9-1-2012

# Analysis of a Genetic Isolate: The Case of Carloforte (Italy)

R. Robledo
*Department of Biomedical Sciences, University of Cagliari*

L. Corrias
*Department of Life and Environment Sciences, University of Cagliari*

V. Bachis
*Department of Life and Environment Sciences, University of Cagliari*

N. Puddu
*Department of Philology, Literature and Linguistics, University of Cagliari*

A. Mameli
*Carabinieri, Reparto Investigazioni Scientifiche (RIS), Cagliari*

***See next page for additional authors***

**Authors**
R. Robledo, L. Corrias, V. Bachis, N. Puddu, A. Mameli, G. Vona, and C. M. Calò

# Analysis of a Genetic Isolate: The Case of Carloforte (Italy)

R. ROBLEDO[1], L. CORRIAS[2], V. BACHIS[2], N. PUDDU[3], A. MAMELI[4], G.VONA[2], and C. M. CALÒ[2]*

1. Department of Biomedical Sciences, University of Cagliari
2. Department of Life and Environment Sciences, University of Cagliari
3. Department of Philology, Literature and Linguistics, University of Cagliari
4. Carabinieri, Reparto Investigazioni Scientifiche (RIS), Cagliari

*Correspondence to: C. M. Calò, Department of Life and Environment Sciences, University of Cagliari, Monserrato 09042. Email: cmcalo@unica.it

KEY WORDS: GENETIC ISOLATE, LANGUAGE VARIETY, ENDOGAMY, FOUNDER EFFECT, Y CHROMOSOME MARKERS, mtDNA, HAPLOGROUPS, SAMPLING STRATEGY, ITALY, CARLOFORTE

**Running title:** Carloforte (Italy), a genetic isolate

**Abstract**

We reviewed data collected during several studies concerning the genetic isolate of Carloforte (Sardinia, Italy) and analyzed new data on Y-chromosome markers. Carloforte is also a language island, where people still speaks Tabarchino, an archaic form of Ligurian dialect. Demographic data indicate that, in the early years of its history, Carloforte population was characterized by a high degree of endogamy and consanguinity rates that started to decrease around 1850, when marriages with Sardinian people began to occur more frequently. Cultural factors, mainly language, account for the high endogamy. Genetic data from classical markers, mtDNA and Y-chromosome markers confirmed the strong isolation of Carloforte population, which appears significantly different from the neighboring population of Sardinia. Analysis of mtDNA emphasizes the crucial aspect of sampling strategy: two different samplings of the same population, one based on founder surnames, while the other based on grandparents' criterion, gave different results. Founder surnames sampling is not affected by recent events, and therefore better describes the ancestral population. Whereas, grandparents' criterion sampling gives a picture of the present population, shaped by more recent events, like migration and gene flow. This review further supports the notion that a comprehensive approach, including a detailed knowledge of the history of the population and the collection of different samplings, is essential in anthropology for reconstructing past and recent events that contributed to establish the present genetic structure of the population. Likewise, it is essential in medical genetics to identify genes involved in complex diseases. An ideal scenario is offered by a genetic isolate with a recent, and well documented, history, like Carloforte, which can be a paradigm for this type of investigations.

## Introduction

Genetic isolates are subpopulations derived from a small number of founders that have been isolated, due to geographical and/or cultural barriers, for many generations with very restricted genetic exchange with other subpopulations (Arcos-Burgos and Muenke 2002). Population isolates have long been a subject of interest in different areas, like anthropology, population and medical genetics. In anthropology, genetic isolates proved to be very useful to identify peopling events, past migrations, demographic behavior, cultural and linguistic features (Colonna et al. 2007; Boattini et al. 2011). In population genetics, the reduced genetic diversity, due to the limited number of founders, together with the availability of detailed historical and genealogical records, made isolated populations ideal in identifying which factors account for genome variation, like assortative mating, genetic drift, bottleneck effect, etc. (Pardo et al. 2005; Colonna et al. 2007; Pope et al. 2011).

Genetic isolate are also very important in medical genetics studies: indeed, isolated populations, like Finnish, Old Order Amish, Jewish, Sardinian, have proved to be an invaluable resource for mapping genes involved in rare diseases showing a Mendelian recessive mode of inheritance. Rare alleles are detectable in isolated population derived from a small number of founders, where such rare alleles (if present in the founding group) are enriched, thus producing homozygous affected individuals. Several genes were mapped and identified, through linkage analysis (Nikali et al. 1995) or homozygosity mapping (Neufeld et al. 1997) in such populations. More recently, the same populations proved to be advantageous also for identifying alleles involved in complex diseases, provided that an in-depth knowledge of the genetic structure is available. Genome-wide association (GWA) studies, based

on single nucleotide polymorphisms (SNPs), have being used to detect haplotypes shared in complex traits (Kristiansson et al. 2008). In an isolated population a reduced number of segregating haplotypes are expected. Whereas, in outbred populations a causative allele is likely to be present in several distinct haplotypes, therefore diluting its signal, and it may be overlooked.

An additional advantage is provided by population isolates with a recent history. The knowledge of the number of founders, the demographic history, the degree of genetic and cultural isolation, and the availability of detailed records, allow a multi-disciplinary approach in determining the features of the genetic landscape of a genetic isolate. The importance of a comprehensive approach in this type of analysis is confirmed by recent works, based of polymorphic loci with different inheritance (Y chromosome, mitochondrial DNA, autosomes), that support the role played by socio-cultural factors in determining patterns of sex-related gene flow (Oota et al. 2001; Destro-Bisol et al. 2008; Marchani et al. 2008). In this regards, Sardinia represents an ideal scenario, since the island has been a theater of intensive studies in different research areas and much is known about its archaeology, history, language and genetics (Cavalli Sforza et al. 1994). Among the different aspects of human culture, language is probably the most important one, playing a major role in the processes of cultural learning, transmission and communication (Destro Bisol et al. 2008). Despite this, few studies have systematically addressed the influence of language on the genetic structure of human populations using polymorphisms with different inheritance, like uniparental or biparental markers. A more focused insight into the role of language could be provided by the investigation of linguistic isolates. Carloforte is an excellent example, being, at the same time, a genetic isolate and a

language island. Moreover, detailed historical and demographical information since the foundation of Carloforte is available.

In the present review, we report and discuss data previously published on Carloforte, concerning the areas of linguistics, biodemography, classical markers, autosomal markers and mtDNA (see methods for references). In addition, we also present and analyze new data on Y-chromosome short tandem repeats (STRs).

***A brief history of Carloforte***

Carloforte is the only town located on the small island of San Pietro, off the Southwestern coast of Sardinia (Italy), at 39°8' N 8°18' E, with a surface area of 50.24 $km^2$. Its distinctive history begins in 1738 when, on April 4$^{th}$, a group of 467 people started settling the deserted island (Ferraro 1989). The great majority, 388 people, arrived on San Pietro after leaving the island of Tabarka, in Tunisia (Fig. 1). They were descendants of Ligurian migrants who had left the small town of Pegli, now part of the city of Genoa in Northern Italy, around 1540 to colonize the uninhabited island of Tabarka, where they had been running a successful business of coral and tuna fishing.

During the two centuries spent in Tabarka, the community prospered and reached about 2,000 individuals. Due to cultural barriers, including language and religion, throughout that time those Genoese migrants kept themselves culturally, as well as genetically, separate. There is no evidence of interbreeding with the neighboring populations of North Africa (Vallebona 1988). At the beginning of XVIII century, clear signs of decline were evident: the impoverishment of the coral reefs severely affected the community, whose economy was dependent on coral fishing. Additional factors were repeated incursions by pirates and the deteriorated relationship with the Beys of Tunis and Algier, to whom the community regularly paid an annual fee.

Around that time, King Carlo Emanuele III of Savoy, ruling the kingdom of Piedmont and Sardinia, granted their request to colonize San Pietro; therefore some members of the community left Tabarka and arrived on the island together with 79 additional people from Pegli. On San Pietro they founded a village, which they named Carloforte, expressing gratitude to King Carlo Emanuele III (Vallebona 1988; Ferraro 1989). Later on, additional families left Tabarka: the majority joined their relatives in Carloforte, while other families established themselves in Sant'Antioco (Sardinia), facing the island of San Pietro, where they founded Calasetta (fig. 1). The last families to leave Tabarka, around 1770, were about 400 people that have been kidnapped and enslaved first by the Bey of Tunis, and then kept enslaved by the Bey of Algier. They were freed by King Carlos III of Spain who paid the ransom and moved them to Spain, where they colonized Illa de Sant Pau (island of S. Paul), renamed Nueva Tabarca.

