Evaluating mtDNA patterns of genetic isolation using a resampling procedure: a case study on Italian populations.

Paolo Anagnostou[1,2*], Marco Capocasa[3,2], Valentina Dominici[1], Francesco Montinaro[4], Valentina Coia[5] and Giovanni Destro-Bisol[1,2]

[1]Dipartimento di Biologia Ambientale, Sapienza University of Rome, 00185, Rome, Italy

[2]Istituto Italiano di Antropologia, 00185, Rome, Italy

[3]Dipartimento di Biologia e Biotecnologie "Charles Darwin", Sapienza University of Rome, 00185, Rome, Italy

[4]Department of Zoology, University of Oxford,South Parks Road,OX1 3PS, Oxford, UK

[5]Istituto per le Mummie e l'Iceman, Accademia Europea di Bolzano (EURAC-Research), 39100, Bolzano, Italy

*** Corresponding Author:** Paolo Anagnostou – paolo.anagnostou@uniroma1.it

Keywords: Human isolates, population genetics statistics, mtDNA, European populations

ABSTRACT

Background: A number of studies which have investigated isolation patterns in human populations rely on the analysis of intra- and inter-population genetic statistics of mtDNA polymorphisms. However, this approach makes it difficult to differentiate between the effects of long–term genetic isolation and the random fluctuations of statistics due to reduced sample size.

Aim: Overcoming the confounding effect of sample size when detecting signatures of genetic isolation.

Subjects and methods: A resampling based procedure was employed to evaluate reduction in intra-population diversity, departure from surrounding genetic background and demographic stationarity in 34 Italian populations subject to isolation factors.

Results: We detected signatures of genetic isolation for all three statistics in seven populations: Pusteria valley, Sappada, Sauris, Timau settled in the eastern Italian Alps and Cappadocia, Filettino and Vallepietra settled in the Appenines. On the other hand, we were unable to find signals for any of the statistics analysed in 19 populations. Finally, we found eight populations showing signals of isolation for two or one statistic.

Conclusion: Our analysis revealed that the use of population genetic statistics combined with resampling procedure can help detect signatures of genetic isolation in human populations even using a single, although highly informative, locus like mtDNA.

Introduction

Human population isolates are of particular interest for population geneticists for two reasons. Firstly, the investigation of such groups of individuals makes it possible to gain a more detailed picture of the genetic structure and spatial patterns of genetic diversity of human populations (Lauc, 2003; Brandstätter et al., 2007; Boattini et al., 2011; Esko et al., 2013; Capocasa et al., 2014). Secondly, genetic isolation in humans may be associated with differences in language, religion and ethnicity (e.g. see Rosenberg et al., 2002; Bosch et al., 2006; Capocasa et al., 2013), which gives the opportunity to study the influence of cultural factors on the genetic structure of populations (Laland et al., 2010). In addition, their high homogeneity and therefore high Linkage disequilibrium make them ideal initial candidates for Genome Wide Association studies.

However, detecting signatures of genetic isolation in human populations is a challenging task. In fact, the recent origin of *Homo sapiens* and its high level of mobility and admixture have limited the overall impact of genetic isolation on genetic structure (Tishkoff and Kidd, 2004). Consequently, the effects of long-term isolation on population allele frequencies may be subtle, and hence, difficult to detect.

A number of studies have used maternally inherited polymorphisms of mitochondrial DNA (mtDNA) to search for signatures of isolation in the genetic make-up of human populations subject to cultural and/or geographic isolation factors (e.g. Tolk et al., 2001; Bosch et al., 2006; Messina et al., 2010; Cilli et al., 2011; van Oven et al., 2011). These investigations have compared within- and among-group diversity between candidate population isolates and large neighbouring outbred groups, with only a couple of exception (Brandstätter et al., 2007, Capocasa et al., 2013). More in particular, they employed specific population genetic statistics to look for lower intra-population diversity levels,

departures from the surrounding genetic background and lack of signatures of demographic expansion (mainly by means of Haplotype Diversity, Fixation index and Fu'FS, respectively).

