Linguistic, geographic and genetic isolation: a collaborative study of Italian populations

Marco Capocasa2,3, Paolo Anagnostou1,2, Valeria Bachis4, Cinzia Battaglia1, Stefania Bertoncini5, Gianfranco Biondi6, Alessio Boattini7, Ilaria Boschi8, Francesca Brisighelli1,9, Carla Maria Calò4, Marilisa Carta7, Valentina Coia10,11, Laura Corrias4, Federica Crivellaro12, Sara De Fanti7, Valentina Dominici1,2, Gianmarco Ferri13, Paolo Francalacci14, Zelda Alice Franceschi15, Donata Luiselli7, Laura Morelli14, Giorgio Paoli5, Olga Rickards16, Renato Robledo17, Daria Sanna14, Emanuele Sanna4, Stefania Sarno7, Luca Simeo18, Luca Taglioli5, Giuseppe Tagarelli19, Sergio Tofanelli5, Giuseppe Vona4, Davide Pettener7 & Giovanni Destro Bisol1,2

1) Sapienza Università di Roma, Dipartimento di Biologia Ambientale, Roma, Italy
2) Istituto Italiano di Antropologia, Roma, Italy
3) Sapienza Università di Roma, Dipartimento di Biologia e Biotecnologie “Charles Darwin”, Roma, Italy
4) Università di Cagliari, Dipartimento di Scienze della Vita e dell’Ambiente, Cagliari, Italy
5) Università di Pisa, Dipartimento di Biologia, Pisa, Italy
6) Università dell’Aquila, Dipartimento di Scienze Ambientali, L’Aquila, Italy
7) Università di Bologna, Dipartimento di Scienze Biologiche, Geologiche e Ambientali, Bologna, Italy
8) Università Cattolica del Sacro Cuore, Istituto di Medicina Legale, Roma, Italy
9) Universidade de Santiago de Compostela, Instituto de Ciencias Forenses, Santiago de Compostela, Spain
10) Università di Trento, Dipartimento di Filosofia, Storia e Beni Culturali, Trento, Italy
11) Accademia Europea di Bolzano, Istituto per le Mummie e l’Iceman, Bolzano, Italy
12) Sezione di Antropologia, Museo Nazionale Preistorico Etnografico “Luigi Pigorini”, Roma, Italy
13) Università di Modena e Reggio Emilia, Dipartimento di Medicina Diagnostica, Clinica e di Sanità Pubblica, Modena, Italy
14) Università di Sassari, Dipartimento di Scienze della Natura e del Territorio, Sassari, Italy
15) Università di Bologna, Dipartimento di Discipline Storiche, Antropologiche e Geografiche, Bologna, Italy
16) Università degli Studi di Roma “Tor Vergata”, Centro di Antropologia Molocolare per lo studio del DNA Antico, Dipartimento di Biologia, Roma, Italy
17) Università di Cagliari, Dipartimento di Scienze Biomediche, Cagliari, Italy
18) Università di Palermo, Dipartimento di Biologia Ambientale e Biodiversità, Palermo, Italy
19) Istituto di Scienze Neurologiche, CNR, Mangone, Cosenza, Italy

§ Present affiliation

e-mail: marco.capocasa@uniroma1.it; davide.pettener@unibo.it; destrobisol@uniroma1.it

doi: 10.4436/JASS.92001
Summary - The animal and plant biodiversity of the Italian territory is known to be one of the richest in the Mediterranean basin and Europe as a whole, but does the genetic diversity of extant human populations show a comparable pattern? According to a number of studies, the genetic structure of Italian populations retains the signatures of complex peopling processes which took place from the Paleolithic to modern era. Although the observed patterns highlight a remarkable degree of genetic heterogeneity, they do not, however, take into account an important source of variation. In fact, Italy is home to numerous ethno-linguistic minorities which have yet to be studied systematically. Due to their difference in geographical origin and demographic history, such groups not only signal the cultural and social diversity of our country, but they are also potential contributors to its bio-anthropological heterogeneity. To fill this gap, research groups from four Italian Universities (Bologna, Cagliari, Pisa and Roma Sapienza) started a collaborative study in 2007, which was funded by the Italian Ministry of Education, University and Research and received partial support by the Istituto Italiano di Antropologia. In this paper, we present an account of the results obtained in the course of this initiative. Four case-studies relative to linguistic minorities from the Eastern Alps, Sardinia, Apennines and Southern Italy are first described and discussed, focusing on their micro-evolutionary and anthropological implications. Thereafter, we present the results of a systematic analysis of the relations between linguistic, geographic and genetic isolation. Integrating the data obtained in the course of the long-term study with literature and unpublished results on Italian populations, we show that a combination of linguistic and geographic factors is probably responsible for the presence of the most robust signatures of genetic isolation. Finally, we evaluate the magnitude of the diversity of Italian populations in the European context. The human genetic diversity of our country was found to be greater than observed throughout the continent at short (0-200 km) and intermediate (700-800 km) distances, and accounted for most of the highest values of genetic distances observed at all geographic ranges. Interestingly, an important contribution to this pattern comes from the “linguistic islands” (e.g. German speaking groups of Sappada and Luserna from the Eastern Italian Alps), further proof of the importance of considering social and cultural factors when studying human genetic variation.

Keywords – Genetic structure, Linguistic diversity, Minority languages, Linguistic islands.

Introduction

The plant and animal biodiversity found on Italian territory is among the richest in the Mediterranean basin and in Europe as a whole. This depends on the presence of different biomes, with Alpine tundra and the Mediterranean arid zones at the extremes, whose variety may be seen as a consequence of the natural role of Italy as a bridge connecting central European and North African environments (Blasi et al., 2005). As has happened throughout the continent, biodiversity has been influenced by human activities, such as cattle breeding and the agricultural exploitation of the lands since the Neolithic (Goudie, 2013). However, the remarkable ethno-cultural diversity of human populations which have settled on Italian territory since prehistory has led to different ways of managing natural resources (Padovani et al., 2009). This has probably reduced the loss of biodiversity produced by man-driven modifications of natural ecosystems to some extent. The latter points introduce us to another important aspect, which is more directly concerned with the anthropological dimension, i.e. biodiversity of human populations.

Looking at biodiversity from a holistic perspective, it is worth raising the question: does the great variety observed in Italy for plants and animals hold for human populations? Since prehistory, numerous peopling events have occurred on Italian territory. During the Last
Glacial Maximum, most of the country provided a refugium not only for animal and plant species (e.g. see Taberlet et al., 1998; Petit et al., 2003; Grassi et al., 2009), but also for human groups (Banks et al., 2008). Migrations during the Late Paleolithic, the Neolithic and, even more so, the Metal Ages characterized the complex peopling process of pre-historic Italy, leaving more (e.g. Greeks; see Scheidel, 2003) or less (e.g. Etruscans; see Barker & Rasmussen, 1998; Beekes, 2003) clear signatures. Under Roman rule, there were heavy demographic reshuffles caused by warfare and slavery and this had a significant impact on population composition (Hin, 2013). In more recent times, further contribution to the population heterogeneity came from invasions of the Barbarians (Lombards and Normans among others; see Jörg, 2002; Donald, 2008) and Arabs (Aghlabids; see Metcalfe, 2009). All these events explain at least part of the genetic structure of extant Italian populations, structured in two main geographic clusters (North-West and South-East) for Y-chromosome markers but more homogeneous for mtDNA polymorphisms (Boattini et al., 2013). However, the investigations carried out so far have yet to provide an exhaustive picture of the genetic diversity of Italian populations, since they could not take into account another important component of the human biodiversity of our country. We are referring to the historic ethno-linguistic minorities which total about 5% of the population today. They differ according to area of origin (central and southeastern Europe) and demographic history, and the majority of them settled in their present locations in relatively recent times (from the Middle Ages to the 18th century). The linguistic isolation from neighboring populations and, in many cases, the settlement as small communities in secluded areas make such groups very important when evaluating the entire spectrum of biodiversity of Italian populations. At present, twelve ethno-linguistic minorities are officially safeguarded by Italian legislation: Albanian, Catalan, Croatian, French, Franco-provençal, Friulian, German, Greek, Ladin, Occitan, Slovene and Sardinian (Toso, 2008). Furthermore, other minor linguistic groups are recognized at regional level (e.g. Tabarkian in Sardinia) and ancient linguistic substrates can be still recognized in marginal areas.