The population of Carloforte has been the focus of several studies by the scientific community and it has been analyzed from different points of view: anthropological, demographic, genetic and linguistic (Vona et al. 1996; Sanna et al. 2006). All studies showed a strong degree of isolation and revealed genetic and linguistic features that differentiate the Carloforte population from the Sardinia population: indeed, the community of Carloforte still speaks today a variety of Ligurian, called Tabarchino, which is completely different from any other language variety spoken in Sardinia. Therefore, Carloforte is a language island and a genetic isolate.

According to the 2001 census, the population of Carloforte reached 6,444 inhabitants. Some are descendants of the 126 founder families, others are descendants of migrants from neighboring Sardinian populations that, starting from the half of the 19th century, moved to San Pietro (Calò et al. 2012). The availability of historical

documents listing the surnames of the founding families, and the availability of genealogical records, make it very easy to identify any individual as a descendant from either a founder, or a Sardinian or a mixed population. Indeed, like the language variety, surnames of the founder families (from Northern Italy) are totally different from Sardinian families' surnames.

**Methods**

Methodologies regarding classical markers (Vona et al. 1996), autosomal markers (Robledo et al. 2009), mtDNA (Calò et al. 2012), biodemography (Vona et al. 1996; Latini et al. 2004; Sanna et al. 2006), and linguistics (Toso 2003) were previously described. To analyze Y-chromosome markers, DNA was extracted from peripheral blood or buccal swab and quantified by Quantiblot human DNA quantification kit (Applied Biosystems). Amplification of 0.5 ng template DNA was carried out with the AmpFlSTR Yfiler PCR amplification kit (Applied Biosystems) in a Geneamp®PCR System 9700 thermal cycler, using the thermal cycling conditions recommended by the manufacturer. Separation and detection of the Y-STR multiplex PCR products were accomplished with the ABI Prism 3100 Genetic Analyzer sixteen-capillary array system (Applied Biosystems). Genotypes were determined by GeneScan v. 3.1. Y-haplogroups were inferred by subjecting each haplotype to the software package Haplogroup Predictor (http://www.hprg.com/hapest5/) (Athey et al. 2005). Statistical analyses were performed using Arlequin program, ver. 3.5, (Excoffier and Lischer 2010) for calculating Slatkin's genetic distances (Slatkin 1995), and Phylip, ver. 3.76, (Felsenstein 1989) for examining relationship of Y-chromosome STRs among Sardinian and other Italian and Mediterranean populations

through Neighbor Joining Tree. The robustness of the tree was assessed by bootstrap analysis.

## Results

### *Linguistic data*

Carloforte, together with nearby Calasetta, represents an interesting case for studying linguistic variation in relation to geographic boundaries and social context. In Sardinia, Italian is the official language but a minority variety called Tabarchino survives in this area. Tabarchino, which is part of the Ligurian group of Romance dialects, has a very high diffusion among Carloforte and Calasetta people. According to a socio-linguistic enquiry (Sitzia et al. 1998), in Carloforte 87% of adults and 72% of children (from 6 to 15 years) declared both active and passive competence in Tabarchino. In Calasetta, where mixed marriages between Tabarchino and Sardinian speakers are more common, these percentages decrease to 65% for adults and 62% for children. These high percentages indicate that Tabarchino has high status within the two towns and this is for two main reasons. First, as minority languages usually do, it has a strong identity value since it clearly differentiates the Ligurian speakers from the Campidanese Sardinian speakers. Second, Tabarchino, during the centuries, held a crucial value, since the main economic activity, especially of Carloforte people, was tuna fishing and trading with Genoa (Toso 2003). Consequently, Tabarchino speakers considered urban Genoese as a "high" model to conform with, and these contacts had more influence on Tabarchino than the contacts with Campidanese Sardinian.

On the other hand, inhabitants of Calasetta, who were mainly vine growers, had a relatively higher degree of integration with the Sardinian community. This exposed

the Calasetta variety to a higher level of borrowing from Campidanese Sardinian (Blasco Ferrer 1994; Toso 2003).

However, the Sardinian influence on Tabarchino is limited mainly to the lexical level, especially in agricultural terms, plant names and interjections (Toso 2003). From a phonetic and morpho-syntactic point of view, Tabarchino clearly continues to pertain to Ligurian dialects. In particular, Toso (2003) considers Tabarchino an example of "Colonial Genoese". Colonial varieties are usually considered conservative, since they maintain forms long after their obsolescence in the mother language (this is the case, for instance, of Bonifacian, another Ligurian dialect that is spoken in Corsica; see Bottiglioni 1928). On the one hand, Tabarchino preserves some archaic and rural features, which derive directly from the language spoken by the settlers who came mainly from Pegli. On the other hand, as Toso (2003) clearly shows from a phonetic and morpho-syntactic point of view, Tabarchino cannot be considered truly archaic because it is deeply influenced by urban Genoese. It also has some peculiar innovations, and it seems that these have increased quite recently when direct contact with Genoa started to decrease. At the lexical level Tabarchino preserves some archaic words, but overall it reflects the 18$^{th}$-19$^{th}$ century Genoese lexicon (Toso 2003). Instead, Tabarchino of Calasetta is richer in Sardinian borrowings because of the geographic and economic closeness to the Sardinian community of Sant'Antioco, but even in this case the core lexicon is undoubtedly Ligurian.

The case of Tabarchino confirms the importance of extra-linguistic factors, like geographic boundaries and socio-economical relationships, in explaining language variation. Moreover, Tabarchino is a very good example of how language is a powerful tool to preserve identity and how a minority variety can successfully survive when its status is high.

Perhaps surprisingly, Tabarchino did not survive in Nueva Tabarca. As reported by Ferraro (1989), already in 1975, about two centuries after colonization, Tabarchino was not spoken, apparently being lost with the last generation. The community has completely blended with the local communities, acquiring Valencian, a variety of Catalan. A likely interpretation is that, in Nueva Tabarca massive contact with the Catalan speaking community, together with the weakening of relationship with the motherland, led to a progressive loss of Tabarchino, which had substantially disappeared in the twentieth century (Toso 2003).

***Endogamy and consanguinity***

Biodemographic data were collected from the Parish registers (Quinque Libri) kept in the Episcopal administration in Iglesias and from the Parish of Carloforte, for the period from the city foundation in 1738 to 2001. The number of marriages per year, number of marriages between people related by blood with their degree of kinship, and place of birth of the spouses were recorded (Vona et al. 1996; Sanna et al. 2006). Endogamy and exogamy rate were studied from 1738 to 2001 (Sanna et al. 2006; Calò and Vona 1994). The data are summarized in figure 2. During those 264 years, a total of 8,322 marriages were celebrated in Carloforte. Endogamy rate for the whole period was 74.26%, while the exogamy rate was 25.74%. The exogamy was not constant, but varied from a low of 8.05% (1825-49) to a maximum of 57.44% (1975-2001). Exogamy was high at the two extremes of the period examined (44.44% and 57.44% respectively): in the early years (1738-1749) it was due to the arrival of additional founders coming mainly from Pegli, whereas in the years 1975-2001 it was due to increasing marriages with Sardinians. Marriages between spouses who are

both non-local are 5.02%. Marriages between descendants of Carloforte founders and Sardinians have increased steadily: from 7.14% (1738-49) to 79.30% (1950-93).

The average total matrimonial distances, over the entire time period, is 42.36 km, with an extreme value of 193.77 km between 1940 and 1944. However, if we consider exclusively the weddings between people from Carloforte and Sardinians, the total average value decreases to 7.17 km.

Spouses in exogamous marriages are predominantly from the areas of Sulcis-Iglesiente and Campidano in Sardinia, but also from regions with a longstanding tradition of fishing, like Liguria, Campania and Sicily. Over time, however, provenance of spouses progressively narrows down to Sardinia only. From the foundation of Carloforte in 1738 there has been an elevated percentage of marriages between blood relatives, with a peak of 56.90% reached in the years 1775-79. From 1960 there are no records of marriages between blood relatives. The average value of Bernstein's coefficient, calculated over the entire period, is $\alpha=1.63 \times 10^{-3}$ while the highest value was recorded between 1785 and 1789 with $\alpha=5.17 \times 10^{-3}$. The main contribution to the α value was provided by marriages between first cousins and by marriages with multiple consanguinity (Calò and Vona 1994). Comparing the values of both endogamy and consanguinity in Carloforte with other Sardinian villages (Moroni et al. 1972; Morelli and Vona 1993; Calò and Vona 1994) Carloforte shows values greater than most Sardinian villages. The α value of Carloforte is slightly higher than the one reported for the whole of Sardinia ($\alpha=1.54 \times 10^{-3}$) for the period 1765-1969 (Moroni et al. 1972).

***Isonymy***

The population of Carloforte has been examined for the comparison between expected and observed marital isonymy using surnames (Latini et al. 2004). The maximum value of observed isonymy (0.018) was found in 1900-1949. Until 1850-1899 the expected isonymy assuming random mating showed higher values than the observed isonymy, while from 1900 the expected isonymy decreases, reflecting the introduction of new surnames from outside, and the opening of the population toward marriages with individuals belonging to other families (Fig. 3).