However, two problems arise with this approach. Firstly, the relatively small census size and the high level of consanguinity of many isolated groups often reduce the availability of unrelated individuals (in the majority of these studies defined according to the grandparent rule), which limits the accuracy of population diversity estimates (Nei, 1987), especially when a single, although hypervariable, locus is employed. Secondly, there are no reference values which can help distinguish signatures of isolation effects from random fluctuations of population genetic statistics due to sample size variation.

In an attempt to overcome the two limitations mentioned above, we have applied a resampling based analysis (Fernandes et al., 2003, Beja-Pereira et al., 2006, Veeramah et al., 2011, Coia et al., 2012). Through this approach, we simulate a condition of small sample size in large outbred populations, and test the null hypothesis: "values of genetic isolation statistics observed in populations subject to geographic and/or cultural barriers fall within the range of values for broad and non-isolated groups with an equal sample size". When this hypothesis is rejected, a more robust evidence for genetic isolation is obtained. We decided to carry out our study on Italian populations, since their marked ethno-cultural heterogeneity makes them an optimal case study for the investigation of genetic isolation (see Destro Bisol et al., 2008). In fact, there are twelve linguistic minorities formally recognized by the Italian constitution, which represent 5% of the population (Toso, 2008, 2014). Furthermore, due to the presence of two main mountain chains, Alps and Apennines, and several islands, the territory may provide geographic conditions for further population isolation.

The comparison between the observed values and unbiased expectations for non-isolated populations obtained by resampling made it possible to evaluate signatures of genetic isolation which are discussed in the light of the historical and demographic background of the populations under study.

METHODS

In order to test the above hypothesis, we built a large mtDNA dataset of populations subject to cultural and/or geographical isolation factors, all settled in Italy. Finally, we collected data relative to 2632 individuals belonging to 57 Italian populations obtained from current literature and open online databases (Congiu et al., 2012). Of these, 34 populations (1692 individuals) are subject to geographic and/or cultural isolation factors (see Table I for the list of populations), whereas the other 23 (940 individuals) are large outbred populations. We classified all these populations in four groups according to their geographical location (North-east, Central and South Italy and Sardinia) (see Table I and Supplementary Table S1 for the list of outbred populations).

In order to avoid biases related to different sampling strategies, we selected only populations whose individuals were sampled with the standard "grandparents" criterion. Regarding the outbred populations, we selected only those whose individuals had all four grandparents born in a certain Italian province or a restricted geographical area. Furthermore, we considered only data regarding mitochondrial DNA hypervariable region 1 (mtDNA HVR-1; 16033-16365 np) in order to maximise the total number of human isolates analysed.

In order to build a set of expectations for some population genetic diversity statistics we proceeded as follows. At first, we evaluated the genetic homogeneity within Italian geographic regions by means of AMOVA analysis. For all four geographic regions we found low and statistically not significant among population variation values (FST) (see Supplementary Table 2 for AMOVA results)

which allowed us to pool the genetic data of these populations to assemble our reference populations. Thereafter, from each reference population we extracted 10,000 random sub-samples with no replacement, using an *ad hoc* script. The whole process exploits R and BASH environment scripts; the former creates the desired number of samples of N randomly sampled sequences, using the functions “replicate” and “sample”, respectively. The latter then converts the resulting tab-delimited output into Arlequin input files, which are loaded into the software using the “batch mode” option. We extracted samples with N={5...100} with incremental increases of 5 up to 50 and of 10 up to 100. For each sub-sample, we computed Haplotype Diversity (HD) and average genetic distances (A-Fst) (Reynolds et al., 1983). For each geographic region, this latter statistic was calculated against the neighbouring populations (North-east Italy, South Italy and Sardinia vs Central Italy; Central Italy vs South Italy).

The following procedure was repeated for each geographic region. In order to define the threshold for rejecting the null hypothesis, for the HD and A-Fst statistics we calculated the 95% confidence interval for each - distribution. We then used these values in a regression analysis in order to understand how they change as the sample size increases and identify the best regression function. In order to evaluate whether the null hypothesis could be accepted or not, for each candidate population isolate, we employed the standard deviations of HD and A-Fst – calculated as for the reference population - to compute standard scores (Z-scores) using the formula:

$$z - score = \frac{x_N - x}{\sigma}$$

where X_N is the value of the threshold at a given sample size, x is the observed value for each population isolate and σ is the latter’s standard deviation. We employed the z-scores to compute the probability, set at a 66.7% level, to reject the null hypothesis. To verify signals of demographic stationarity, we calculated the Fu’s FS (FS) (Fu, 1997). We chose this statistic because, as highlighted