The interest in ethnic minorities and local communities is considerable in the present-day international debate concerning the preservation of cultural diversity and its influence on the biodiversity in the globalized world. Since 1999, the United Nations Environment Programme (UNEP) and the United Nations Educational, Scientific and Cultural Organization (UNESCO) have turned their attention to the “indigenous” peoples for the assessment of global biodiversity. UNEP focused on the relationships between biodiversity and cultural diversity with the publication of the report Cultural and Spiritual Values of Biodiversity (Addison Posey, 1999). After having defined Cultural Diversity as the “common heritage of humanity” (UNESCO, 2001), UNESCO has recently pointed out to the risk of extinction of 2500 languages, among which most of the minority languages spoken in Italy (Moseley, 2010). The message emerging from these initiatives is that paying attention to ethno-linguistic minorities is important not only for their contribution to the overall human biodiversity, but also for their role in maintaining cultural traditions which have an impact on biodiversity. This seems to expand the view, shared by many evolutionary Anthropologists, that biological and cultural diversity should be studied together to understand how the past has shaped the present of human species: “The subjects are (on the one hand) data, and (on the other) the cultural history surrounding the collection and interpretation of those data” (Marks, 1995).

In order to achieve a more complete characterization of the genetic structure of Italian populations, four research units of four Italian Universities (Bologna, Cagliari, Pisa and Roma Sapienza) initiated a collaborative research on linguistic minorities in 2007. During the realization of the project, the working group started further collaborations with researchers from other Italian Universities. A first part of work was done in the framework of the project “Isolating the isolates:
geographic and cultural factors of human genetic variation” supported by the Italian Ministry of Education, University and Research (PRIN project 2007 and 2009). In this initiative, our aim was to understand the relationships between cultural and geographical factors and the genetic structure of the Italian isolates (Destro Bisol et al., 2008). The research work continued under the umbrella of the project “Human biodiversity in Italy: micro-evolutionary patterns” (PRIN project 2009-2011). In this second step, we also focused on the role of social structures in shaping the human biodiversity of isolated groups1. Further support to the whole initiative was provided by the Istituto Italiano di Antropologia (project “Bio-cultural Atlas of Italian populations”).

As genetic tools, we used mtDNA and Y chromosome polymorphisms, a choice based on their power to detect microevolutionary events on a wider timescale than bi-parentally transmitted polymorphisms, their relatively low costs, abundance of comparative data both at micro- and macro-geographic scale and the possibility to estimate gender-biased processes. Notably, the outputs of the study are not limited to scientific papers (e.g. Destro Bisol et al., 2008; Robledo et al., 2009; Boattini et al., 2010; Bertoncini et al., 2012; Capocasa et al., 2013b; see Supplementary File S1 for a complete list of papers and communications to congress), but also include the return of scientific results to the investigated communities (e.g. Capocasa et al., 2012, 2013a) and a dedicated genetic online database (http://www.isita-org.com/Anthro-Digit/italianisolates_PRINproject/index.html).

In this paper, we present an account of the results obtained by the research units (RUs) involved in the collaborative study. To this aim, four case studies spanning from the eastern Alps to southern Italy are described and discussed, with a focus on their micro-evolutionary and anthropological implications. Thereafter, we present the results of a systematic analysis of the relations between linguistic, geographic and genetic isolation combining the data obtained in the course of the long-term study with literature and unpublished results. More in particular, we: (i) analyze the genetic structure of numerous Italian populations in order to evaluate the association of linguistic and geographic factors with genetic isolation; (ii) compare the extent of genetic diversity observed in Italy and Europe taking into account both open populations and geographically and/or linguistically isolated groups.

Materials and methods

The dataset

Our original dataset includes data from 873 individuals for mtDNA and 795 individuals for Y chromosome from 19 populations (see Table 1) studied in the course of the PRIN projects. It was integrated with other published and unpublished results, reaching a total of 2875 and 1811 entries for mtDNA and Y chromosome polymorphisms, respectively (see Appendix; Congiu et al., 2012).

Buccal swabs were collected in apparently healthy donors (see Tab. 1). A detailed description of laboratory analyses for DNA extraction, amplification, purification and genotyping is reported in Supplementary File S2.

Statistical analyses

Haplotype diversity (HD) and its standard errors were calculated according to Nei 1987. Pairwise differences among all the populations of the datasets were calculated using the genetic distance measure Fst (Reynolds et al., 1983; Slatkin, 1995). Analyses of molecular variance (AMOVA) were performed in order to test genetic differences among Sardinian populations. Demographic descriptive indices (Tajima’s D, Fu’s Fs and Harpending’s raggedness) were calculated to test for the presence of signs of demographic expansion (Tajima, 1989; Harpending, 1994; Fu, 1997). An analysis of haplotype sharing was performed in order to evaluate the specific contribution of private haplotypes to the observed patterns of genetic differentiation among populations.

---

1 As a rule, by the term “isolated populations” we mean linguistic and/or geographic isolates.


Tab. 1. Populations analyzed in the course of the collaborative study. Abbreviations: ALT, altitude; FS, founder surnames; G, geographic isolate; GL, geographic/linguistic isolate; GP, grandparents; L, linguistic isolate; O, open population.

<table>
<thead>
<tr>
<th>POPULATION</th>
<th>ABB.</th>
<th>STATUS</th>
<th>SAMPLING STRATEGY</th>
<th>CENSUS SIZE*</th>
<th>ALT.°</th>
<th>N</th>
<th>REFERENCE</th>
<th>POPuLATION ABB.</th>
<th>STATuS</th>
<th>SAMPLING STRATEGY</th>
<th>CENSUS SIZE*</th>
<th>ALT.°</th>
<th>N</th>
<th>REFERENCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arbereshe (Calabria)</td>
<td>ARC</td>
<td>GL</td>
<td>FS</td>
<td>28034</td>
<td>535</td>
<td>87</td>
<td>This study</td>
<td>87 Boattini et al., 2011; Pettener, pers. com.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arbereshe (Sicily)</td>
<td>ARS</td>
<td>L</td>
<td>FS</td>
<td>7875</td>
<td>645</td>
<td>40</td>
<td>This study</td>
<td>42 Pettener, pers. com.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Benetutti</td>
<td>BEN</td>
<td>G</td>
<td>FS</td>
<td>2010</td>
<td>406</td>
<td>50</td>
<td>This study</td>
<td>45 Calò et al., 2013</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cadore</td>
<td>CAD</td>
<td>O</td>
<td>GP</td>
<td>10797</td>
<td>984</td>
<td>32</td>
<td>Caglià, pers. com.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Carloforte</td>
<td>CFT</td>
<td>GL</td>
<td>FS/GP</td>
<td>6420</td>
<td>10</td>
<td>51</td>
<td>Calò et al., 2012</td>
<td>41 Robledo et al., 2012</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cimbrians (Lessinia)</td>
<td>LES</td>
<td>GL</td>
<td>GP</td>
<td>13455</td>
<td>758</td>
<td>40</td>
<td>Capocasa et al., 2013</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cimbrians (Luserna)</td>
<td>LUS</td>
<td>GL</td>
<td>GP</td>
<td>286</td>
<td>1333</td>
<td>21</td>
<td>Coia et al., 2012</td>
<td>25 Coia et al., 2013</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cricello</td>
<td>CIR</td>
<td>G</td>
<td>FS</td>
<td>2501</td>
<td>700</td>
<td>27</td>
<td>This study</td>
<td>34 Tofanelli, pers. com.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cosenza</td>
<td>COS</td>
<td>O</td>
<td>GP</td>
<td>69131</td>
<td>238</td>
<td>42</td>
<td>Pettener, pers. com.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fiemme Valley</td>
<td>FIE</td>
<td>G</td>
<td>GP</td>
<td>18990</td>
<td>1033</td>
<td>41</td>
<td>Coia et al., 2012</td>
<td>41 Coia et al., 2013</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ladins (Fassa Valley)</td>
<td>LVF</td>
<td>GL</td>
<td>GP</td>
<td>9894</td>
<td>1345</td>
<td>47</td>
<td>Coia et al., 2012</td>
<td>47 Coia et al., 2013</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lucca plain</td>
<td>PIL</td>
<td>O</td>
<td>GP</td>
<td>154928</td>
<td>81</td>
<td>50</td>
<td>This study</td>
<td>50 This study</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>North Sardinia</td>
<td>NSA</td>
<td>O</td>
<td>GP</td>
<td>67253</td>
<td>440</td>
<td>40</td>
<td>This study</td>
<td>47 Calò et al., 2013</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sappada</td>
<td>SAP</td>
<td>GL</td>
<td>GP</td>
<td>1307</td>
<td>1217</td>
<td>59</td>
<td>Capocasa et al., 2013b</td>
<td>36 Coia et al., 2013</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sauris</td>
<td>SAU</td>
<td>GL</td>
<td>GP</td>
<td>429</td>
<td>1212</td>
<td>48</td>
<td>Capocasa et al., 2013b</td>
<td>29 Coia et al., 2013</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sulcis-Iglesiente</td>
<td>SGL</td>
<td>O</td>
<td>GP</td>
<td>128614</td>
<td>96</td>
<td>50</td>
<td>Robledo et al., 2012</td>
<td>46 Robledo et al., 2012</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Timau</td>
<td>TIM</td>
<td>GL</td>
<td>GP</td>
<td>500</td>
<td>830</td>
<td>46</td>
<td>Capocasa et al., 2013b</td>
<td>22 Coia et al., 2013</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vagli</td>
<td>VAG</td>
<td>G</td>
<td>FS</td>
<td>995</td>
<td>575</td>
<td>22</td>
<td>This study</td>
<td>23 Tofanelli pers. com.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

° meters above sea level.
All parameters of intra- and inter-population genetic diversity were calculated using the Arlequin 3.5 software (Excoffier & Lischer, 2010). Multidimensional scaling (MDS) was applied to genetic distance matrices to visualize genetic differentiation among populations using the SPSS software (release 16.0.1 for Windows, S.P.S.S. Inc.). In order to evaluate the effect of census size on the intra- and inter-population diversity, we performed a stepwise multiple regression analysis using the current census size and altitude as covariates. Census size was log-transformed in order to linearize its relation to HD and average Fst. Unless otherwise stated, analyses were performed using mtDNA HVR-1 region (from 16033 to 16365 n.p.) and 15 Y chromosome STRs (DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS437, DYS438, DYS439, DYS448, DYS456, DYS458, DYS635 and GATA H4.1).