The total inbreeding coefficient calculated from isonymy values (Fit) shows a trend similar to observed isonymy, with the highest value (0.0045) between 1900 and 1949 (Fig. 4). The random component (Fr) is the major contributor in determining the total inbreeding until 1850, demonstrating the absence of intentionality in the choice of partner. So the high value of consanguinity observed until 1850 is determined by the isolation of Carloforte. On the contrary, starting from 1850, an intentional choice of the partner within the group seems to be prevalent: indeed, the non random component (Fn) values increase to reach the highest values (0.0018) between 1900-1949 (Calò and Vona 2005).

Values of inbreeding calculated by consanguinity and isonymy are not identical: inbreeding values estimated by isonymy are higher than inbreeding estimated by consanguinity determined by pedigree analysis, in agreement with previous research (Gagnon and Toupance 2002).

***Classical markers***

Analysis of 12 classical markers, blood groups AB0, Rh, MN; erythrocyte enzymes phosphogluconate dehydrogenase (PGD), adenylate kinase (AK), phosphoglucomutase (PGM1), diaphorase (DIA), acid phosphatase 1 (ACP1),

esterase D (ESD), and serum proteins haptoglobin (HP), vitamin D-binding protein (GC) and complement protein (C3), shows that allele frequencies in Carloforte's population are clearly different from those observed in Sardinian and Italian populations (Vona et al. 1996). Indeed, a neighbor-joining tree (Fig. 5) shows that the greatest degree of divergence is between Sardinians and continental Italians, whereas Carloforte's population lies midway.

In particular, comparison between Sardinians and Carloforte for blood groups, shows a higher frequency of haplotype RH*cde and a lower frequency of alleles MN*M, ESD*1, and haplotype RH*CDe in Carloforte. The frequency of PGM1*2S and HP*1 in Carloforte are above the upper limit of the frequency reported for Sardinians. On the other hand, the frequency of GC*1S in Carloforte is below the lower limit of the frequency observed in Sardinians (Tab. 1). Differences between Carloforte and other Sardinian populations for RH, ESD, and GC systems are highly significant ($p<0.001$) (Vona et al. 1996).

***Autosomal DNA markers***

The Carloforte population was also analyzed for autosomal markers: 31 Short Tandem Repeats (STRs) located on chromosomes 19, 20, 21, and 22 were genotyped in 50 individuals from Carloforte selected on the basis of founder surnames. According to this sampling strategy, participants were descendants (from both paternal and maternal sides) of the first founders, and not related up to, at least, two generations. The 31 loci showed a high degree of heterozygosity (Robledo et al. 2009).

Some forensic parameters were calculated, such as the power of discrimination and power of exclusion. These parameters gave very high values (Robledo et al. 2009),

and did not show the decrease in the power of exclusion that is generally observed in small, isolated, villages (de Pancorbo et al. 2000). This surprising finding, together with the high level of heterozygosity shown by the STR markers is likely a consequence of the sampling strategy, which was designed to maximize the genetic differences among the selected participants.

***Y Chromosome markers***

The isolate of Carloforte was also studied through the analysis of 17 STRs located on Y chromosome. Male individuals from Carloforte (N=41) were selected with the same methodology described in the previous section. Y-haplogroups were inferred from the 17-locus haplotypes by haplogroup predictor (http://www.hprg.com/hapest5) (Athey 2005). Among the 17 loci (DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS437, DYS438, DYS439, DYS635, DYS385 a/b, DYS448, DYS456, DYS458, Y GATA H4), 13 loci showed a unimodal allele distribution, whereas DYS438, DYS392, DYS456 and DYS635 displayed a bimodal distribution (unpublished data). Using Arlequin v.3.5 (Excoffier and Lischer 2010) some intra-population parameters were calculated: gene diversity (d=0.975), mean number of pairwise differences ($\Theta_\pi$=9.650) and average gene diversity over all loci (equal to 0.603). The last two parameters appear very high, when compared with values from other European populations (Ehler et al. 2011).

Among the 39 different haplotypes found, five individuals, with the same surname, shared a haplotype that was identical for 16 loci, except DYS458. Clearly, their Y chromosomes derive from a shared ancestral Y chromosome, even though the five individuals are unrelated to at least the grandparental generation. It is interesting to note that their last name coincides with the family name of a couple that, among the

founders, had four sons (Vallebona 1988). Therefore, at the time of foundation, there were multiple copies of that Y chromosome. Moreover, the difference in locus DYS458 was one repeat unit. In particular, two individuals carry allele 17, two individuals carry allele 16, and the fifth individual carries allele 18. It is noteworthy that this STR has the reported highest mutation rate: μ=1.06 (http://www.cstl.nist.gov/biotech/strbase/str_y458.htm).

An inter-population analysis was carried out. The isolate of Carloforte appears significantly differentiated from the neighboring mainland area of Sulcis-Iglesiente ($F_{ST}$ = 0.02). In fact, it shares only one haplotype with Sulcis-Iglesiente. Haplogroup analysis also reveals an unusual distribution: the Carloforte population is characterized by a low frequency of haplogroup I2a1 (2%), which in Sulcis-Iglesiente has a frequency of 38%. This haplogroup is defined by marker M26, and it is very characteristic of Sardinia, and nearly absent in the rest of Italy (Onofri et al. 2007) and in Corsica (Francalacci et al. 2003). The frequency of R1b haplogroup in Carloforte reaches a value of 56%, similar to the value found in Northern Italy (Ferri et al. 2009; Semino et al. 2000), while in Sardinia it has a much lower frequency of 20% (Fig. 6). To analyze the phylogenetic relationship between Carloforte and other Italian populations, a neighbor-joining tree was built. Published data (Presciuttini et al. 2001; Di Gaetano et al. 2009) only allow the use of seven STRs: DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, in the comparison. The genetic tree confirms the peculiarity of Carloforte populations. Carloforte, despite being on a separate branch, is in the cluster of Italian populations (Fig. 7). However, the Ligurian population does not cluster near Carloforte. This result could be an effect of genetic drift, due to the small number of Carloforte founders; an alternative explanation is the presence of stratification in the Ligurian population.

When Tunisia (Khodjet El Khil et al. 2001) is added to the analysis, it is located in a separate branch, as an outlier, confirming that this population did not contribute to the Carloforte gene pool (data not shown). A second tree (Fig. 8) built with the use of 11 STRs, adding loci DYS437, DYS438, DYS439, and DYS635, compared Carloforte with other Mediterranean populations. Again, Carloforte clusters in a separate branch near to Spain and Northern Italy, while it is more differentiated from Sardinian populations (Turrina et al. 2006; Onofri et al. 2007; Pokupcic et al. 2008; Palet et al. 2010; Martin et al. 2004; Frigi et al. 2006; Kovatsi et al. 2009).

***mtDNA***

The same sampling (N=49), selected on the basis of founder surnames, has also been analyzed for mtDNA variation. After sequencing the hyper-variable region I (HVRI, nt 16,024 – 16,383) we obtained ten sequences that were identical to the cCRS revised Cambridge reference sequence, and 34 variable sites. We detected 167 mutations, which included 154 substitutions, (142 transitions (92.2%) and 12 transversions (7.8 %)), and 13 single base insertions with an addition of a cytosine following position 16,193.

An intra-population analysis, performed with Arlequin (v. 3.5), detected a level of gene diversity (d) equal to 1, a mean number of pairwise differences ($\Theta_\pi$) equal to 5.257, and an average gene diversity over all nucleotides equal to 0.015. Haplogroups were estimated with use of Phylotree (van Oven and Kayser 2009), and the following haplogroup frequencies were obtained: H (33%), U (19%), X (16%), T (16%), M (8%), J (6%) and K (2%).

A second sampling (N=50) of the Carloforte population was obtained with the grandparents' criterion: selected individuals were born and resident in Carloforte, and

unrelated for at least three generations (Falchi et al. 2006). Intra-population analysis of this sampling detected a level of gene diversity (d) equal to 0.978, a mean number of pairwise differences ($\Theta_\pi$) equal to 4.46, and an average gene diversity over all nucleotides equal to 0.012 (Calò et al. 2012). Haplogroup analysis gave the following frequencies: H (53%), J (14%), V (9%), HV (5%), U (5%), I (5%), K (5%), X (2%) and M (2%) (Fig. 9).

The difference between the two samplings of Carloforte in mtDNA haplogroups, was highly significant ($P<0.0001$). The major difference detected was the absence of haplogroups HV, V and I in the sample based on founder surnames, and the absence in the sample based on birthplace of haplogroup T.

