui[14], L. Pagani[13], S. Parolo[15], P. Paschou[16] , A. Piazza[5,6], V. Pascali[17], M. Peyret-Guzzon[18], F. Ricaut[19], G. Stamatoyannopoulos[20], F. Cucca[21], A. Angius[21], A. Torroni[1], M. Metspalu[22], O. Semino[1], G. Hellenthal[23], G. Matullo[5,6]*, A. Achilli[1]*, A. Olivieri[1]*, C. Capelli[2]*

1Dept of Biology and Biotechnology "L. Spallanzani", University of Pavia, Pavia, Italy; 2Dept of Zoology, Oxford University, South Parks Road, Oxford OX1 3PS, UK; 3Dept of Chemistry, Biology and Biotechnology, University of Perugia, Perugia, Italy; 4Istituto di Ricerca Genetica e Biomedica (IRGB), CNR, Lanusei, Italy; 5Dept of Medical Sciences, University of Turin, Turin, Italy; 6HuGeF Human Genetics Foundation, Turin, Italy; 7Laboratoire d'Ecologie Humaine Faculté des Sciences Semlalia, Marrakech Maroc; 8Dept of Cardiovascular Research, IRCCS Mario Negri Institute for Pharmacological Research, Milan, Italy; 9Dept of Cerebrovascular Diseases, IRCCS Istituto Neurologico Carlo, Besta, Milan, Italy; 10Catholic University of the Sacred Heart, Institute of Legal Medicine and Insurance, Rome, Italy; 11Molecular Biology Laboratory, Istituto Auxologico Italiano, Milan, Italy; 12Centre d'Anthropologie, UMR 8555, Toulouse, France; 13Division of Biological Anthropology, University of Cambridge, Cambridge, UK; 14Département de Biologie Faculté des Sciences, Oujda Maroc; 15Computational Biology Unit, Institute of Molecular Genetics-National Research Council, Pavia, Italy; 16Department of Molecular Biology and Genetics, Democritus University of Thrace, 68100 Alexandroupolis, Greece; 17Institute of Public Health, Section of Legal Medicine, Università Cattolica del Sacro Cuore, Rome, Italy; 18University of Oxford, Wellcome Trust for Human Genetics, Oxford; 19Laboratoire d'Anthropologie Moléculaire et Imagérie de Synthèse UMR-5288, Université de Toulouse, Toulouse, France; 20Departments of Medicine and Genome Sciences, University of Washington, Seattle, WA 98195; 21Istituto di Ricerca Genetica e Biomedica (IRGB), CNR, Monserrato, Italy; 22Dept of Evolutionary Biology, Estonian Biocentre and University of Tartu, Tartu, Estonia; 23UCL Genetics Institute, University College London, Gower Street, London, UK

e-mail: alessandro.raveane01@universitadipavia.it
*Co-senior authors

P16.13

OVERVIEW

Surrounded by the sea and bounded by the Alps, Italy extends over more than 1,000 km along a North-South axis and comprises the two largest islands of the Mediterranean, Sicily and Sardinia. The combination of this geographic complexity with a rich set of historical events and cultural dynamics has the potential to shape in a unique way the distribution of genetic variation within the Italian population. Recent investigations have reported substantial stratification in Italy when compared to other European countries, but a fine and exhaustive characterisation of its population structure and admixture history has yet to be conducted.
In order to dissect the fine structure and the admixture profile of Italian populations, we genotyped 167 novel samples with the Illumina Infinium Omni2.5 BeadChip and assembled a comprehensive genome-wide SNP dataset which included almost 1,500 individuals representing all of the 20 Italian administrative regions, and data from ~ 300 world-wide reference populations. Preliminary results based on parametric and non-parametric statistical analyses suggest extensive population structure and a complex pattern of admixture episodes over the last few thousand years.