by simulation analyses, differently from both Tajima's D and mismatch distribution, it provides higher statistical power with low sample sizes and it performs equally well with both population growth and bottlenecks (Ramos-Onsins and Rozas, 2002; Ramírez-Soriano et al., 2008). Therefore, we applied the same resampling method to Fu's FS in order to evaluate its robustness regarding variations of sample sizes. All statistics of intra- and inter-population genetic diversity were calculated using the Arlequin software (version 3.5.1.2, Excoffier and Lischer, 2010) whereas the regression analysis were performed with SPSS v. 19.0 (IBM Corp. Released 2010. IBM SPSS Statistics for Windows, Version 19.0. Armonk, NY: IBM Corp).

RESULTS AND DISCUSSION

There are two ways of looking at the results obtained with our resampling approach which may help us distinguish between likely signatures of genetic isolation and results biased by small sample size: (i) comparing the distribution of each single genetic statistic separately (HD, Fst, Fu's FS) using empirical and simulated values: (ii) making inferences regarding genetic isolation in each candidate population taking into account all three statistics together.

Test of the null hypothesis for genetic diversity statistics

Our first concern has been to make sure that the procedure used to build our reference neutral populations did not produce values of genetic diversity statistics that were substantially different from what observed across populations. Therefore, we computed several intra-population and demographic statistics as well as genetic distances for all four reference populations and the relative Italian outbred groups. The results obtained from these analyses suggest that each reference population provide a good representation of the genetic diversity of each of the Italian geographic regions. In fact, we do not observe any significant difference between the values of the genetic diversity statistics tested in the reference population and the distributions of the same statistics in

the open Italian population of the same region (see supplementary Table S1). As regard genetic distances, all the Fst values obtained comparing the reference populations with the open groups of the same region were not statistically significant. The above evaluation reassured us about the validity of our approach and allowed us to proceed with the testing of the null hypothesis for the three population genetic statistics.

Haplotype diversity (HD)

In order to detect signatures of intra-population diversity reduction in the populations subject to isolation factors we defined a threshold, as a function of sample size, representing the limit under which the null hypothesis could be rejected. The regression analyses showed that, for all reference populations, the threshold representing the lowest possible values of HD follow a trend which is an inverse function of sample size (R^2 from 0.971 to 0.978 for South Italy and Sardinia, respectively; p-value<0.001)(see supplementary Figure S1). As shown in Figure 1A, in the north-eastern region, we found 6 out of 18 populations with HD values falling below the threshold. Among them, we were able to reject the null hypothesis for a total of 5 populations as they provided a probability higher than 66.7% to have an HD value lower that what expected in a large outbred group at the same sample size (see Table II). As for the Central Italy region, we found 4 out of 9 populations with HD values falling below the threshold, all of them with probability higher than 66.7% (Figure 1B). None of the populations of the Southern Italy region provided HD falling below the threshold (Figure 1C), whereas we were able to reject the null hypothesis for only one Sardinia population (Figure 1D). Furthermore, the distribution of sample sizes of populations for which the null hypothesis could be accepted or rejected are not significantly different from each other (Mann-Whitney test; p=0.325), with the two groups showing the same median sample size (Supplementary Figure S2). This suggests that resampling methods represent a useful tool to evaluate signatures of genetic isolation for intra-population diversity without being substantially biased by differences in sample size.

Fixation index (Fst)