We compared HVRI sequences of the two different Carloforte samplings with those from other Sardinian and Mediterranean populations (Fadhlaoui-Zid et al. 2004 Falchi et al. 2006; Francalacci et al.1996; Fraumene et al. 2006; Plaza et al. 2003; Varesi et al. 2000; Vona et al. 2001). We constructed a genetic tree in which the two different Carloforte samplings were located very distant from each another (Fig. 10). The sampling collected using the grandparents' criterion clustered with the Italian population and is quite close to other Sardinian communities; however, the sampling based on founder surnames is located on a separate, and distant, branch. We conclude that the contrasting results reflect a different time frame in the population history: sampling for founder surnames describes the early events of the population, while sampling using the grandparents' criterion reflects the influence of migrations or recent gene flow, and therefore provides information on the more recent history (Calò et al. 2012).

**Discussion**

In the present paper we reviewed data obtained from different studies concerning the population isolate of Carloforte, and analyzed new data on Y-chromosome markers. Exogamy and isonymy data (Figs. 2 and 3) show a high degree of isolation that lasted for the first 100 years following the foundation of Carloforte, in spite of the geographic proximity to Sardinia. We believe that the population's unique, troubled but successful, history elicited a strong feeling of self identity among the Carloforte people that favoured endogamous marriages. The peculiar local language has been an additional cultural factor in the differentiation of Carloforte from Sardinians. It is notable that Tabarchino is still spoken today by educated people and even by the younger generation, whereas in the rest of Italy local language varieties are generally spoken only by elderly, or by the less educated people. Those cultural factors favoured assortative matings that, in turn, were the main contributing factor to the differentiation of Carloforte from Sardinians observed first with classical markers and afterwards confirmed with autosomal, mtDNA and Y-chromosome markers. On the contrary, an extensive gene flow occurred between the descendants of Nueva Tabarca founders and the local Spaniard community, which led to the disappearance of Tabarchino. Indeed, the only remnant of the founder group is represented by the surnames that survived, even though sometimes slightly modified: for example the original surnames Borghero and Luxoro became Burguero and Luchoro, respectively (Ferraro 1989).

Data from autosomal markers, Y chromosome loci, and mtDNA sequencing are concordant in describing a high level of heterogeneity within Carloforte (Fig. 6 and 9). We believe that this result is due to the sampling strategy based on founder surnames: by selecting proven descendants of the village founders, but with no ancestors in common to, at least, the grandparental generation, we intended to capture

the maximum differences. A similar conclusion was reached in a different study on Y chromosome lineages in population isolates located in the Italian Western Alps (Boattini et al. 2010). Moreover, the high level of gene and haplotype diversity indicates a negligible, if any, effect of genetic drift in shaping Carloforte's population (Fraumene et al. 2003; Casas Vargas et al. 2011), consistent with the young history of the population, and the absence of a bottleneck effect (Robledo et al. 2009), also in agreement with historical records.

mtDNA analysis emphasizes the crucial aspect of sampling strategy. Comparing the sampling through founder surnames with the sampling through the grandparents' criterion, significant differences in haplogroup distribution in Carloforte population were evident (Calò et al. 2012). Moreover, the analysis of HVRI sequences of the two different Carloforte samplings and those of other Sardinian and Mediterranean populations produced a genetic tree in which the two Carloforte samplings were located very distant from one another (Fig 10), leading to contrasting results and different conclusions. We believe that the discrepancy reflects the different time frame in the population analysis: founder surnames sampling is not affected by recent gene flow and is therefore a signature of the ancestral population, whereas the grandparents' criterion is a signature of the present population, shaped by migration and gene flow, mainly from nearby Sardinia. Indeed, only the sampling through the grandparents' criterion locates Carloforte next to other Sardinian populations, in agreement with the biodemographic data, which indicates an increase in exogamous marriages with Sardinians. It would be interesting to collect Carloforte male individuals through the grandparents' criterion and see if the same pattern will be observed also analyzing Y-chromosome markers.

We believe that, in studying a population's dynamics, both sampling strategies through the grandparents' criterion and founder surnames, whenever possible, should be employed, to correctly infer past and recent events. This is especially important in association studies aimed to identify genes involved in complex phenotypes: lack of knowledge about the current genetic structure of the populations studied are likely to produce false negative or false positive associations, which may explain why several reports on loci associated with complex phenotypes turned out to be not reproducible.

In conclusion, genetic isolates may represent a powerful tool in anthropological studies, population genetics, and in mapping genes involved in Mendelian or complex disorders, provided that detailed information is available. However, genetically isolated populations differ from each other in several parameters, like number of founders, population age, bottlenecks, and migration. Only a multidisciplinary approach allows the knowledge of all those parameters, which is essential to understand the present genetic structure of the population.

An ideal scenario is offered by a genetic isolate with a recent, and well documented, history, like Carloforte, where, in addition, the availability of official records allows the construction of very large, multi-generation, pedigrees that may help in identifying shared ancestral alleles or shared haplotypes.

**Acknowledgements**

The authors are grateful to John Mitchell for helpful suggestions and for revising the manuscript. We also wish to thank Paolo Francalacci and Antonietta Marra for discussions and criticisms.

This work was supported by grant from the Italian Ministry of Research, MIUR (PRIN 2007 2007TYXE3X_003, and 2009 200975T9EW_002).

**Literature Cited**

Arcos-Burgos M., and M. Muenke. 2002. Genetics of population isolates. *Clin. Genet.* 61:233–247.

Athey T., W. 2005. Haplogroup prediction using an allele frequency approach. *J. Genet. Geneal.* 1:1-7.

Blasco Ferrer, E. 1994. Contributo alla conoscenza del ligure insulare. Il tabarchino di Sardegna. *Zeitschrift für romanische Philologie (ZrP)*. 110:153–194.

Boattini, A., M.E. Pedrosi, D. Luiselli, and D. Pettener. 2010. Dissecting a human isolate: Novel sampling criteria for analysis of the genetic structure of the Val di Scalve (Italian Pre-Alps). *Ann. Hum. Biol.* 37:604–609.

Boattini, A., C. Griso, and D. Pettener. 2011. Are ethnic minorities synonymous for genetic isolates? Comparing Walser and Romance populations in the Upper Lys Valley (Western Alps). *J. Anthropol. Sci.* 89:161-173.

Bottiglioni, G. 1928. L'antico Genovese e le isole linguistiche sardo-corse. *L'Italia dialettale*. 4:1–78.

Calò, C.M., and G. Vona. 1994. Carloforte: evoluzione della struttura matrimoniale. *Rendiconti del Seminario della Facoltà di Scienze dell'Università degli Studi di Cagliari*, 64: 73-87.

Calò, C., and G. Vona. 2005. Genetic and demographical analysis of San Pietro Island (Sardinia, Italy). *Hum. Evol*. 20:99–106.

Calò, C.M., L. Corrias, G. Vona et al. 2012. Sampling strategies in a linguistic isolate: results from mtDNA analysis. *Am. J. Hum. Biol.* 24:192–194.

Casas-Vargas, A., A. Gómez, I. Briceño et al. 2011. High genetic diversity on a sample of pre-Columbian bone remains from Guane territories in northwestern Colombia. *Am. J. Phys. Anthropol*. 146:637–649.

Cavalli Sforza, L., P. Menozzi, and A. Piazza. 1994. The history and geography of human genes. Princeton University Press, Princeton.

Colonna, V., T. Nutile, M. Astore, et al. 2007. Campora: A young genetic isolate in South Italy. *Hum. Hered*. 64: 123-135.

de Pancorbo, M., N. Cuevas, M. Bujan et al. 2000. Micro-satellite DNA and forensic parameters in isolated populations. *Progr. Forensic. Genet*. 8:181–183.

Destro-Bisol G., P. Anagnostou, C. Batini et al. 2008. Italian isolates today: geographic and linguistic factors shaping human biodiversity. *J. Anthropol. Sci.* 86:179–188

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

Ehler, E., D. Vaněk, V. Stenzl, and V. Vančata. 2011. Y-chromosomal diversity of the Valachs from the Czech Republic: model for isolated population in Central Europe. *Croat Med J*. 52:358–367.

Excoffier, L., and H.E.L. Lischer. 2010. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol*

*Ecol Resour*. 10:564–567.

Fadhlaoui-Zid, K., S. Plaza, F. Calafell et al. 2004. Mitochondrial DNA heterogeneity in Tunisian Berbers. *Ann. Hum. Genet.* 68:222–233.

Falchi, A., L. Giovannoni, C.M. Calo et al. 2006. Genetic history of some western Mediterranean human isolates through mtDNA HVR1 polymorphisms. *J. Hum. Genet.* 51:9–14.

Felsenstein, J. 1989. PHYLIP -- Phylogeny Inference Package (Version 3.2). *Cladistics.* 5: 164-166.

Ferraro, G. 1989. Da Tabarka a San Pietro. Nasce Carloforte. *Tipografia Artigiana Musanti Editrice*, Cagliari.

Ferri, G., M. Alù, B. Corradini et al. 2009. Slow and fast evolving markers typing in Modena males (North Italy). *Forensic Sci. Int. Genet*. 3:e31–33.