THE ITALIAN GENETIC STRUCTURE

Our analysis has the aim to detect different levels of structure in the Italian population. Principal components (PCA) and ADMIXTURE analyses have been performed using allelic frequencies.
Fig. 1 shows the PCA of 1,588 individuals belonging to all the 20 administrative regions of Italy. In line with recent investigations [1]-[3], a well defined North-South cline is evident comprising all Sicilian and continental samples spread from the bottom left towards the upper right of the plot. Sardinian samples are separated from the rest of the samples, as the result of their mostly Neolithic ancestry [4] and long term isolation.
Discordances between geographic origin and position in the plot observed for a few samples (i.e the clustering of Piedmont samples (ITN-PIE) togheter with South-Italy clusters (ITS-)) are probably due to different methods of sampling. We plan to explore this aspect, in order to assess recent migration patterns in the country. Similarly, samples in intermediate positions (i.e. IT-U between ITS-SAR and the other samples), might represent admixed subjects.

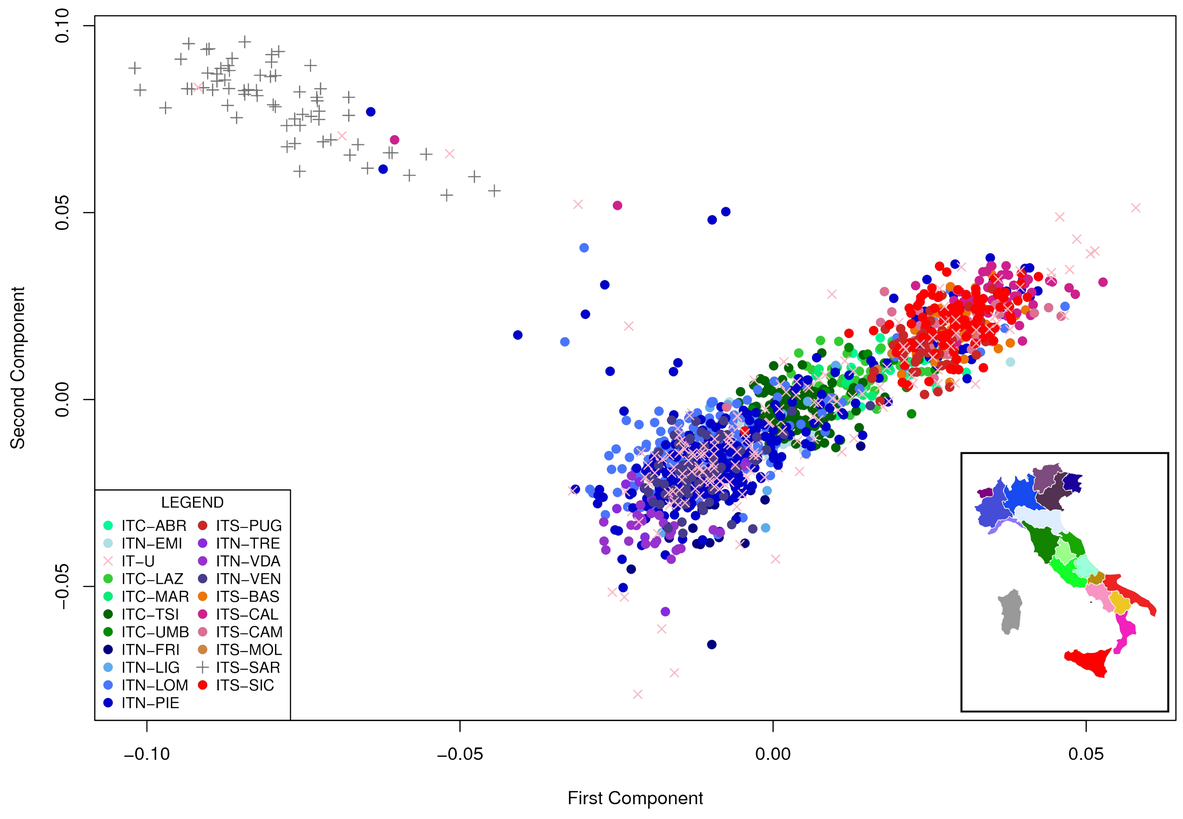

Figure 1. Genome-wide principal component analysis of the Italian samples. Each individual is depicted by a symbol and a color representing the administrative region where he/she was collected . The labels for the regions are as follows: ITN-VDA, Valle D'Aosta; ITN-PIE, Piedmont; ITN-LOM, Lombardy; ITN-TRE, Trentino Alto Adige; ITN-VEN, Veneto; ITN-FRI, Friuli Venezia Giulia; ITN-EMI, Emilia Romagna; ITC-TSI, Tuscany; ITC-UMB, Umbria; ITC-MAR, Marche; ITC-LAZ, Lazio; ITC-ABR, Abruzzo; ITS-MOL, Molise; ITS-CAM, Campania; ITS-PUG, Apulia; ITS-BAS, Basilicata; ITS-CAL, Calabria; ITS-SIC, Sicily; ITS-SAR, Sardinia; IT-U, Unassigned (samples with mexed/unknown origins). The inset map provides a key to the labels.