Genetic relationships among Arbereshe, putative source populations (Albania, Greece) and sink populations (Calabria, Sicily) were analysed as follows: populations were first grouped in clusters using a non-hierarchical EM algorithm (mclust R package; see Fraley & Raftery, 2002, 2006), then posterior membership probabilities (for each population and for each cluster) were calculated using Discriminant Analysis of Principal Components (DAPC, adegenet R package; see Jombart et al., 2010) and graphically represented with a bar-plot.

To assess which interval of stepwise mutational differences between the male specific region of Y chromosome (MSY) haplotypes is the most suitable to represent the Ligures contribution to Samnium, we used the Equation 31 described by Walsh (2001) implemented in the software ASHEs 1.1 (ashes.codeplex.com, Tofanelli et al., 2011). The Bayesian posterior distribution of expected pairwise mutation differences was calculated for haplotypes separated by 78 generations (2,190 years ago using 28 years per generation), assuming a strict stepwise mutational model, a mutation rate of 3.09x10^{-3} per locus per generation (averaged values from observed single-locus germ-line mutations following Ballantyne et al., 2010) and a lambda value of 0.0002.

**Genes, geography and language: a tale of four stories**

The collaborative project covered a substantial part of the linguistic and/or geographic isolates settled on Italian territory. What follows is a selection of the results obtained by each of the four RUs, which illustrates the variety of factors to be taken into account when studying human isolated populations and the diversity of approaches needed to carry out these investigations.

**The Arbereshe: between Italy and the Balkans**

The Arbereshe are one of the largest linguistic minorities settled in Italy (~100,000 individuals; see Fig. 1). They originate from migratory waves from Albania between the end of the 15th and beginning of the 16th century in response to the invasion of the Balkans by the Ottoman Empire. Historians agree that most of the immigrants came from the south of Albania (Toskeria), often passing along the Peloponnese peninsula (Zangari, 1941). This hypothesis is supported by evidence that Arberisht, the language spoken by the Arbereshe, is part of the Tosk dialect group originally spoken in Toskeria. At present, there are 50 Arbereshe communities in Italy. They are located in Southern Italy, where they are separated from each other and interspersed among Italian villages. As such, they constitute an interesting case study allowing us to explore the effects of cultural isolation on genetic variability, as well as their relationships with the source (Southern Balkans) and sink populations (Southern Italy).

The Arbereshe of Calabria are probably the most important and interesting group, due to the high number of villages (30, with a total of ~ 60,000 individuals) and an exceptional preservation of their cultural identity. The latter feature is evident not only in the language (Arberisht), but also by the Greek Orthodox religion and a common mytho-history (Fiorini et al., 2007). The investigated populations are scattered around the Pollino massif, an area that was historically characterised by strong geographic isolation and an economy based on sheep breeding and crop cultivation. During
Fig. 1 - Geographic location of the populations analysed in this study. Filled and empty symbols indicate populations studied in the PRIN projects and literature data, respectively. Population abbreviations as in Table 1 and Appendix. The colour version of this figure is available at the JASs website.
the second half of the 20th century, the whole area was heavily struck by emigration which led to a partial loss of original cultural traits (breakdown of isolates) (Fiorini et al., 2007).

In Sicily, the Arbereshe are currently found in only three municipalities of the Palermo province (Contessa Entellina, Piana degli Albanesi and Santa Cristina Gela). Characterized by a relatively low population size (~15,000 individuals), they are settled in less geographically isolated areas than Calabrians. Their migration history is particularly complex, involving various intermediate steps (both in the Balkans and in Italy) before their definitive settlement. Furthermore, there is documented evidence of partial re-peopling events from Greece, the most well-known case being that of approximately 100 families coming from the island of Andros in 1520 AD (Giunta, 2003).

Calabrian Arbereshe were sampled in 13 different villages (Acquaformosa, Cerzeto, Civita, Firmo, Frascineto, Lungro, Plataci, S. Basile, S. Cosmo Albanese, S. Demetrio Corone, S. Giorgio Albanese, S. Sofia d’Epiro and Spezzano Albanese). Sicilian Arbereshe were sampled in two villages (Contessa Entellina, Piana degli Albanesi). In both cases, founder surnames were used as inclusion criterion (Boattini et al., 2011) and “control” samples were collected from open neighbouring populations. All the sampled individuals have been unrelated for at least three generations. Data for Albanian and Greek populations were obtained from literature (Bosch et al., 2006).

We first detected the genetic signatures of ethno-linguistic affiliation of the Arbereshe from Calabria by means of biodemographic analyses (Fiorini et al., 2007; Tagarelli et al., 2007). We showed that Arbereshe populations are clearly distinguishable from their Italian neighbours, and observed a correlation between geographic- and surname-based distances between the Arbereshe groups. In addition, Arbereshe groups from the Pollino area showed considerably higher marital isonymy levels than in Italians (0.080 and 0.050, respectively). All these lines of evidence suggest that Arbereshe of Calabria are a group of populations that have undergone cultural and geographical isolation. Importantly, diachronic analyses revealed that such isolation features declined progressively during the last two centuries, and more rapidly in the last 50 years. A Y-chromosome study based on a first sampling of Calabrian Arbereshe (a total of 40 individuals) confirmed and extended these findings (Boattini et al., 2011). Arbereshe were found to provide a signal of discontinuity in the Italian genetic landscape, while showing a strong affinity with some modern Balkan populations (in particular Albanians and Kosovars).

Further insights are provided by taking into account unpublished results for Y-chromosomal and mitochondrial polymorphisms from a larger sampling of Calabrian Arbereshe (104 and 101 individuals, respectively) together with new data on Sicilian Arbereshe (44 and 57 individuals). Haplotype diversity values observed for male and female lineages in Calabrian Arbereshe (0.976 and 0.995, respectively) and Sicilian Arbereshe (0.979 and 0.992) are very close to those of Italian neighbour populations of Cosenza (0.994 and 1.000) and Trapani-Enna (0.985 and 0.999). These results are in line with our previous conclusions that Arbereshe have “conserved much of [their] ancestral genetic diversity along with [their] founder surnames and cultural features” (Fiorini et al., 2007; Boattini et al., 2011) and that their Y-chromosomal gene pool may mirror “the genetic structure of the migrants that came to Italy from the Southern Balkans (Albania) five centuries ago” (Boattini et al., 2011). As shown in Fig. 2, Calabrian Arbereshe cluster (with high membership probabilities) with Albanians both for male and female lineages, as predicted by the isolation pattern inferred from surname analysis and their similarity with the source population. Intriguingly, our results suggest a different and more complex scenario for the Arbereshe from Sicily. Their Y-chromosome clustering pattern supports a significantadmixture with Greeks, which agrees with historical information. On the other hand, mtDNA results highlight a substantial similarity to populations from Southern Italy, a likely consequence of admixture processes (and/or sexually imbalanced migration patterns).
These findings suggest a noticeable difference in the ethnogenesis of Calabrian and Sicilian Arbereshe, with a more evident continuity with source populations for the former and lower impact of isolation for the latter. The study of Albanian-speaking minorities of Southern Italy illustrates, therefore, how the genetic structure of populations sharing the same ethno-linguistic label might have been shaped by diverse micro-evolutionary histories.

An analysis of isolated populations in Sardinia

Populations from Sardinia provide a paradigmatic example of genetic differentiation in an insular context. The combined effect of small population size, endogamy, and the associated consanguinity, resulting from geographical and cultural barriers, has increased among-group diversity, mainly via long-term genetic drift. A further source of genetic heterogeneity was provided by the different incidence of malaria morbidity. Thus, the unique genetic structure and high level of isolation of some Sardinian villages create opportunities to map genes involved in multifactorial diseases (Zavattari et al., 2000, Angius et al., 2001). However, the interplay between microevolutionary, demographic and cultural factors makes Sardinian populations of great interest also for anthropological studies.