To evaluate the divergence of populations subject to isolation factors from the surrounding genetic background, we used the Fst statistic. This is the most commonly used statistic to compute genetic distances using mtDNA data and it has been shown to have high discrimination power (Kalinowski, 2005). The regression analyses showed that, for all four Italian regions, the best fitting line representing the higher possible A-Fst obtainable for an outbred populations at a given sample size is an inverse function of sample size itself (R^2 from 0.984 to 0.999 for Central Italy and South Italy, respectivelly; p-value<0.001) (see supplementary Figure S3). As shown in Figure 2A, we found that 6 out of 18 North-eastern Italian populations provided A-Fst point estimates higher than the threshold value, for which the null hypothesis could be rejected (see Table II). For the majority of these latter populations (4 out of 6) the probability to reject the null hypothesis was highly significant (>95%). As for the Central Italy region, most of the populations (7 out of 9) provided A-Fst values falling above the threshold, all of them with a probability higher than 66.7% to have higher values that expected in a large outbred group at the same sample size. On the other hand, none of the South Italian and Sardinia populations provided A-Fst values falling above the threshold. Similarly to what we observed for HD, the distributions of sample sizes of population showing or lacking signals of departure from the surrounding genetic background are not significantly different (Mann-Whitney test; p=0.404), with the median sample size value of the latter group being slightly higher (Supplementary Figure S4). Therefore, the use of resamplings seems to be exempt from biases due to differences in sample size.

Fu's FS

As concerns the analysis of this statistic though our resamplings we found that in North-east, Central and South Italy the expected signal of population expansion is always observed for sample sizes above 20, whereas for Sardinia for sample sizes above 25 (data not shown). It is worth noting that

this limit is close to the one identified by Ramos-Onsins and Rozas (2002) who showed that the Fu's FS test starts to lose statistical power with sample sizes of less than 15-20 (see figure 3 of the mentioned paper). Less than one fourth of populations (8 out of 34) subject to isolation factors analysed show a non significant value for Fu's FS ($p>0.02$) (Table II) indicating a stationary demographic size.

Evaluation of signals of genetic isolation

In current literature the concept of human population isolate is usually used to refer to groups, subject to geographic and/or cultural isolation factors (Arcos-Burgos and Muenke, 2002), that have evolved following a specific demographic model in terms of size of the founder group, population growth and gene flow with other groups. Under this model populations arise "from the founder effect of a small number of individuals as a consequence of some type of bottleneck", remain in "isolation over many generations without genetic interchange from other subpopulations" (Arcos-Burgos and Muenke, 2002) and have been subject to slow expansion after their foundation (Neel, 1992) or may have "experienced bottlenecks alternating with periods of rapid growth" (Peltonen et al., 2000). Populations that went through this kind of demographic history are expected to have increased levels of endogamy compared to large outbred groups (Peltonen et al., 2000; Varilo and Peltonen, 2004). Overall, their gene pools are more exposed to the effects of genetic drift and assortative mating, which results in population differentiation, lower heterozygosity and deeper gene genealogies compared to non-isolated populations (Schierup et al., 2000; Charlesworth and Wright, 2001; Arcos-Burgos and Muenke, 2002). Therefore, for populations evolved under such demographic model the null hypothesis is expected to be rejected for both A-Fst and HD statistics while the FS statistic should be insignificant.

The analysis of results at population level revealed the presence of five patterns which differentiated according to the number and the type of statistics for which the null hypothesis could be rejected (Table II).

The first pattern, found in 7 out of 34 populations, namely Cappadocia, Filettino, Pusteria valley, Sappada, Sauris, Timau, and Vallepietra, consists in the rejection of the null hypothesis for all three statistics, thus fully complying with what expected from a population evolved under the demographic model of human population isolates. With the exception of Pusteria valley, all these groups have very small census sizes (from 208 for Vallepietra to 1307 for Sappada) suggesting that their founder groups were also small. Notably, 4 out of 7 populations are settled in the Alpine mountain range (between 830 and 1217 meters above sea level for Timau and Sappada, respectively), an area which is characterized by physical barriers to gene flow, and three of them are also linguistic isolates (Sappada, Sauris and Timau; see Capocasa et al., 2013, 2014; Coia et al., 2012, 2013). The remaining three populations (Filettino, 1075 m.a.s.l.; Cappadocia, 1108 m.a.s.l.; Vallepietra, 825 m.a.s.l.) are still located in a mountainous environment (Central Italian Apennines). Interestingly, a certain degree of isolation for Sappada and Sauris was detected through autosomal microsatellites (Montinaro et al., 2012).