ADMIXTURE EVENTS IN ITALY

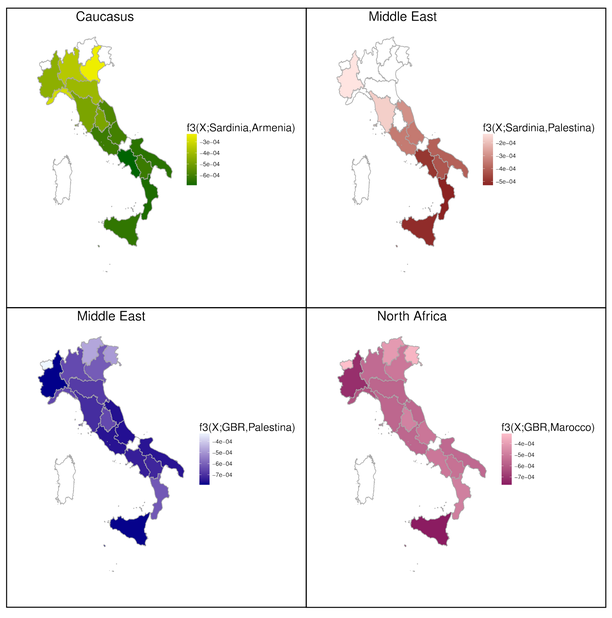

Figure 2. Distribution of the f3 statistic in Italy: the density of the colours displays the degree of admixture that a population X (Italian region) had with the two source populations A and B (the legend of each image displays the populations involved and, the main titles indicate the most likely source region). The Middle East region has been identified in relation to two other sources, one with a North-European (GBR) component and another with a South-European component (Sardinia).

In order to detect the signature of admixture events in the Italian population we performed the f3 analyses in the form f3(X; A, B), in which A and B are represented by an European and a non-European source [5].
Briefly, the value f3(X; A, B) is negative when X does not fit onto a simple tree with A and B, but it is the result of admixture between the two sources. We considered f3 significantly lower than zero when the Z score associated was < -5.

Virtually all the Italian populations show signature of gene flow involving at least one extra European population.
The most significant admixture signals are shown in Fig. 2 and they involve five main sources: three from outside Europe, the Caucasus (Armenia), Middle East (Palestine) and North Africa (Morocco) and two Europeans, possibly associated to the North-South differentiation also highlighted in Europe differentiating pre-Neolithic and Neolithic populations [6].
The results align a North-South cline. This is particularly evident when the Caucasus and Middle East f3 value distributions are considered, possibly reflecting a similar temporal and geographic source for these events.
On the other hand, the signal for admixture involving North African populations is stronger in Sicily, likely due to the Arab occupation of the region between the IX and XI century.
Piedmont is characterized by negative f3-values, comparable to those of Sicily and in striking contrast with neighbouring populations.
This is probably due to the fact that only for a subset of the individuals from Piedmont, the origin of the grandparents was known (and from same region), and that Piedmont was one of the region most affected by migration from Southern Italy in the second half of the 20th century. The large number of individuals from Piedmont and Lombardy included in the analyses will allow us to assess this important socio-cultural phenomenon from a genetic perspective.