During the collaborative project, the RU of the University of Cagliari focused on the isolates of Carloforte (linguistic) and Benetutti (geographic), in order to assess the degree of isolation and the relative contribution of cultural and geographical barriers (see Fig. 1). Carloforte (6420 inhabitants), located in the small island of San Pietro, off the Southwestern coast of Sardinia, is an alloglot community founded in 1738 AD by Ligurian migrants coming from Tabarka, a Tunisian island (Vallebona, 1974). As a historical imprint, the present-day Carloforte inhabitants speak the ancient dialect of Pegli (Liguria), also referred to as “tabarkino” (a minority language officially recognized by Sardinian regional legislation). Benetutti, a small village of 2010 inhabitants, is located in an area of Northern Sardinia, Goceano, which is characterized by strong geographic isolation. The first historical-demographic documents regarding Benetutti may be found in the “Quinque Libri”, dating back to 1618 AD.
By studying Y chromosome and mtDNA polymorphisms, we were able to detect signatures of genetic isolation and of a constant demographic state in both populations (Robledo et al., 2012; Calò et al., 2013; present study). In fact, haplotype diversity observed in Carloforte (0.975 and 0.921 for Y chromosome and mtDNA, respectively) and Benetutti (0.975 and 0.918 for Y chromosome and mtDNA, respectively) are lower than the ones observed in open Sardinian populations (Robledo et al., 2012). In fact, using both founder surnames and grandparent rule as a selection criteria, we were able to select individuals descending from village founders and capture a large extent of within-population genetic diversity at the same time. A similar conclusion was reached in a different study on Y chromosome lineages in isolates located in the Italian Western Alps (Boattini et al., 2010). In any case, the lack of a substantial drop in within-population diversity is consistent with historical records which do not point to any dramatic change in population size. This is also supported by demographic inferences based on mtDNA analysis. In fact, the negative but statistically insignificant values of Tajima’s D and Fu’s index together with values of Harpending’s raggedness suggest that these populations have kept their sizes small and constant over time.

The analysis of inter-population variation proved useful to shed light on other aspects of the genetic history of the two populations under study. Effects of genetic isolation were detected with both uniparental markers, since genetic distances of Benetutti and Carloforte from neighbouring open populations were found to be statistically significant (p<0.05). In the multi-dimensional scaling (MDS) plot based on Y chromosome data (Fig. 3a), Carloforte lies in a rather isolated position, although not very distant from other Italian populations. Genetic differentiation between the sampled Ligurian population and the founders of Carloforte may account for the former evidence. Benetutti is rather separated from other groups, including Sardinians. The mtDNA based plot (Fig. 3b) confirms both the peculiar behaviour of the Carloforte, which is the only Sardinian population located in the x-positive values of the plot, and the genetic differentiation of Benetutti from other Sardinians. However, differently from what we observed using paternal lineages, we found that Carloforte lies close to African populations, particularly Tunisia. Although no support comes from historical sources, a hypothesis of admixture of Carloforte males with Tunisian females is worth considering due to the occurrence at appreciable frequency (8%) of the mtDNA haplogroup M, characteristic of Northern African and Western Asian populations but so far undetected in Sardinia and Italy (Gonzalez et al., 2007; Fraumene et al., 2003, 2006; Morelli et al., 2000).

An analysis of molecular variance among Sardinian populations was carried out to evaluate the apportionment of genetic diversity. Consistent with previous research work in Sardinia, only a small portion of variation could be attributed to differences between populations (3.78% for mtDNA and 3.31% for Y chromosome; p<0.05). Interestingly, data on preferential male mobility in the Carloforte (Calò & Vona, 1994) and, more in general, the historically documented practice of matrilocal behaviour of some areas of Sardinia (Murru Corriga, 1995) might explain the slightly larger variance observed for maternal lineages.

**Genetic and cultural isolation in Northern Apennines**

Recent studies suggest an ethno-linguistic continuity between the human groups currently inhabiting the inner valleys of North-West Tuscany (Lunigiana, Garfagnana and orographically linked areas) and populations settled in this area in pre-Roman times. The genetic evidence of this connection was formerly reported by Piazza et al. (1988), who associated one of the main principal components obtained from blood protein alleles to the inheritance of those ancient tribes named Ligures according to the Latin etymology. This possible link is supported by recent studies of variation in the male-specific portion of the Y chromosome, which have identified North Western Tuscany as an area where lineages
Fig. 3 - Multi-dimensional scaling plots of Fst genetic distances among Sardinian and Mediterranean populations based on (A) 7 Y chromosome STRs (DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392, and DYS393; stress value=0.046) and (B) mtDNA HVR-1 sequences (stress value=0.130). Abbreviations and references: (a) Benetutti, BEN; North Sardinia, NSA (Calò et al., 2013). Carloforte, CFT; Sulcis-Iglesiente, SGL (Robledo et al., 2012). Emilia Romagna, ERM (Ferri et al., 2009). Lazio, LAZ; Liguria, LIG; Puglia, PUG; Toscana, TUS; Umbria, UMB (Presciuttini et al., 2001). Lombardia, LOM (Cerri et al., 2005). Marche, MAR (Onofri et al., 2007). Sicilia, SIC (Ghiani et al., 2004); Tunisia, TUN (Frigi et al., 2006). Veneto, VEN (Turrina et al., 2006). (b) Algeria, ALG; Catalona, CAT (Cortereal et al., 1996). Andalusia, ADL; Balearic Islands, BAL; Gallura, GAL; Morocco, MOR; Nuorese, NUO; South Corsica, SCO; Trexenta, TRX (Falchi et al., 2006). Benetutti, BEN, North Sardinia, NSA (this study). Carloforte, CFT (Calò et al., 2012). Central Corsica, CCO (Varesi et al., 2000). Egypt, EGY (Kings et al., 1999). Middle East, MID (Di Rienzo & Wilson, 1991). North Italy, ITA (Stenico et al., 1996). Ogliastro, SAO (Fraumene et al., 2006). Sicilia, SIC (Vona et al., 2001). Sulcis Iglesiente, SGL (Robledo et al., 2012). Toscana, TUS (Francalacci et al., 1996). Tunisia, TUN (Plaza et al., 2003). Tunisian Berbers, BRB (Fadhlaoui-Zid et al., 2004). Turchia, TUR (Comas et al., 1996).
with a putative post-Neolithic date of origin persist at relatively high frequencies (Busby et al., 2012, Boattini et al., 2013).

Modern historical sources consider Ligures Apuani to be not entirely eradicated by the Roman military campaigns of the II century BC and the following mass deportations to Ager Taurasinum (Samnium) of at least 47,000 people (Titus Livius, In: Storia Romana XXXVIII). Rather, they depict the fate of natives in the form of confederated tribes surviving in hilly or mountainous areas, until they were eventually absorbed in the demo-territorial institutes later imposed by the Roman authorities (i.e. res publicae, arcifinalis, excepta), often maintaining a certain degree of autonomy and ownership (Dilke, 1971; Gagliardi, 2006; Marcuccetti, 2012).

From a linguistic point of view, many hamlets remained in the Apuan Alps and the Northern Apennines show similar forms of resistance to the “tuscanization” of the spoken language, maintaining a sort of bilingualism exerted through the use of lexemes and phonemes which are traceable to a pre-Latin matrix (Ambrosi, 1956; Giacomelli, 1975; ALT-Web). Moreover, the distribution of many toponyms with pre-Latin roots, in particular those referring to general orographic names such as rivers, mountain reliefs, open places (i.e. nava, asc, var, dur, mat, penn, aus, kar), is limited over Europe to conservative areas and have a peak of density in the territory once inhabited by Ligures tribes (Marcuccetti, 2008).

The hypothesis of a genetic legacy between Ligures Apuani and present Apuan and Samnite isolates (Ligures Legacy Model, LLM) was tested by the RU of the University of Pisa through the genetic characterization of uni-linear markers in a sample of unrelated donors from the communities of Vagli (Province of Lucca) and Circello (Province of Benevento) selected according to founder surnames. Vagli is located in the core of the area which has many archeological records linked to Ligures Apuani. Its elder inhabitants still speak a language characterized by a number of linguistic relicts (Ambrosi, 1956; Guazzelli, 2001). Circello lays in close proximity to the remains of Macchia, the town in the Samnium around which the deported Ligures (L. Baebiani) were forced to settle in 180 BC (Patterson, 2009).

A slight reduction of HD at both mtDNA and Y Chromosome has been observed for Vagli (0.948 and 0.984) and Circello (0.960 and 0.975) relative to the neighbouring populations of Piana di Lucca (0.983 and 0.999) and Benevento (0.989 and 1.000). As expected for communities of Indo-European ancestry, usually practicing prevalence of female vs male mobility (patrilocal), genetic distances based on mtDNA are weaker discriminators than distances based on MSY haplotypes (see Supplementary Tab. S1). Preliminary comparative assessments of MSY profiles suggest that the diversity of Apuans might be due to an excess of R-U152 haplotypes, whose diffusion in Italy is thought to coincide with the diffusion of Ligures cultural features in the Middle-Late Bronze Age (Bertoncini et al., 2012).