The most frequent of the other patterns (pattern 2) consists in the failure to reject the null hypothesis for all statistics and was observed in 19 out of 34 populations. Although it may indeed point to the lack of substantial barriers to gene flow in these populations, we cannot exclude completely the possibility that some of them may have experienced genetic isolation. Due to the rationale underlying the construction of the test, we can identify robust isolation signatures. However, we are unable to discriminate between random fluctuations of genetic statistics due to a reduced sample size and effects of isolation phenomena which have not been sufficiently intense or prolonged in time. In addition to the existence of this sort of grey area, we should also consider

the limited power which comes from the use of a single locus. Undoubtedly, the analysis of a broad panel of independent loci and the application of methods to estimate gene flow could help shed light on most of these cases.

Three other patterns are worthy of discussion. For the populations of Gardena valley Ladins and Jenne we could not reject the null hypothesis only for the FS statistic (pattern 3). The more obvious explanation is that these populations have probably been subject to some degree of demographic expansion. However, two alternative explanations are worth taking into account. The first implies a failure of the statistic (FS) to detect stationarity. In fact, very low effective population sizes, that would be reasonably expected in isolated populations which have originated from a small number of founders, may produce gene genealogies with very few mutations that can largely affect the power of haplotype-based demographic parameters (Ramírez-Soriano et al., 2008). The second scenario instead involves recent and severe bottleneck events, possibly followed by rapid growth, after which it is most likely that no lineages survive without coalescing resulting in a reduced genealogy size with a star-like shape (Ramírez-Soriano et al. 2008). As far as we know, none of the above populations experienced recent bottlenecks which means that the first of the two scenario is the most likely. For the populations of Isarco valley, Piglio, Saracinesco and Trevi we were able to reject the null hypothesis only as concerns the A-Fst statistic (pattern 4). One possible scenario is that these populations have evolved under a demographic model which differs from the one we considered here as regards the dimension of the founding group. In fact, the genetic diversity reduction for populations originated by a relatively large number of individuals will be less severe or even negligible. This will result in HD estimates comparable to that of large outbred groups. Furthermore, these populations will keep on expanding thus maintain signals of demographic expansion. On the other hand, the process of differentiation from their surrounding genetic background will be still ongoing, mediated by mutation and related to the time since foundation,

and admixture dynamics which involve neighbour populations but not the “isolated” one. The above demographic model has been used in current literature to describe the evolutionary history of “primary isolates” isolates by Neel (1992). This latter scenario may be suitable for the populations of Isarco valley, Piglio, and Trevi but not for Saracinesco whose census size is the lowest of the entire dataset (164).

A possible explanation for Saracinesco is related to the introgression of a small amount of haplotypes which are substantially different from those of the receiving population. In fact, the presence in this village of haplotypes belonging to haplogroups U3 and R0a suggests gene flow from the Eastern Mediterranean (Messina et al. 2015). We found one population (Luserna Cimbrians), for which the null hypothesis could only be rejected for FS but not for A-Fst and HD (pattern 5). A possible explanation for this pattern is that severe bottleneck events can produce an excess of rare frequency variants, an excess of haplotypes and an under-representation of major haplotypes, resulting in a higher level of haplotype diversity than what would be expected under a more relaxed bottleneck (Depaulis et al., 2005). Furthermore, the excess of haplotypes resulting from this mechanism may also help maintain a shared genetic diversity between the population undergone to the bottleneck and the parental one, resulting in low Fst values. This could be a reasonable explanation of the observed pattern for the Luserna Cimbrians, since historical records and previous genetic investigations suggests that this community was founded by reduced number of families (Coia et al., 2013 and citations therein).

Conclusion

In conclusion, our study suggests that the use of population genetic statistics combined with resampling analysis can provide the means to detect signatures of genetic isolation in human populations even when using a single, although highly informative, locus such as mtDNA. This may

turn out to be useful in two ways. Firstly, unambiguous evidence of genetic isolation may be retrospectively searched using the large body of data which are today available for mtDNA variation in human populations. This may be particularly important for groups whose DNA is difficult to resample or reanalyse, e.g. due to isolation breakdown, population dispersal or ethical concerns. Secondly, our approach may be used as a tool to select populations which are more likely to show strong signatures of genetic isolation also for autosomal loci and, which therefore, can be potentially more informative for gene mapping and gene-disease association studies.

References

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

Babalini C, Martinez-Labarga C, Tolk HV, Kivisild T, Giampaolo R, Tarsi T, Contini I, Barać L, Jančijević B, Martinović Klarić I, Peričić M, Sujoldžić A, Villems R, Biondi G, Rudan P, Rickards O. 2005. The population history of the Croatian linguistic minority of Molise (southern Italy): a maternal view. *Eur J Hum Genet.* 13:902-912.