Francalacci, P., J. Bertranpetit, F. Calafell, and P.A. Underhill. 1996. Sequence diversity of the control region of mitochondrial DNA in Tuscany and its implications for the peopling of Europe. *Am. J. Phys. Anthropol.* 100:443–460.

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

Fraumene, C., E. Petretto, A. Angius, and M. Pirastu. 2003. Striking differentiation of sub-populations within a genetically homogeneous isolate (Ogliastra) in Sardinia as revealed by mtDNA analysis. *Hum. Genet.* 114:1–10.

Fraumene, C., E.M.S. Belle, L. Castrì et al. 2006. High resolution analysis and

phylogenetic network construction using complete mtDNA sequences in sardinian genetic isolates. *Mol. Biol. Evol.* 23:2101–2111.

Frigi, S., F. Pereira, L. Pereira et al. 2006. Data for Y-chromosome haplotypes defined by 17 STRs (AmpFLSTR Yfiler) in two Tunisian Berber communities. *Forensic Sci. Int.* 160:80–83.

Gagnon, A., and B. Toupance. 2002. Testing isonymy with paternal and maternal lineages in the early Québec population: the impact of polyphyletism and demographic differentials. *Am. J. Phys. Anthropol.* 117:334–341.

Khodjet El Khil, H., R. Triki Marrakchi, B. Yacoubi Loueslati et al. 2001. Y chromosome microsatellite variation in three populations of Jerba Island (Tunisia). *Ann. Hum. Genet.* 65:263–270.

Kovatsi, L., J.L. Saunier, and J.A. Irwin. 2009. Population genetics of Y-chromosome STRs in a population of Northern Greeks. *Forensic Sci Int Genet*. 4:e21–22.

Kristiansson, K., J. Naukkarinen, and L. Peltonen. 2008. Isolated populations and complex disease gene identification. *Genome Biol.* 9:109.

Latini, V., C.M. Calò, M.E. Ghiani et al. 2004. Carloforte, comunità alloglotta: struttura genetica attraverso l'analisi dei cognomi e dei marcatori genetici. *Studiostampa*, Nuoro. 66–74.

Marchani E., W. Watkins, K. Bulayeva et al. 2008. Cultures create genetic structure in the Caucasus: Autosomal, mitochondrial, and Y-chromosomal variation in Daghestan. *BMC Genet.* 9:47

Martín, P., J. García-Hirschfeld, O. García et al. 2004. A Spanish population study of

17 Y-chromosome STR loci. *Forensic Sci. Int.* 139:231–235.

Morelli, L., and G. Vona. 1993. Studio genetico della popolazione di Villanova Monteleone (Sardegna). Nota I. Struttura matrimoniale. *Rendiconti del Seminario della Facoltà di Scienze dell'Università degli Studi di Cagliari*, 63: 75–86.

Moroni, A., A. Anelli, W. Anghinetti et al. 1972. La consanguineità umana nell'isola di Sardegna dal XVIII al XX secolo. *Ateneo Parmense*. 8:69–92.

Neufeld E., H. Mandel, T. Raz et al. 1997. Localization of the gene for thiamine-responsive megaloblastic anemia syndrome, on the long arm of chromosome 1, by homozygosity mapping. *Am. J. Hum. Genet.* 61: 1335–1341.

Nikali K., A. Suomalainen, J. Terwilliger et al. 1995. Random search for shared chromosomal regions in four affected individuals: the assignment of a new hereditary ataxia locus. *Am. J. Hum. Genet.* 56: 1088–1095.

Onofri, V., F. Alessandrini, C. Turchi et al. 2007. Y-chromosome genetic structure in sub-Apennine populations of Central Italy by SNP and STR analysis. *Int. J. Legal Med.* 121:234–237.

Oota H., W. Settheetham-Ishida, D. Tiwawech et al. 2001. Human mtDNA and Y-chromosome variation is correlated with matrilocal versus patrilocal residence. *Nat. Genet.* 29: 20–21.

Palet, L., C. Coudray, C. Galey et al. 2010. Y-STR genetic diversity in Moroccans from the Figuig oasis. *Forensic Sci. Int. Genet*. 4:e139–141.

Pardo, L.M., I. MacKay, B. Oostra et al. 2005. The effect of genetic drift in a young genetically isolated population. *Ann. Hum. Genet.* 69:288–295.

Plaza, S., F. Calafell, A. Helal et al. 2003. Joining the pillars of Hercules: mtDNA sequences show multidirectional gene flow in the western Mediterranean. *Ann. Hum. Genet.* 67:312–328.

Pokupcić, K., S. Cukrov, I.M. Klarić et al. 2008. Y-STR genetic diversity of Croatian (Bayash) Roma. *Forensic Sci. Int. Genet*. 2:e11–13.

Pope, A.M., S.M. Carr, K.N. Smith, and H.D. Marshall. 2011. Mitogenomic and microsatellite variation in descendants of the founder population of Newfoundland: high genetic diversity in an historically isolated population. *Genome*. 54:110–119.

Presciuttini, S., A. Caglià, M. Alù et al. 2001. Y-chromosome haplotypes in Italy: the GEFI collaborative database. *Forensic Sci. Int.* 122:184–188.

Robledo, R., I. Piras, W. Beggs, and C. Calò. 2009. Analysis of 31 STR loci in the genetic isolate of Carloforte (Sardinia, Italy). *Genet. Mol. Biol.* 32:462–465.

Sanna, E., A. Melis, and M.C. Iovine. 2006. Endogamy and exogamy in an alloglot linguistic community of Sardinia (Carloforte) from 1738 to 2001. In Human Genetic Isolates. Carla Calò and Giuseppe Vona. *Research Signpost*, India, 121–130.

Semino, O., G. Passarino, P.J. Oefner et al. 2000. The genetic legacy of Paleolithic Homo sapiens sapiens in extant Europeans: a Y chromosome perspective. *Science*. 290:1155–1159.

Sitzia, P. 1998. Le comunità tabarchine della Sardegna meridionale: un'indagine sociolinguistica. *Condaghes*, Cagliari.

Slatkin, M. 1995. A measure of population subdivision based on microsatellite allele frequencies. *Genetics.* 139: 457-462.

Toso, F. 2003. I tabarchini della Sardegna. *Le mani*, Genova.

Turrina, S., R. Atzei, and D. De Leo. 2006. Y-chromosomal STR haplotypes in a Northeast Italian population sample using 17plex loci PCR assay. *Int. J. Legal Med.* 120:56–59.

Vallebona, G. 1988. Carloforte. Storia di una colonizzazione. *Edizioni della Torre*, Cagliari.

van Oven, M., and M. Kayser. 2009. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Hum. Mutat.* 30:E386–394.

Varesi, L., M. Memmí, M.-C. Cristofari et al. 2000. Mitochondrial control-region sequence variation in the Corsican population, France. *Am. J. Hum. Biol.* 12:339–351.

Vona, G., C.M. Calò, G. Lucia et al. 1996. Genetics, geography, and culture: the population of S. Pietro Island (Sardinia, Italy). *Am. J. Phys. Anthropol.* 100:461–471.

Vona, G., M.E. Ghiani, C.M. Calò et al. 2001. Mitochondrial DNA sequence analysis in Sicily. *Am. J. Hum. Biol.* 13:576–589.

Table 1. Allele frequencies of classical genetic markers in the Carloforte and Sardinian population (modified from Vona et al. 1996) and heterogeneity G-test with Sardinian mean value: *p<0.05, **p<0.01.

| Markers | alleles | Carloforte | Sardinia | | |
|---|---|---|---|---|---|
| | | | mean | min | max |
| RH** | CDE | 0.015 | 0.008 | 0.004 | 0.016 |
| | CDe | 0.514 | 0.696 | 0.683 | 0.709 |
| | cDE | 0.102 | 0.106 | 0.083 | 0.114 |
| | cDe | 0.083 | 0.069 | 0.023 | 0.187 |
| | Cde | 0.000 | 0.001 | 0.000 | 0.003 |
| | cdE | 0.001 | 0.000 | 0.000 | 0.000 |
| | cde | 0.285 | 0.119 | 0.000 | 0.174 |
| MN | M | 0.589 | 0.676 | 0.651 | 0.697 |
| | N | 0.411 | 0.324 | 0.303 | 0.329 |
| ESD* | 1 | 0.797 | 0.882 | 0.847 | 0.914 |
| | 2 | 0.203 | 0.118 | 0.092 | 0.153 |
| PGM1 | 1S | 0.640 | 0.671 | 0.655 | 0.689 |
| | 1F | 0.079 | 0.090 | 0.080 | 0.094 |
| | 2S | 0.243 | 0.206 | 0.173 | 0.233 |
| | 2F | 0.037 | 0.034 | 0.025 | 0.049 |
| GC** | 1S | 0.536 | 0.702 | 0.680 | 0.739 |
| | 1F | 0.145 | 0.053 | 0.009 | 0.112 |
| | 2 | 0.349 | 0.245 | 0.164 | 0.266 |
| HP | 1 | 0.436 | 0.382 | 0.348 | 0.415 |
| | 2 | 0.564 | 0.618 | 0.585 | 0.652 |

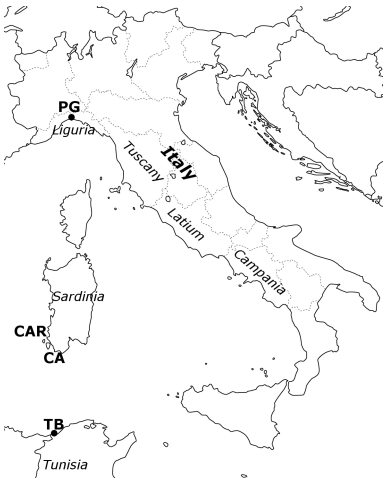


Figure 1. Geographical location of Carloforte (Cf) and Calasetta (Ca) in Sardinia. Locations of the Ligurian town of Pegli (Pg) and Tabarka (Tb) are also indicated.