As a whole, the two communities under study (Vagli and Circello) showed a genetic pattern which is compatible with a long history of isolation but also with quite diverging microevolutionary histories after contacts implied by the LLM. As a more direct test of a genetic continuum with Ligures tribes (Fig. 4), we assessed whether an enrichment of matches compatible with the Titus Livius deportation hypothesis is detectable when comparing any MSY haplotype of the local population (Vagli) with haplotypes of both, the putative displaced (Circello) and the open Samnite population (Benevento). The enrichment of LLM-compatible matches in the Vagli-Circello curve totaled about 80% and was extremely statistically significant (Fisher exact test, p<0.0001).

The hypothesis of direct descent of the resident males in the Apuan and in the Circello area from members of Ligures tribes, who escaped deportation, cannot be ruled out. Further data from wider samples and haplogroup diagnostic markers, as well as more extensive simulation analyses will help achieve more robust inferences. Nonetheless, our case study shows that even mild geographical and cultural isolation may lead to the preservation of a long genetic thread connecting present populations to ancient layers of
pre-Roman Italy. As a corollary, it suggests that many isolated Italian communities other than Northern Apennine ones may escape simplistic classification schemes (i.e. linguistic vs geographic isolate), owing to the gradual fringing and recent oblivion of a common ancient cultural identity. Finally, our study highlights the usefulness of accurate non random genetic samplings to uncover genetic layers obscured by recent reshalings within and among human populations.

Genetic and cultural isolation in the Eastern Italian Alps

Despite the presence of several linguistic islands, the Eastern Alps had not been thoroughly investigated from a bio-anthropological perspective before the start of the collaborative project. The RU from the University of Rome “La Sapienza” focused on six populations: Ladins from the Fassa Valley, Cimbrians from Lessinia and Luserna and the ethno-linguistic “islands” of Sappada, Sauris and Timau (see Fig. 1). Dolomitic Ladins probably originated from a founder group of pre-Indoeuropean speakers that underwent a process of “Latinization” under the pressure of the Roman Empire. At the beginning of the Middle Ages, they fragmented as a result of the arrival of German people. The Ladin community of the Fassa Valley consists of approximately 6,000 individuals settled in seven villages (Toso, 2008). The other populations under study are German-speaking groups thought to be in continuity with peoples migrated from Carinthia, Bavaria and Tyrol in the high and late Middle Ages (Denison, 1971; Rizzi, 1993). After their initial settlement, these ethno-linguistic isolates maintained close social and cultural links (De Concini, 1997; Navarra, 2002). At present, their common background is evident due to their sharing of cultural aspects, such as language and traditions, which persist despite a certain degree of cultural exchanges with the surrounding neo-latin groups. The two Cimbrian settlements have different social and environmental characteristics. Luserna is a small town located over 1,300 meters a.s.l. in the Trentino region, and inhabited by approximately 300 people. Conversely, Lessinia is populated by over 13,000 inhabitants who are distributed in a wider and less “hostile” mountainous territory in the province of Verona in the Veneto region. Sappada is located at an altitude of 1,245 m.a.s.l. in the Veneto region (province of Belluno; 1,307 inhabitants). Sauris and Timau are two small
villages (429 and 500 inhabitants, respectively) located in the Carnic Alps. The former is placed at 1,212 m.a.s.l. within the province of Udine, in the Friuli Venezia Giulia region. Timau is situated at 830 m.a.s.l. in the But valley.

The analysis of genetic variation suggests a certain degree of genetic isolation from surrounding populations for Luserna, Sappada, Sauris and Timau. In fact, in accordance with the James Neel’s statement that genetic isolates are “derived from a relatively small population sample, which then slowly expand, with very little recruitment from outside the group” (Neel, 1992), we observed low haplotype diversity values and lack of any signal of demographic expansion (Fu’s Fs and Harpending’s raggedness) (Coia et al., 2012, 2013; Capocasa et al., 2013b). Analyses of genetic distances highlight the differentiation among the four communities and their diversity from other European populations. Furthermore, genetic differentiation among Alpine populations was detected even at individual level using a Bayesian method to cluster multilocus genotypes based on 15 autosomal microsatellites (Montinaro et al., 2012). A different pattern was detected for Ladins from Fassa and Cimbrians from Lessinia, who showed neither an evident reduction of mtDNA and Y chromosome intra-population diversity nor a significant departure from the European genetic background.

Although patterns of genetic variation and the lack of signature of demographic expansion observed in Sappada, Sauris and Timau are compatible with what is to be expected in human isolates, it is possible that these results may simply reflect their small initial effective size. As a more direct test of genetic isolation, we performed an analysis of gene flow using a Bayesian approach, applying an Isolation with Migration model (Hey & Nielsen, 2007). Through this approach, we inferred a reduced incoming gene flow for Sappada, Sauris and Timau, both from a wide Central Western European and a neighbouring open Italian-speaking population (Fig. 5). Therefore, our results support the hypothesis that the peculiar linguistic and geographical factors acting on the communities under study might have determined a substantial degree of genetic isolation, so shaping their genetic structure (Capocasa et al., 2013b).

Another aspect which we noticed was the extreme genetic differentiation among Sappada, Sauris and Timau, an unexpected finding considering their relative geographic proximity and closeness in terms of language and traditions (De Concini, 1997; Navarra, 2002). In order to evaluate the magnitude of the genetic differentiation among the three linguistic islands, we first compared them with other European language minorities (Cimbrians and Ladins from Eastern Alps, Aromuns from Albania and Macedonia) whose members are settled in geographically close but distinct locations. Interestingly, Sappada, Sauris and Timau showed inter-population distance values which were two (Y chromosome) or three (mtDNA) times higher than in other groups (Capocasa et al., 2013b). We then wondered whether cultural factors might help explain this pattern. This seems to be worth testing since members of the Eastern Alps linguistic islands do not self-identify as belonging to the same community despite their cultural homogeneity (Steinicke et al., 2011). Such behaviour seems to be in sharp contrast with the sense of belonging of Cimbrians, Ladins and Aromuns and their respective ethnic groups. Therefore, we went on by hypothesizing that this “local ethnicity” may have played a role in marriage strategies, decreasing the genetic exchange among the three linguistic islands. Carrying out coalescent simulations, we inferred that the combined effect of a small initial effective size, as indicated by historical sources (Petris, 1980; Brunettin, 1998; Peratoner, 2002), and a substantial reduction of gene flow, as implied by a local ethnicity model, can explain the pattern observed for Sappada, Sauris and Timau communities (Capocasa et al., 2013b). More in general, our case study shows that complementing classical measures of genetic diversity with Bayesian estimates of gene flow and simulations of micro-evolutionary models may help us better understand patterns of genetic isolation and its relations with demographic and cultural factors in human populations.
Studies of cultural and geographical factors of genetic isolation have been so far mainly concerned with single or a limited number of groups, focusing on specific populational contexts (e.g. Jeran et al., 2009; van Oven et al., 2011; Veeramah et al., 2011). Combining data relative to numerous isolated groups, produced in the course of this collaborative project, with literature and unpublished results gives us the opportunity to draw inferences on the relations between culture, geography and genetics on a larger geographic scale. To this aim, we built large mtDNA and Y Chromosome databases (see Tab. 1 and Appendix), including 57 populations (29 isolates and 28 open groups) and 46 populations (22 isolates and 24 open), respectively. In order to reach the maximum coverage of the Italian territory/populations, we decided to include geographic and/or linguistic isolates studied with different sampling schemes.

As a preliminary step, we compared the genetic structure of isolated and open populations, without considering any further subdivision. We decided to focus on HD and Fst values, since these parameters are little influenced by differences in sample size among those commonly used to detect genetic isolation (Anagnostou et al., in preparation). The HD of isolated populations was found to be significantly lower (Mann-Whitney test p<0.001) than those of open groups, both for mtDNA (from 0.778 to 0.990 vs 0.903-0.995) and Y Chromosome (from 0.590 to 0.998 vs 0.993-1.000) (see Supplementary Tab. S2 and S3; Supplementary Fig. S1). Most of the highest average genetic distances from open populations were observed in geographic and geographic/linguistic isolates, which is another expected effect of genetic isolation (Fig. 6; Supplementary Fig. S2). The difference between open and isolated population groups was slightly more evident for Y chromosome (0.011-0.322 for isolates vs 0.004-0.046 for open populations) than mtDNA (0.000-0.085 vs 0.000-0.013), but highly statistically significant for both polymorphisms (Mann-Whitney test p<0.001). Private haplotypes do not contribute in any significant way to...
these patterns (Supplementary Fig. S3), a finding which is consistent with the relatively recent age of most isolated populations under study.