Beja-Pereira A, Caramelli D, Lalueza-Fox C, Vernesi C, Ferrand N, Casoli A, Goyache F, Royo LJ, Conti S, Lari M, Martini A, Ouragh L, Magid A, Atash A, Zsolnai A, Boscato P, Triantaphylidis C, Ploumi K, Sineo L, Mallegni F, Taberlet P, Erhardt G, Sampietro L, Bertranpetit J, Barbujani G, Luikart G, Bertorelle G. 2006. The origin of European cattle: evidence from modern and ancient DNA. *Proc Natl Acad Sci U S A*. 103:8113-8118.

Boattini A, Griso C, Pettener D. 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.

Bosch E, Calafell F, González-Neira A, Flaiz C, Mateu E, Scheil HG, Huckenbeck W, Efremovska L, Mikerezi I, Xirotiris N, Grasa C, Schmidt H, Comas D. 2006. Paternal and maternal lineages in the Balkans show a homogeneous landscape over linguistic barriers, except for the isolated Aromuns. *Ann Hum Genet.* 70:459-487.

Brandstätter A, Egyed B, Zimmermann B, Duftner N, Padar Z, Parson W. 2007. Migration rates and genetic structure of two Hungarian ethnic groups in Transylvania, Romania. *Ann Hum Genet.* 71:791-803.

Brisighelli F, Álvarez-Iglesias V, Fondevila M, Blanco-Verea A, Carracedo A, Pascali VL, Capelli C, Salas A. 2012. Uniparental markers of contemporary Italian population reveals details on its pre-Roman heritage. *PLoS One.* 7:e50794.

Calò CM, Corrias L, Vona G, Bachis V, Robledo R. 2012. Sampling strategies in a linguistic isolate: results from mtDNA analysis. *Am J Hum Biol.* 24:192-194.

Capocasa M, Battaggia C, Anagnostou P, Montinaro F, Boschi I, Ferri G, Alu M, Coia V, Crivellaro F, Destro Bisol G. 2013. Detecting genetic isolation in human populations: a study of European language minorities. *PLoS One.* 8:e56371.

Capocasa M, Anagnostou P, Bachis V, Battaggia C, Bertoncini S, Biondi G, Boattini A, Boschi I, Brisighelli F, Caló CM, Carta M, Coia V, Corrias L, Crivellaro F, De Fanti S, Dominici V, Ferri G, Francalacci P, Franceschi ZA, Luiselli D, Morelli L, Paoli G, Rickards O, Robledo R, Sanna D, Sanna E, Sarno S, Sineo L, Taglioli L, Tagarelli G, Tofanelli S, Vona G, Pettener D, Destro Bisol G. 2014. Linguistic, geographic and genetic isolation: a collaborative study of Italian populations. J *Anthropol Sci.* 92:201-231.

Charlesworth D, Wright SI. 2001. Breeding systems and genome evolution. *Curr Opin Genet Dev.* 11:685-690.

Cilli E, Delaini P, Costazza B, Giacomello L, Panaino A, Gruppioni G. 2011. Ethno-anthropological and genetic study of the Yaghnobis; an isolated community in Central Asia. A preliminary study. *J Anthropol Sci.* 89:189-194.

Coia V, Boschi I, Trombetta F, Cavulli F, Montinaro F, Destro Bisol G, Grimaldi S, Pedrotti A. 2012. Evidence of high genetic variation among linguistically diverse populations on a micro-geographic scale: a case study of the Italian Alps. *J Hum Genet.* 57:254-260.

Coia V, Capocasa M, Anagnostou P, Pascali V, Scarnicci F, Boschi I, Battaggia C, Crivellaro F, Ferri G, Alù M, Brisighelli F, Busby GB, Capelli C, Maixner F, Cipollini G, Viazzo PP, Zink A, Destro Bisol G. 2013. Demographic histories, isolation and social factors as determinants of the genetic structure of Alpine linguistic groups. *PLoS One.* 8:e81704.