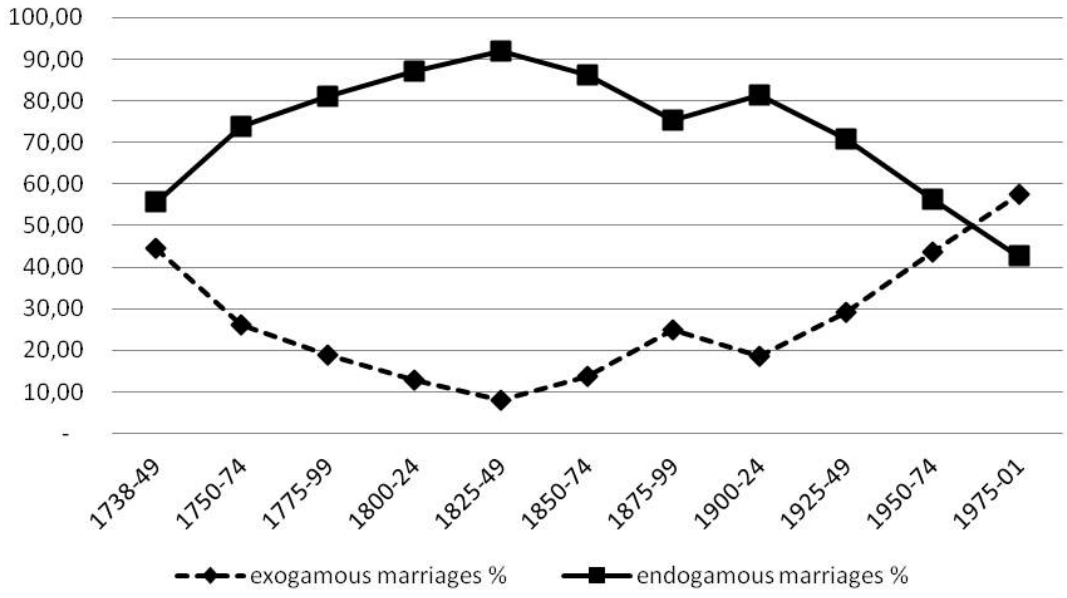


Figure 2. Frequencies of exogamous and endogamous marriages recorded in Carloforte population between 1738 and 2001 (modified from Vona et al. 1996).

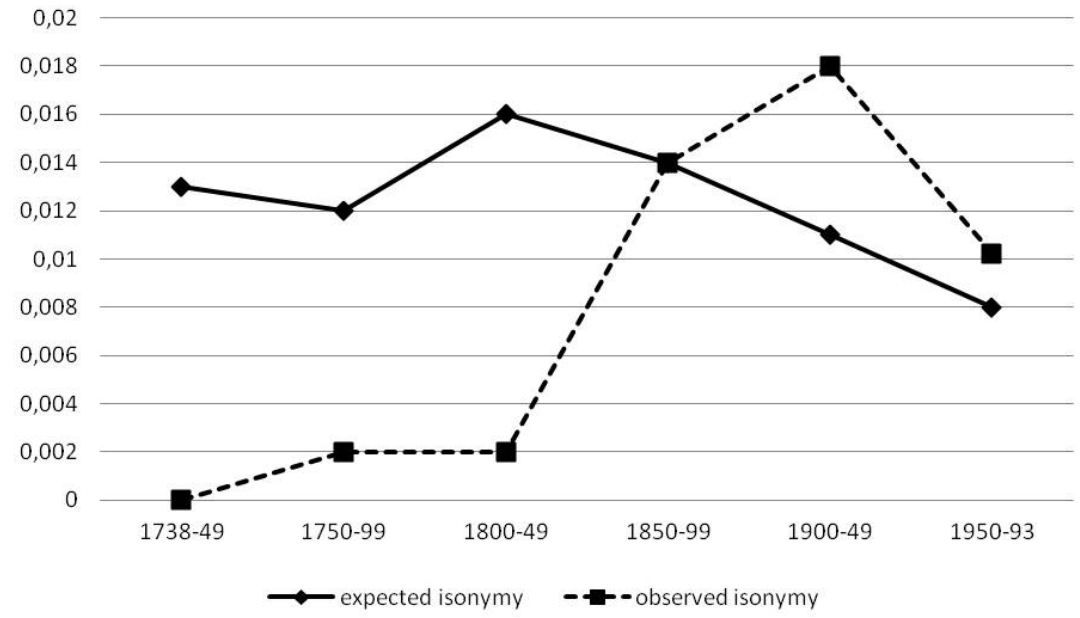


Figure 3. Expected and observed isonymy in Carloforte population between 1738 and 1953 (from Latini et al. 2004, modified).

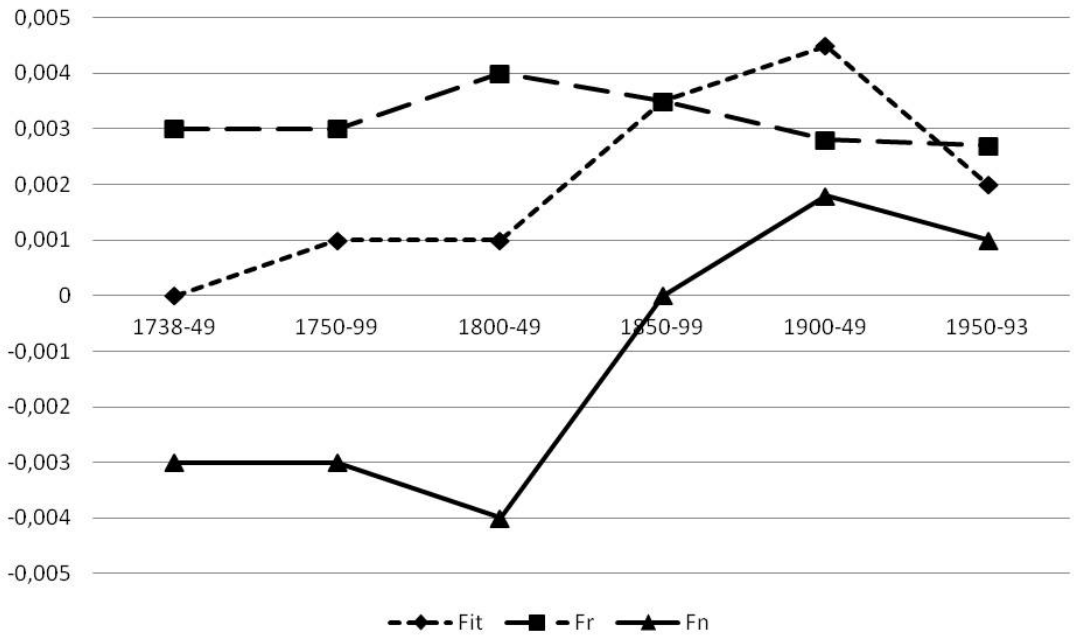


Fig. 4. Temporal variation of total inbreeding coefficient (Fit), with non random (Fn) and random component (Fr), in Carloforte population between 1738 and 1953.