On the whole, these results suggest that populations subject to geographic and linguistic isolation are more likely to show lower HD and higher average Fst than open groups. However, they do not clarify whether these differences may be actually regarded as an effect of genetic isolation and the relative weight of geographic and linguistic barriers. To explore these issues, we carried out further analyses in three steps.

First, we used the departure from the pattern of open populations as a criterion to distinguish between simple fluctuations of values and signatures of genetic isolation. To do this, we used the inter-quartile method to identify weak and strong outliers for HD and average Fst distributions. To be considered as bearing a signature of genetic isolation, a given population should display a departure falling in the “weak outlier” interval for both parameters. Furthermore, they were labelled as strong only when both reduction in HD and increase in average Fst exceeded the threshold delimiting the strong outlier area. Using these criteria, we identified 11 populations (out of 29 geographic and/or linguistic isolates) bearing signatures of genetic isolation (10 strong and 1 weak) for mtDNA. Results for Y chromosome do not markedly differ for the number of groups showing signatures of genetic isolation (8 out of 22), but, rather, for the proportion between strong (3) and weak (5) signals (see Fig. 7).

Second, we took into account the possible confounding effect of census sizes in order to validate the obtained inferences. This parameter was preferred to population density, for which no statistically significant correlation with either HD or average Fst was observed (data not shown). To this purpose, we carried out a stepwise multiple regression analysis on all our populations subject to isolation factors using census as a dependent variable. In these calculations, altitude was used as a control variable. This choice was based on two reasons: (i) it may be regarded as one of the ways in which geography may determine isolation between populations, rather than a confounding factor; (ii) we observed a significant correlation between altitude and census size in our dataset (Pearson’s $r=-0.495$, $p=9.201\times10^{-5}$; Spearman’s $\rho=-0.546$, $p=1.566\times10^{-5}$), as already reported in previous studies (Cavalli Sforza & Bodmer, 1971; for the Italian context see Franceschi & Paoli, 1994; Morelli et al., 2002; Capocasa et al., 2014).

We found a significant correlation between census size and average Fst for both polymorphisms. However, they account only for a minor portion of the total variance (10.5% for mtDNA and 31.8% for Y chromosome). Census size and HD were found to be significantly correlated only for Y chromosome, but again with a weak effect (26.5% of variance explained) (see Supplementary Tab. S4). Therefore, the effect of census size, although detectable, does not seem so marked to undermine our inferences. Altitude was found to have a low and significant, but yet negligible effect only on Y chromosome Fst distribution (10.5% of total variance), implying that other factors are responsible for the observed signals of genetic isolation.

Third and finally, we tried to understand whether the conditions of geographic isolation alone, or combined with linguistic isolation, may have different outcomes on the genetic structure of populations under study. Unfortunately, the presence of only 3 linguistic isolates do not make it possible to perform any test of the effect of linguistic barriers alone. In the mtDNA plot (Fig. 7), signatures of genetic isolation were observed more frequently in geographic/linguistic (6 out of 10; 60%) than in geographic isolates (5 out of 16: 31%). A comparable disproportion was shown by Y chromosome polymorphisms, with 2 geographic (out of 9; 22%) and 5 geographic/linguistic isolates (out of 10; 50%) showing the signal. Limiting the comparison to strong signatures (defined as above), the ratio of positives between geographic/linguistic and geographic isolates does not change substantially either for mtDNA (5 out of 10 vs 5 out of 16, respectively) or Y chromosome (3 out of 10 vs 0 out of 9). A further indication of the effect of linguistic isolation on the genetic structure of populations is provided by the demographic inferences based on mtDNA data.
Fig. 6 - Multi-Dimensional Scaling plot of Fst genetic distances among Italian populations based on (A) mtDNA HVR-1 sequences (stress value=0.236) and (B) 15 Y Chromosome STRs (stress value=0.156). Population abbreviations as in Table 1. The colour version of this figure is available at the JASs website.
In fact, four of the 6 groups which showed a lack of signals of demographic expansion (non significant Fu's and high raggedness values) belong to the geographic /linguistic isolates (Cimbrians from Luserna, Sappada, Sauris, Timau), whereas only 2 are geographic isolates (Vallepietra and Val Pusteria) (Supplementary Tab. S2).

What do all these findings tell us? We admit that no simple cause-effect relationship between linguistic barriers and genetic structure can be inferred when studying populations with different ethnogenesis and demographic histories. Nonetheless, it seems clearly evident that the combination of linguistic and geographic isolation factors is frequently associated with the presence of robust signatures of genetic isolation and a condition of slow, if any, demographic expansion, fitting the classical definition of genetic isolate given by James Neel in 1992. This is not trivial since, to date, evidence of a joint effect of geographic and linguistic isolation factors in Italian populations has been provided by biodemographic investigations only (e.g. Fiorini et al., 2007; Boattini et al., 2011). It is also worth noting that the extreme reduction of genetic heterogeneity observed in some geographic isolates in which language has acted as an additional barrier to gene flow makes such groups of particular usefulness for genetic studies of complex diseases (see Heutink & Oostra, 2002).

Re-assessing genetic diversity of Italian populations

Another opportunity offered by our collaborative study was to produce a more complete picture of the diversity of Italian populations, combining data from large populations, which are presumably exempt from isolation factors, with the results obtained for ethno-linguistic minorities and geographical isolates. A contextualization into the European background seemed to us to represent a convenient approach in order to evaluate the magnitude of genetic heterogeneity, a necessary step to understand whether human genetic variation parallels plant and animal biodiversity (see introduction). Also in this case, we started by assembling ad hoc mtDNA and Y chromosome datasets, in which we included populations settled throughout the continent (Supplementary Tab. S5 and S6).

Both Italian and European open populations are found in the middle of the MDS plots of genetic distances, with a more tight clustering observed for mtDNA (Fig. 8). Although strong signals of genetic differentiation for both maternal and paternal lineages were detected for some European isolates (e.g. Aromuns from Stip and Basques), the most evident outliers are represented by some Italian isolates (Sappada and Luserna). Looking at mtDNA only, the genetic distinctiveness of the Isarco Valley, Ogliastra and Vallepietra is also noticeable. To obtain a quantitative assessment of the diversity among Italian and European populations, we carried out an AMOVA using an extended dataset (see Supplementary tables 5 and 6). Including geographically and/or isolated groups in the analyses, we observed increased estimates of diversity among Italians (mtDNA from 0.38% to 1.89%; Y chromosome from 3.19 to 6.89%) which were higher than for Europeans (mtDNA from 0.33% to 1.52%; Y chromosome 8.95% from to 10.60%). It seems, therefore, that inclusion of isolates in the dataset makes the extent of diversity of Italian populations slightly higher than Europeans for mtDNA and reduces the negative difference observed for Y chromosome. However, given the substantial difference in the ratio between open and isolated groups and spatial dispersion between Italy and Europe, we thought it useful to compare genetic and geographic distances using box plots of quartile distributions (Fig. 9). In the case of mtDNA, genetic distances of Italian isolates between each other and with Italian open populations are characterized by wider distributions and lower median values than Europeans for 7 and 8 ranges of geographic distances (out of 10), respectively. However, a high level of differentiation among Italian populations is signalled by the outlying genetic distances observed at practically all geographic ranges. Although this finding may be again related to the availability of a larger number of data on Italian compared to European isolates, it should be noted that large genetic distances
Fig. 7 - Scatter plot of haplotype diversity and average Fst of isolated populations for (A) mtDNA and (B) Y chromosome (LUS not shown HD=0.590; average Fst= 0.322). Dotted lines represent the boundaries defined by the analysis of outliers (interquartile method). Symbols and colors as in Fig. 6. MtDNA labels: 1=ARC, 2=ARS, 3=BEN, 4=CFT, 5=CIR, 6=CRM, 7=FER, 8=FIE, 9=GIU, 10=GRC, 11=JEN, 12=LES, 13=LUS, 14=LVB, 15=LVF, 16=LVG, 17=NON, 18=PRI, 19=SAO, 20=SAP, 21=SAU, 22=SOL, 23=TIM, 24=TOC, 25=VAG, 26=VAP, 27=VAV, 28=VIS, 29=VPU. Y chromosome labels: 1=ARC, 2=ARS, 3=BEN, 4=CFT, 5=CIR, 6=CRM, 7=FER, 8=FIE, 9=GIU, 10=GRC, 11=LES, 12=LUS, 13=LVB, 14=LVF, 15=LVG, 16=NON, 17=PRI, 18=SAP, 19=SAU, 20=SOL, 21=TIM, 22=VAG. The colour version of this figure is available at the JASS website.
Fig. 8 - Multi-Dimensional Scaling plot of Fst genetic distances among Italian and European populations based on (A) mtDNA HVR-1 sequences (stress value=0.226) and (B) 5 Y chromosome STRs (stress value=0.183). Filled and empty dots represent European and Italian populations, respectively. Population abbreviations as in Supplementary Table S3 and S4.
Fig. 9 - Box plot analyses performed on Fst values within different ranges of geographic distances (expressed as thousands of kilometers) for mtDNA (A) and Y chromosome (B). The light gray boxes describe the distribution of all pairwise genetic distances between Italian isolates and those between Italian isolates and open Italian populations; the dark gray boxes indicate the distribution of pairwise genetic distances between European isolates and those between European isolates and open European populations.
between Italian groups are observed even at short geographic distances (<200 km). As predicted by previous analyses (Fig. 7), an important contribution to this finding comes from ethno-linguistic minorities of the Eastern Alps. Another range of geographic distances in which the signal of heterogeneity of Italian populations is more marked is between 700 and 800 km. Here, a substantial effect comes from the genetic distances of Sardinian isolates (Benetutti, Carloforte and Ogliastra). Results obtained using Y chromosome data point to a closer correlation between genetic and geographical distances ($r^2$ value of 0.226, $p<0.001$, for European populations), which is in line with previous observations at a wider geographic scale (Rosser et al., 2000). As observed for mtDNA, a greater differentiation is shown by Italian populations in the ranges between 0-200 and 700-800 km, with a higher frequency of extreme values for all intervals of geographic distances.