Congiu A, Anagnostou P, Milia N, Capocasa M, Montinaro F, Destro Bisol G. 2012. Online databases for mtDNA and Y chromosome polymorphisms in human populations. *J Anthropol Sci.* 90:201-215.

Depaulis F, Mousset S, Veuille M. 2005. Detecting selective sweeps with haplotype tests. In: Nurminsky D, editor. Selective Sweep. Georgetown: Landes Bioscience. 34p.

Destro Bisol G, Anagnostou P, Batini C, Battaggia C, Bertoncini S, Bottini A, Caciagli L, Caló CM, Capelli C, Capocasa M, Castrì, L, Ciani G, Coia V, Corrias L, Crivellaro F, Ghiani ME, Luiselli D, Mela C, Melis A, Montano V, Paoli G, Sanna E, Rufo F, Sazzini M, Taglioli L, Tofanelli S, Useli A, Vona G, Pettener

D. 2008. Italian isolates today: geographic and linguistic factors shaping human biodiversity. *J Anthropol Sci.* 86: 179-188.

Esko T, Mezzavilla M, Nelis M, Borel C, Debniak T, Jakkula E, Julia A, Karachanak S, Khrunin A, Kisfali P, Krulisova V, Aušrelé Kučinskiené Z, Rehnström K, Traglia M, Nikitina-Zake L, Zimprich F, Antonarakis SE, Estivill X, Glavač D, Gut I, Klovins J, Krawczak M, Kučinskas V, Lathrop M, Macek M, Marsal S, Meitinger T, Melegh B, Limborska S, Lubinski J, Paolotie A, Schreiber S, Toncheva D, Toniolo D, Wichmann HE, Zimprich A, Metspalu M, Gasparini P, Metspalu A, D'Adamo P. 2013. Genetic characterization of northeastern Italian population isolates in the context of broader European genetic diversity. *Eur J Hum Genet.* 21:659-665.

Excoffier L, Lischer HEL. 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.

Falchi A, Giovannoni L, Calò CM, Piras SI, Moral P, Paoli G, Vona G, Varesi L. 2006. Genetic history of some western Mediterranean human isolates through mtDNA HVRI polymorphisms. *J Hum Genet.* 51:9-14.

Fernandes AT, Velosa R, Jesus J, Carracedo A, Brehm A. 2003. Genetic differentiation of the Cabo Verde archipelago population analysed by STR polymorphisms. *Ann Hum Genet*. 67:34034-7.

Fu YX. 1997. Statistical tests of neutrality of mutations against population growth, hitchhiking and backgroud selection. *Genetics* 147:915-925.

Kalinowski ST. 2004. Do polymorphic loci require large sample sizes to estimate genetic distances? *Heredity* 94:33-36.

Laland KN, Odling-Smee J, Myles S. 2010. How culture shaped the human genome: bringing genetics and the human sciences together. *Nat Rev Genet.* 11:137-148.

Lauc T, Rudan P, Rudan I, Campbell H. 2003. Effect of inbreeding and endogamy on occlusal traits in human isolates. *J Orthod.* 30:301-308.

Messina F, Scorrano G, Labarga CM, Rolfo MF, Rickards O. 2010. Mitochondrial DNA variation in an isolated area of Central Italy. *Ann Hum Biol.* 37:385-402.

Messina F, Finocchio A, Rolfo MF, De Angelis F, Rapone C, Coletta M, Martínez-Labarga C, Biondi G, Berti A, Rickards O. 2015. Traces of forgotten historical events in mountain communities in Central Italy: A genetic insight. *Am J Hum Biol.* 27:508-519.

Montinaro F, Boschi I, Trombetta F, Merigioli S, Anagnostou P, Battaggia C, Capocasa M, Crivellaro F, Destro-Bisol G, Coia V. 2012. Using forensic microsatellites to decipher the genetic structure of linguistic and geographic isolates: a survey in the eastern Italian Alps. *Forensic Sci Int Genet.* 6:827-833.

Neel J. 1992. Minority populations as genetic isolates: the interpretation of inbreeding results. In: Bittles AH, Roberts DF, editors. Minority Populations: Genetics Demography and Health. London: The MacMillan Press. 1p.

Nei M. 1987. *Molecular evolutionary genetics*. New York: Columbia University Press.