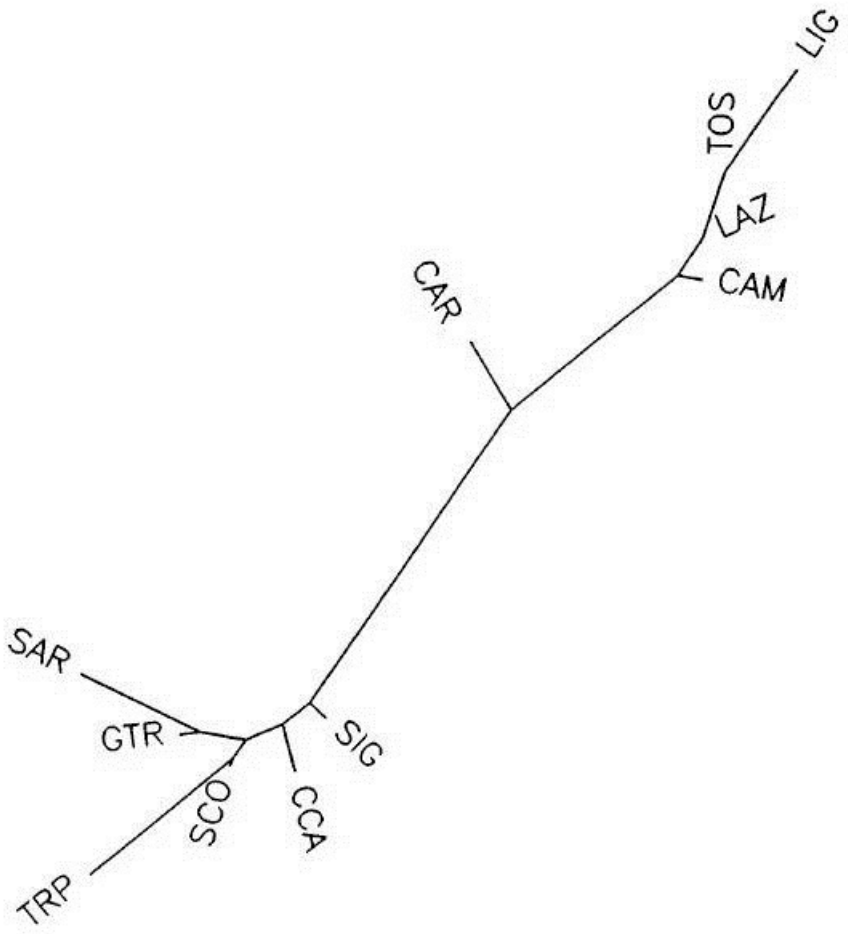


Figure 5. Neighbor-joining tree obtained from the matrix of Reynold's genetic distances. CAM, Campania; CAR, Carloforte; CCA, Campidano Cagliari; GTR, Gerrei-Trexenta; LAZ, Latium; LIG, Liguria; SAR, Sarrabus; SCO, South Campidano Oristano; SIG, Sulcis-Iglesiente; TOS, Tuscany; TRP, Trexenta-Parteolla (figure from Vona et al., 1996).

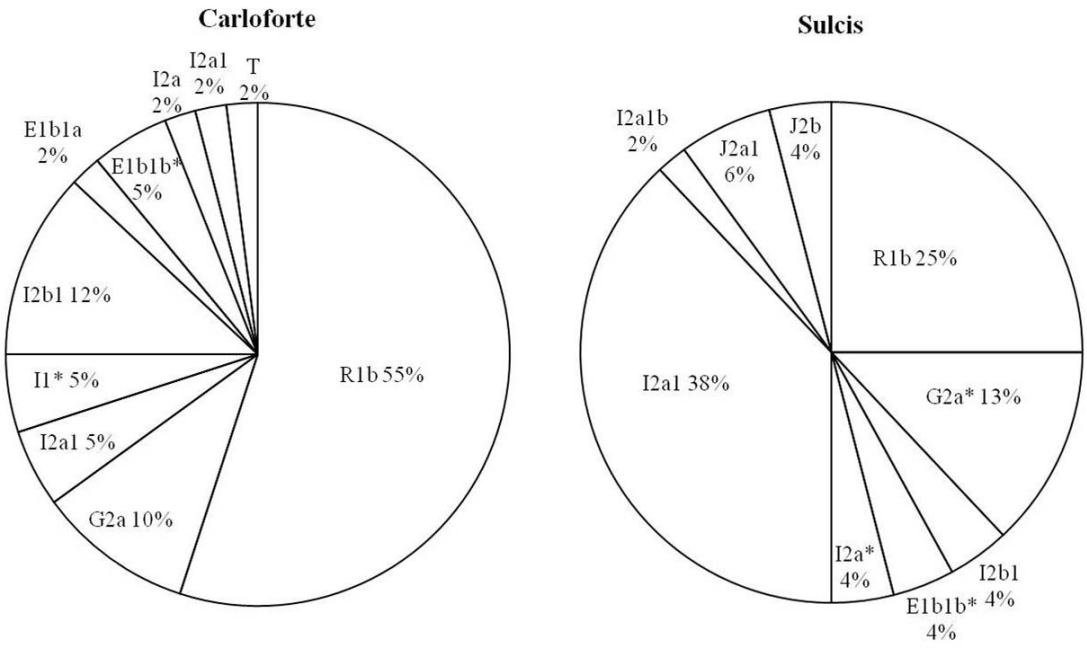


Figure 6. Y- chromosome haplogroup frequencies in Carloforte and Sulcis populations.

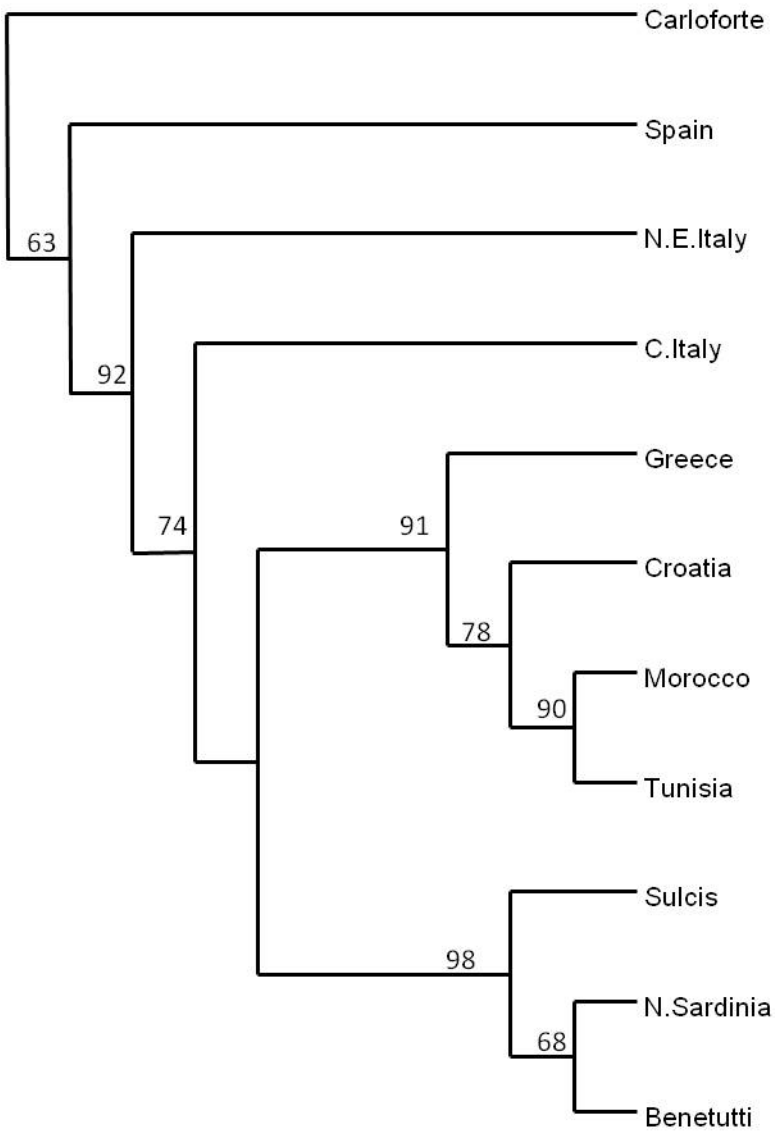


Figure 7. Neighbor-joining tree of Carloforte, Italian and Mediterranean populations (7 Y-STRs), bootstrap values are shown on the branches.