THE ITALIAN GENETIC VARIABILITY IN THE EUROPEAN AND WORLDWIDE CONTEXT

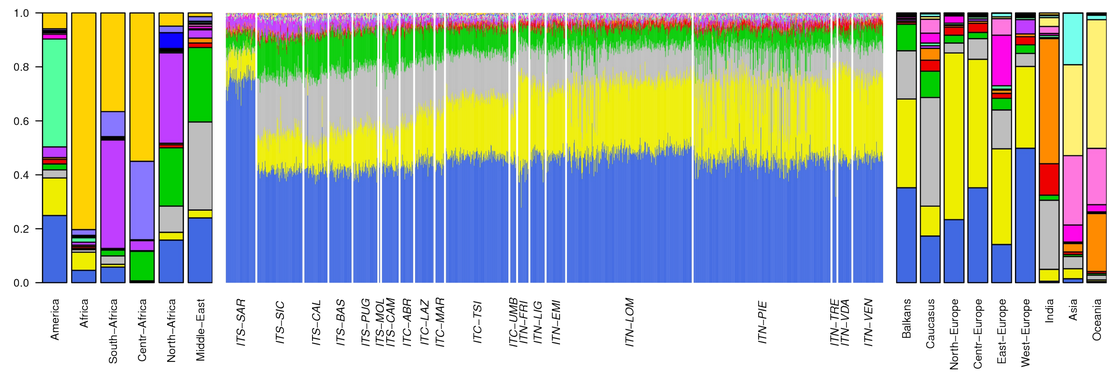

Figure 3. ADMIXTURE plot of individuals and populations at K = 15. For the Italian population we reported the reference regions and display results for each individual, represented as a thin vertical bar, for the other populations the average across individuals is presented. Samples from Americas include individuals of admixed ancestry.

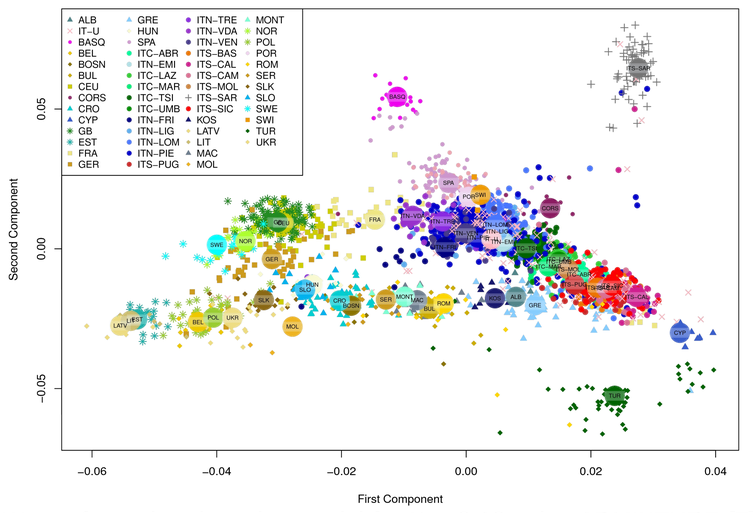

Figure 4. PC analysis of genetic data from 1,588 Italians and 983 Europeans. Small coloured symbols represent individuals and large coloured points represent median PC1 and PC2 values for each country. ALB, Albania; BASQ, Baschi; BEL, Belgium; BOSN, Bosnia-Herzegovina; BUL, Bulgaria; CEU, Central Europe; CORS, Corsica; CRO, Croatia; CYP, Cyprus; GB, United Kingdom; EST, Estonia; FRA, France; GER, Germany; GRE, Greece; HUN, Hungary; SPA, Spain; KOS, Kosovo; LATV, Latvia; LIT, Lithuania; MAC, Macedonia; MOL, Moldova; MONT, Montenegro; NOR, Norway; POL, Polonia; POR, Portugal; ROM, Romania; SER, Serbia; SLK, Slovakia; SLO, Slovenia; SWE, Sweden; SWI, Switzerland; TUR, Turkey; UKR, Ukraine; for the Italian populations the same scheme as in Fig. 1 is used.

We performed a clustering model for K = 2-18 using ADMIXTURE [7] on 1,588 individuals belonging to the 20 italian regions combined with ~300 worldwide populations. The results are shown in Fig. 3.
We identified five major ancestral components: blue, yellow, gray, green and violet.
Overall, they are distributed homogeneously within regions, along a North-South cline for entire Italian peninsula, with the exception of Sardinia.
The blue component is present in all the Italian regions with high frequencies in Sardinia remarking its genetic peculiarity related to its isolation history and ancestry.
The yellow component has high frequencies in the regions of Northern Italy, decreasing values in the Central regions, until reaching low frequencies in Southern Italy. Its distribution is modal in Northern and Central Europe reflecting their genetic affinity with neighbouring regions, as already reported in our f3 statistics analysis.
An opposite distribution is evident in the violet component, in which highest frequencies are observed in the South with decreasing values in Northern Italy. This component is found AT high frequencies in North Africa and might represent the legacy of the Arab rule in Southern Italy.
Finally, the grey and the green components, modal in Caucasian and Middle Eastern groups, could be the results of recent interactions between Italy and Eastern Europe or Western Asia.