Concluding remarks

Our study moves from the assumption that human biological variation is an integral part of what we mean by the term biodiversity. This concept may seem self-evident. However, it is contradicted by many examples in which biodiversity is thought to be necessarily linked to conservation, rather than a value per se, while human factors are considered only in relation to anthropic effects or educational aspects (e.g. Johns, 2009; Marton-Lefèvre, 2010). Not less importantly, the study of human biodiversity is often perceived as a difficult, if not risky, task, given the long tradition of misuse of this concept which still survives in bio-medical literature (e.g. Burchard et al., 2003).

Once shortcomings and prejudices have been removed, the question of whether human genetic diversity observed in a certain territory shows similarities with other layers of biodiversity becomes worth answering. Certainly, comparing observations on a large number of species studied at various ecological scales using different approaches (e.g. morphological and genetic) with results from populations of a single species could appear to be simplistic. Obviously, our survey is intended to be just a first step towards further studies driven by explicit hypotheses regarding the impact of environmental and demographic factors on different layers of biodiversity. We also maintain that studying human populations in the framework of biodiversity has an intrinsic value, since it may help us gain insights into the impact of socio-cultural factors on human genetic variation and on biological diversity as a whole.

The results of this collaborative study may be summarized in two main points. We could shed light on the genetic structure of most of the Italian ethno-linguistic minorities, showing that a combination of linguistic and geographic factors is probably responsible for the presence of the most robust signatures of genetic isolation. Furthermore, drawing a picture of Italian populations which also includes linguistic and/or geographic isolates allowed us to better appreciate the noticeable extent of the genetic diversity of our country. It turned out to be slightly greater than that of Europeans for maternal lineages on the whole, and at specific ranges of geographic distances for both genetic markers. Therefore, our survey provides a first affirmative response to the question of whether the genetic diversity of human populations shows a pattern which is comparable to what has been observed in our country for plant and (the rest of) animal biodiversity. The longitudinal extension and the nature of the natural bridge between Central Europe and the Mediterranean seems to provide an explanation also to the high level of inter-population variation. In fact, some of the isolates which determine the noticeable diversity we observed in our study are located along the boundaries of Italian territory. Furthermore, the small demographic size of some of these groups has probably increased their departure from the surrounding genetic background. From an anthropological point of view, the finding that an important contribution to the genetic diversity of Italy comes from the so called “linguistic islands” (e.g. German speaking groups of Sappada and Luserna from the Eastern Italian Alps) is of
particular significance. In fact, this is a further proof of how taking into account social and cultural factors may help us understand the structure of human genetic variation more in depth.

Obviously, our study should be seen as a first step towards a more complete assessment. Extending the sampling to the few linguistic isolates that have yet to be included in the survey (e.g. Walser, Occitan and Franco-Provençal linguistic minorities from the Western Alps) could offer an even more detailed picture. Not less importantly, some limitations of unilinearly transmitted polymorphisms (e.g. their nature of single loci and the possible confounding effect of small sample sizes) might be overcome by increasing their resolution or, even better, exploiting the power of large panels of SNPs. Our research network has already planned to deepen the analysis of the genetic structure of isolated populations using a genome-wide approach implemented in the new genotyping array “GenoChip” (Elhaik et al., 2013). Enforcing the interdisciplinary collaborations already initiated in the course of this study (e.g. Fiorini et al., 2007; Robledo et al. 2012; Coia et al., 2013) will be a key-step to really take advantage of these new genomic data.

Author Contributions

Designed the research: MCP CMC VC GDB DP GP GV ST.
Collected the samples: PA AB CB GB SB CMC LC MCP VC FC GDB ZAF DL LM GP OR RR ES LS SS GT ST GV.
Performed the experiments: MA CB FB IB SB VB VC MCP MCR SDF GF PF SS.
Analyzed the data: AB PA CMC MCP VD FM LT ST.
Wrote the paper: PA AB CMC MCP GDB DP ST GV.
Read and approved the manuscript: All authors.

Acknowledgements

This paper is dedicated to our friend and colleague Laura Morelli. We would like to thank: the medical and nursing staff of the Azienda Provinciale per i Servizi Sanitari, Trento; the volunteers of the Associazione Volontari Italiani Sangue del Trentino (AVIS), Biblioteca Intercomunale di Primiero, Istitut Cultural Ladin and Kulturinstitut Lusern; all the people who kindly contributed to the collection of
samples in Trentino (http://laboratoriobagolini.it); Marcella Benedetti (Municipality of Sappada), Nino Pacilè and Lucia Protop (Municipality of Saurus), Vito Masalongo (Giazza), Ottaviano Matiz and Velia Plozner (Timau) for their warm assistance in sample collection and hospitality; all the people from Benetutti, Circello, Carloforte and Vagli who agreed to participate to this study; Silvano Zaccone (Consorzio “Il Cigno”), Sauro Mattei, Alfonso Tatavitto and Massimo Del Grosso for their fundamental help in collecting samples in Samnite and Apuan areas; Lorenzo Marcuccetti for his commentary to Ligures’ history and linguistics; Michele Musacchia (Azienda Sanitaria Locale 4, Corleone) and Giorgio Ales (MD, Piana degli Albanesi) for their help in collecting Sicilian Arbereshe.

This research was made possible by grants from the Italian Ministry of Education, University and Research (PRIN projects 2007 and 2009, Principal Investigator DP). Research work in the Eastern Alps was further supported by the Istituto Italiano di Antropologia (project “Atlante bioculturale Italiano”; funds to GDB), the University of Rome “La Sapienza” (project “L’isolamento genetico in popolazioni europee”, funds to GDB) and the Provincia Autonoma di Trento (Project “Biostre”; funds to VC).

References


Côte-Real H.B., Macaulay VA., Richards M.B., Hariti G., Issad M.S., Cambon-Thomsen A.,


Frigi S., Pereira P., Pereira L., Yacoubi B., Gusmao L., Alves C., Khodjet el Khil H.L., Cherni A., Amorim A. & El Giaied A. 2006. Data for Y-chromosome haplotypes defined by 17 STRs...


Walsh B. 2001. Estimating the time to the most recent common ancestor for the Y chromosome or mitochondrial DNA for a pair of individuals. *Genetics*, 158: 897-912.


Associate Editor, Rita Vargiu
Exploring biodiversity of Italian populations

Appendix - Other unpublished and literature data used in this study. Abbreviations: Alt, altitude; G, geographic isolate; GL, geo-linguistic isolate; L, linguistic isolate; O, open population.