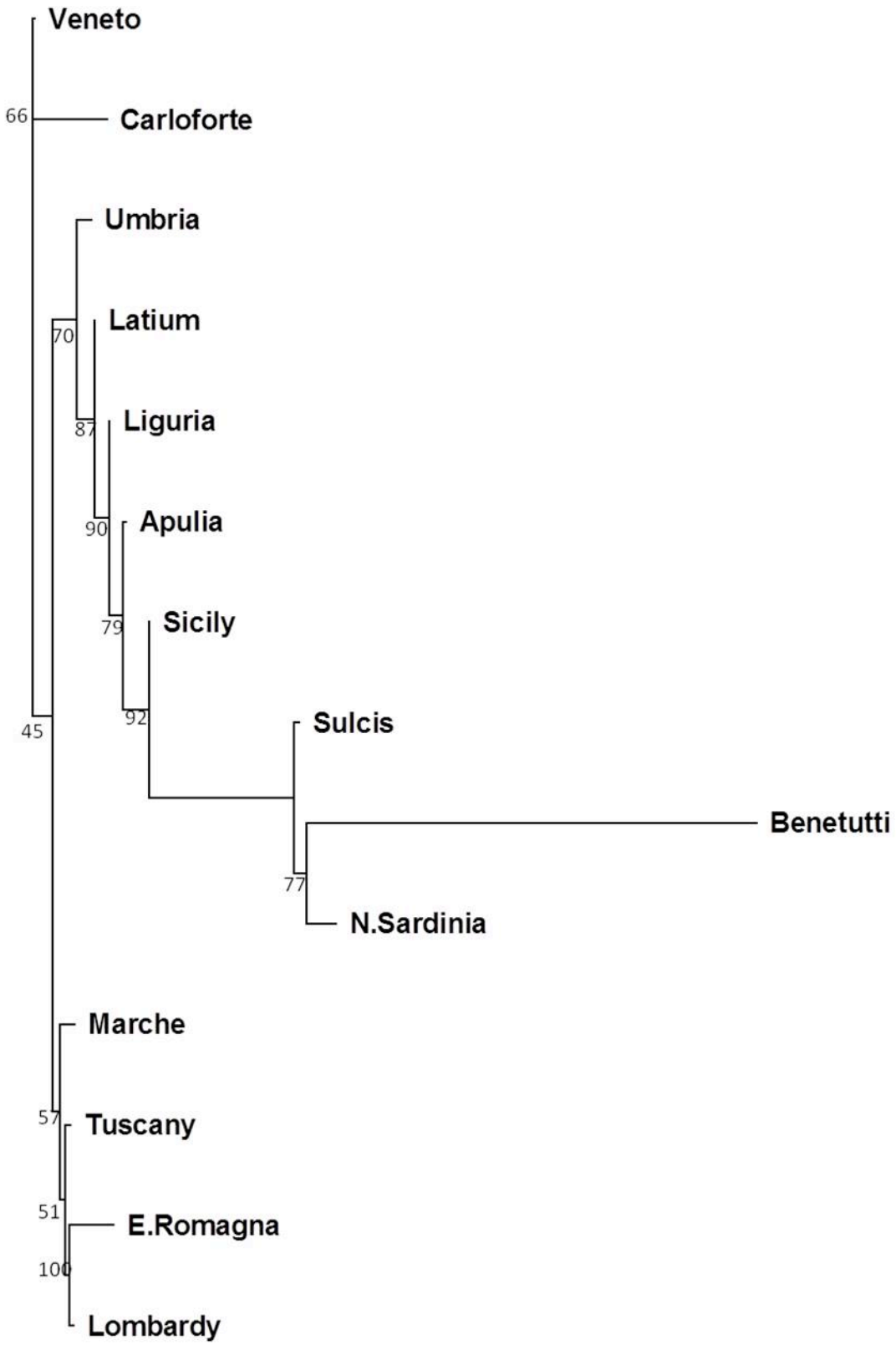


Figure 8. Neighbor-joining tree of the Carloforte sample and Italian populations (11 Y-STRs), bootstrap values are shown on the branches.

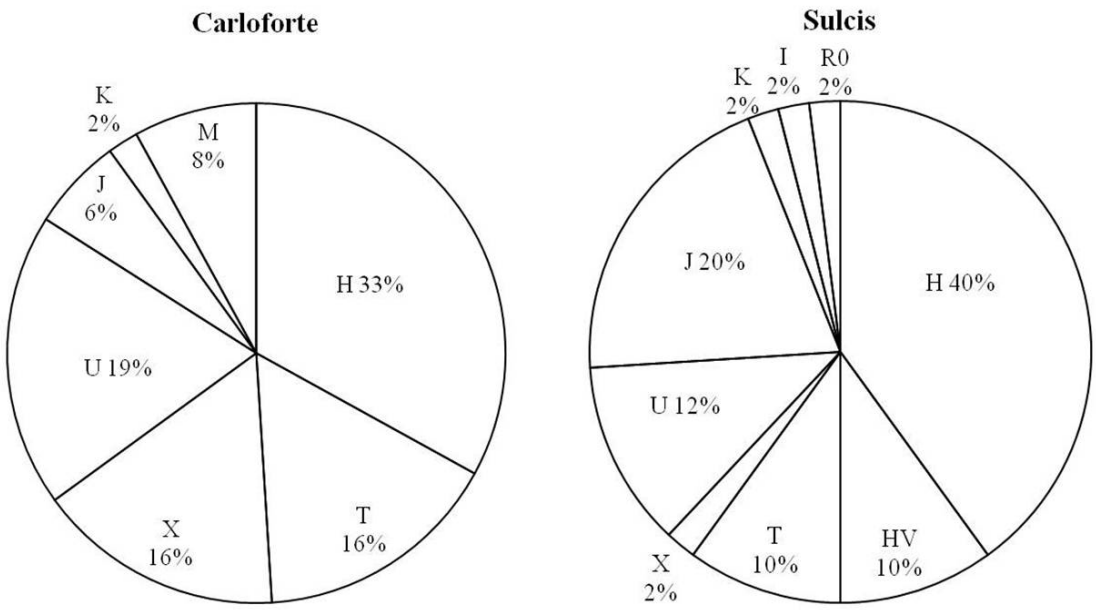


Figure 9. mtDNA haplogroup frequencies in the Carloforte and Sulcis populations.

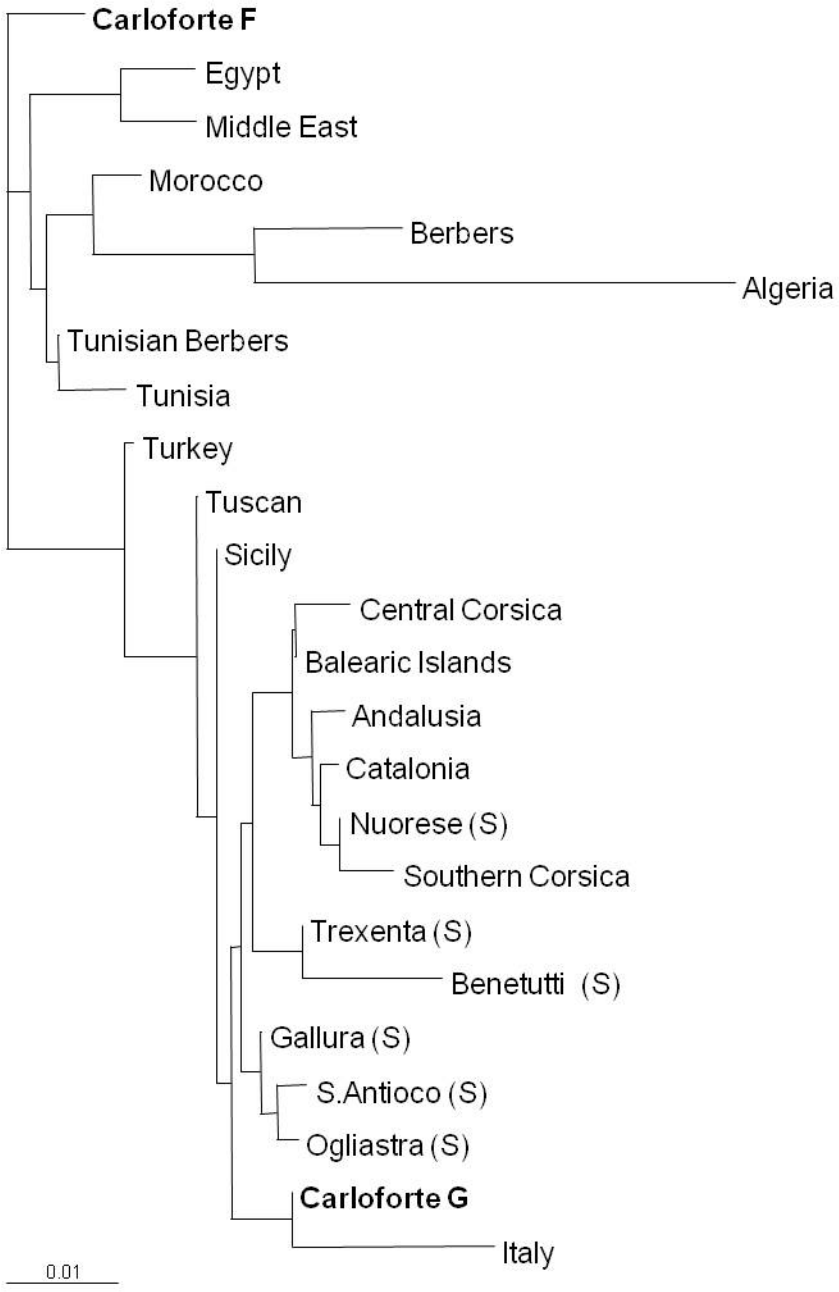


Figure 10. mtDNA-Neighbor-joining tree of the two Carloforte samples, Sardinian (S) and Mediterranean populations (figure from Calò et al., 2012).