The PC analysis in Fig. 4 is a summary of the genetic variability of Italy compared to the one of the Europe. Italy is stretched between the Mediterranean area (Cyprus-CYP) and West/Central Europe (Spain-SPA and France-FRA) with high affinity with Portugal, Switzerland and Corsica.
Interestingly, we noticed a partial overlap between France and Valle d'Aosta, while individuals from Friuli Venezia Giulia are spread toward the Central Balkanic area.
The Sardinian and Basque are well-known examples of genetically differentiated populations [8] and diverge clearly from the rest of the continental samples in our plot.
Similarly to the distribution of the Italian populations, a North-South cline of individuals from the Balkanic regions is observed, with populations in the South and North closer to Southern and Central Europe,respectively.

SAMPLES AND METHODS

We collected and genotyped 167 novel Italian samples (for whom all four grandparents were born in the same Italian administrative region) and assembled a dataset composed by 4,688 samples from more than 300 worldwide populations (1,588 of these from the 20 Italian regions). All of these samples were analysed with either Infinium Omni 2.5-8, Human OmniExpress, Human660W-Quad, or Human610-Quad chip.

We used Plink1.9 [9] for data merging and quality control pruning, filtering missing data for both genotypes (missing data higher than 0.02) and individuals (higher than 0.01), with a total of 218,725 SNPs remaining for the analysis involving the whole dataset.
The f3 statistics were calculated with the threepop program implemented in TreeMix [10] with -k 500. f3 significant values (Z score <-5) were summarized as a choropleth map.
Principal Component Analyses (PCA) was performed using Plink1.9 as a non-parametric assessment of genetic clustering using both the Italian samples and European subset. We carried out ADMIXTURE analysis of the whole dataset after pruning for linkage disequilibrium in PLINK with parameters --indep-pairwise 200 25 0.4, which retained 144,553 SNPs.

FUTURE PERSPECTIVES

Our preliminary analyses confirmed the separation of Sardinia from the rest of Italy, while a more continuum distribution of genetic variation along a North-South axis was observed for the remaining samples. In the future we plan to further characterize this pattern exploring the related genetic structure by identifying clusters of homogeneous individuals and investigating their admixuture history as well as by including in the analyses data from ancient genomes.

REFERENCES

[1] Parolo et al. (2015). "Characterization of the biological processes shaping the genetic structure of the Italian population"; BMC Genetics. 16: 132.
[2] Fiorito et al.(2015). "The Italian genome reflects the history of Europe and the Mediterranean basin"; European Journal of Human Genetics. 24: 1056-1062.
[3] Sazzini et al. (2016). "Complex interplay between neutral and adaptive evolution shaped differential genomic background and disease susceptibility along the Italian peninsula"; Scientific Reports. 6: 32513.
[4] Lazaridis et al. (2016). "Genomic insights into the origin of farming in the ancient Near East"; Nature. 536: 419–424.
[5] Reich et al. (2009). "Reconstructing Indian population history"; Nature. 461: 489-494.
[6] Skoglund et al. (2012). "Origins and genetic legacy of Neolithic farmers and hunter-gatherers in Europe"; Science. 336: 466-469.
[7] Alexander et al. (2009). "Fast model-based estimation of ancestry in unrelated individuals" Genome Research. 19: 1655-1664.
[8] November et al. (2008). "Genes mirror geography within Europe". Nature. 455: 861.
[9] Pickrell and Pritchard (2012). "Inference of population splits and mixtures from genome-wide allele frequency data". PLoS Genetics. 8: e1002967.
[10] Chang et al. (2015). "Second-generation PLINK: rising to the challenge of larger and richer datasets". Gigascience 4: 1.