<table>
<thead>
<tr>
<th>POPULATION</th>
<th>ABB.</th>
<th>STATUS</th>
<th>CENSUS*</th>
<th>ALT.°</th>
<th>mtDNA N</th>
<th>REFERENCE</th>
<th>Y CHROMOSOME N</th>
<th>REFERENCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adige Valley</td>
<td>TVA</td>
<td>O</td>
<td>166394</td>
<td>199</td>
<td>56</td>
<td>Coia et al., 2012</td>
<td>56</td>
<td>Coia et al., 2013</td>
</tr>
<tr>
<td>Ascoli Piceno</td>
<td>PIC</td>
<td>O</td>
<td>49892</td>
<td>154</td>
<td>53</td>
<td>Brisighelli et al., 2012</td>
<td>38</td>
<td>Brisighelli et al., 2012</td>
</tr>
<tr>
<td>Aviano</td>
<td>AVI</td>
<td>O</td>
<td>9270</td>
<td>159</td>
<td>29</td>
<td>Boattini et al., 2013</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Belvedere</td>
<td>BEL</td>
<td>O</td>
<td>9388</td>
<td>150</td>
<td>-</td>
<td>-</td>
<td>27</td>
<td>Brisighelli et al., 2012</td>
</tr>
<tr>
<td>Benevento</td>
<td>BNV</td>
<td>O</td>
<td>61700</td>
<td>207</td>
<td>36</td>
<td>Boattini et al., 2013</td>
<td>33</td>
<td>Boattini et al., 2013</td>
</tr>
<tr>
<td>Bologna</td>
<td>BOL</td>
<td>O</td>
<td>373020</td>
<td>54</td>
<td>100</td>
<td>Bini et al., 2003</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Bolzano</td>
<td>BOZ</td>
<td>O</td>
<td>105774</td>
<td>262</td>
<td>59</td>
<td>Thomas et al., 2008</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Brescia</td>
<td>BRE</td>
<td>O</td>
<td>188602</td>
<td>149</td>
<td>40</td>
<td>Boattini et al., 2013</td>
<td>35</td>
<td>Boattini et al., 2013</td>
</tr>
<tr>
<td>Campobasso</td>
<td>CMB</td>
<td>O</td>
<td>48479</td>
<td>701</td>
<td>37</td>
<td>Boattini et al., 2013</td>
<td>29</td>
<td>Boattini et al., 2013</td>
</tr>
<tr>
<td>Casentino</td>
<td>CAS</td>
<td>O</td>
<td>46039</td>
<td>465</td>
<td>122</td>
<td>Achilli et al., 2007</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Como</td>
<td>COM</td>
<td>O</td>
<td>83132</td>
<td>201</td>
<td>39</td>
<td>Boattini et al., 2013</td>
<td>41</td>
<td>Boattini et al., 2013</td>
</tr>
<tr>
<td>Croatian Molise</td>
<td>CRM</td>
<td>L</td>
<td>1884</td>
<td>493</td>
<td>41</td>
<td>Babalini et al., 2005</td>
<td>15</td>
<td>This study</td>
</tr>
<tr>
<td>Cuneo</td>
<td>CUN</td>
<td>O</td>
<td>55627</td>
<td>534</td>
<td>40</td>
<td>Boattini et al., 2013</td>
<td>30</td>
<td>Boattini et al., 2013</td>
</tr>
<tr>
<td>Fersina Valley</td>
<td>FER</td>
<td>G</td>
<td>51478</td>
<td>1066</td>
<td>25</td>
<td>Coia et al., 2012</td>
<td>26</td>
<td>Coia et al., 2013</td>
</tr>
<tr>
<td>Foligno</td>
<td>FOL</td>
<td>O</td>
<td>58363</td>
<td>234</td>
<td>-</td>
<td>-</td>
<td>29</td>
<td>Boattini et al., 2013</td>
</tr>
<tr>
<td>Giudicarie Valley</td>
<td>GIU</td>
<td>G</td>
<td>36282</td>
<td>680</td>
<td>52</td>
<td>Coia et al., 2012</td>
<td>51</td>
<td>Coia et al., 2013</td>
</tr>
<tr>
<td>Greccanici Salento</td>
<td>GRC</td>
<td>L</td>
<td>10000</td>
<td>80</td>
<td>47</td>
<td>Brisighelli et al., 2012</td>
<td>46</td>
<td>Brisighelli et al., 2012</td>
</tr>
<tr>
<td>Isarco Valley</td>
<td>VIS</td>
<td>G</td>
<td>44500</td>
<td>801</td>
<td>34</td>
<td>Pichler et al., 2006</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Jenne</td>
<td>JEN</td>
<td>G</td>
<td>407</td>
<td>834</td>
<td>103</td>
<td>Messina et al., 2010</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>La Spezia</td>
<td>SPZ</td>
<td>O</td>
<td>95378</td>
<td>3</td>
<td>50</td>
<td>Brisighelli et al., 2012</td>
<td>45</td>
<td>Brisighelli et al., 2012</td>
</tr>
<tr>
<td>Ladins (Badia Valley)</td>
<td>LVB</td>
<td>GL</td>
<td>10644</td>
<td>1341</td>
<td>56</td>
<td>Thomas et al., 2008</td>
<td>44</td>
<td>Coia et al., 2013</td>
</tr>
<tr>
<td>Ladins (Gardena Valley)</td>
<td>LVG</td>
<td>GL</td>
<td>10198</td>
<td>1320</td>
<td>46</td>
<td>Thomas et al., 2008</td>
<td>51</td>
<td>Coia et al., 2013</td>
</tr>
<tr>
<td>L'Aquila</td>
<td>AQU</td>
<td>O</td>
<td>110268</td>
<td>714</td>
<td>25</td>
<td>Boattini et al., 2013</td>
<td>27</td>
<td>Boattini et al., 2013</td>
</tr>
<tr>
<td>Lazio</td>
<td>LAZ</td>
<td>O</td>
<td>5536292</td>
<td>-</td>
<td>52</td>
<td>Babalini et al., 2005</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Lecce</td>
<td>LEC</td>
<td>O</td>
<td>89839</td>
<td>49</td>
<td>39</td>
<td>Boattini et al., 2013</td>
<td>35</td>
<td>Boattini et al., 2013</td>
</tr>
</tbody>
</table>
## Appendix - Continued.

<table>
<thead>
<tr>
<th>POPULATION</th>
<th>ABB.</th>
<th>STATUS</th>
<th>CENSUS*</th>
<th>ALT.°</th>
<th>mtDNA N</th>
<th>mtDNA REFERENCE</th>
<th>Y CHROMOSOME N</th>
<th>Y CHROMOSOME REFERENCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Macerata</td>
<td>MAC</td>
<td>O</td>
<td>43019</td>
<td>315</td>
<td>39</td>
<td>Boattini et al., 2013</td>
<td>34</td>
<td>Boattini et al., 2013</td>
</tr>
<tr>
<td>Matera</td>
<td>MAT</td>
<td>O</td>
<td>59973</td>
<td>401</td>
<td>36</td>
<td>Boattini et al., 2013</td>
<td>24</td>
<td>Boattini et al., 2013</td>
</tr>
<tr>
<td>Non Valley</td>
<td>NON</td>
<td>G</td>
<td>37832</td>
<td>816</td>
<td>48</td>
<td>Coia et al., 2012</td>
<td>48</td>
<td>Coia et al., 2013</td>
</tr>
<tr>
<td>Ogliastra</td>
<td>SAO</td>
<td>G</td>
<td>57959</td>
<td>392</td>
<td>175</td>
<td>Fraumene et al., 2003</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Oristano</td>
<td>ORI</td>
<td>O</td>
<td>32156</td>
<td>5</td>
<td>39</td>
<td>Boattini et al., 2013</td>
<td>40</td>
<td>Boattini et al., 2013</td>
</tr>
<tr>
<td>Primiero Valley</td>
<td>PRI</td>
<td>G</td>
<td>9959</td>
<td>763</td>
<td>40</td>
<td>Coia et al., 2012</td>
<td>41</td>
<td>Coia et al., 2013</td>
</tr>
<tr>
<td>Pusteria Valley</td>
<td>VPU</td>
<td>G</td>
<td>73000</td>
<td>997</td>
<td>37</td>
<td>Pichler et al., 2006</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Sole Valley</td>
<td>SOL</td>
<td>G</td>
<td>15235</td>
<td>884</td>
<td>63</td>
<td>Coia et al., 2012</td>
<td>65</td>
<td>Coia et al., 2013</td>
</tr>
<tr>
<td>Terni</td>
<td>TER</td>
<td>O</td>
<td>109482</td>
<td>130</td>
<td>31</td>
<td>Boattini et al., 2013</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Tocco da Casauria</td>
<td>TOC</td>
<td>G</td>
<td>2782</td>
<td>356</td>
<td>50</td>
<td>Verginelli et al., 2003</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Treviso</td>
<td>TRE</td>
<td>O</td>
<td>82535</td>
<td>15</td>
<td>39</td>
<td>Boattini et al., 2013</td>
<td>28</td>
<td>Boattini et al., 2013</td>
</tr>
<tr>
<td>Udine</td>
<td>UDI</td>
<td>O</td>
<td>100514</td>
<td>113</td>
<td>51</td>
<td>Brisighelli et al., 2012</td>
<td>45</td>
<td>Brisighelli et al., 2012</td>
</tr>
<tr>
<td>Vallepietra</td>
<td>VAP</td>
<td>G</td>
<td>308</td>
<td>825</td>
<td>21</td>
<td>Messina et al., 2010</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Venosta Valley</td>
<td>VAV</td>
<td>G</td>
<td>34307</td>
<td>1024</td>
<td>112</td>
<td>Pichler et al., 2006</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Vicenza</td>
<td>VIC</td>
<td>O</td>
<td>113639</td>
<td>39</td>
<td>40</td>
<td>Boattini et al., 2013</td>
<td>33</td>
<td>Boattini et al., 2013</td>
</tr>
</tbody>
</table>

* Source: ISTAT (2011) (http://demo.istat.it)
° meters above sea level.