Peltonen L, Palotie A, Lange K. 2000. Use of population isolates for mapping complex traits. *Nat Rev Genet.* 1:182-190.

Pichler I, Mueller JC, Stefanov SA. De Grandi A, Beu Volpato C, Pinggera GK, Mayr A, Ogriseg M, Ploner F, Meitinger T, Pramstaller P. 2006. Genetic structure in contemporary South Tyrolean isolated populations revealed by analysis of Y-chromosome, mtDNA and Alu polymorphisms. *Hum Biol.* 78:441-464.

Ramírez-Soriano A, Ramos-Onsins SE, Rozas J, Calafell F, Navarro A. 2008. Statistical power analysis of neutrality tests under demographic expansions, contractions and bottlenecks with recombination. *Genetics.* 179:555-567.

Ramos-Onsins SE, Rozas J. 2002. Statistical properties of new neutrality tests against population growth. *Mol Biol Evol.* 19:2092-2100.

Reynolds J, Weir BS, Cockerham CC 1983. Estimation of the coancestry coefficient: Basis for a short-term genetic distance. *Genetics*. 105: 767–779.

Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW. 2002. Genetic structure of human populations. *Science.* 298:2381-2385.

Schierup MH, Charlesworth D, Vekemans X. 2000. The effect of hitch-hiking on genes linked to a balanced polymorphism in a subdivided population. *Genet Res.* 76:63–73.

Thomas MG, Barnes I, Weale ME, Jones AL, Forster P, Bradman N, Pramstaller P. 2008. New genetic evidence support isolation and drift in the Ladin communities of the South Tyrolean Alps but not an ancient origin in the Middle East. *Eur J Hum Genet.* 16:124-134.

Tishkoff SA, Kidd KK. 2004. Implications of biogeography of human populations for 'race' and medicine. *Nat Genet.* 36:S21-S27.

Tolk HV, Pericić M, Barać L, Klarić IM, Janićijević B, Rudan I, Parik J, Villems R, Rudan P. 2001. MtDNA haplogroups in the populations of Croatian Adriatic Islands. *Coll Anthropol.* 24:267-280.

Toso F. 2008. Le minoranze linguistiche in Italia. Il Mulino, Bologna.

Toso F. 2014. The study of language islands: an interdisciplinary approach. J Anthropol Sci. 92:I-IV.

van Oven M, Hämmerle JM, van Schoor M, Kushnick G, Pennekamp P, Zega I, Lao O, Brown L, Kennerknecht I, Kayser M. 2011. Unexpected island effects at an extreme: reduced Y chromosome and mitochondrial DNA diversity in Nias. Mol Biol Evol. 28:1349-1361.

Varilo T, Peltonen L. 2004. Isolates and their potential use in complex gene mapping efforts. Curr Opin Genet Dev. 14:316-323.

Veeramah KR, Tönjes A, Kovacs P, Gross A, Wegmann D, Geary P, Gasperikova D, Klimes I, Scholz M, Novembre J, Stumvoll M. 2011. Genetic variation in the Sorbs of eastern Germany in the context of broader European genetic diversity. Eur J Hum Genet. 19:995-1001.

Table I. List of populations subject to geographical and/or linguistic isolation under study.

Abbreviations: N=sample size, G, geographic isolate; GL, geo-linguistic isolate; L, linguistic isolate.

Population	Label	N	Status	Census size*	Reference
North-east Italy					
Badia valley Ladins	LVB	56	GL	10644	Thomas et al., 2008
Fassa valley Ladins	LVF	47	GL	9894	Coia et al., 2012
Fersina valley	FER	25	G	2575	Coia et al., 2012
Fiemme valley	FIE	41	G	18990	Capocasa et al., 2014
Gardena valley Ladins	LVG	46	GL	10198	Thomas et al., 2008
Giudicarie valley	GIU	52	G	36282	Coia et al., 2012
Isarco valley	VIS	34	G	47492	Pichler et al., 2006
Lessinia Cimbrians	LES	40	GL	13455	Capocasa et al., 2013
Luserna Cimbrians	LUS	21	GL	286	Coia et al., 2012
Non valley	NON	48	G	37832	Coia et al., 2012
Primiero valley	PRI	40	G	9959	Coia et al., 2012
Pusteria valey	VPU	37	G	76149	Pichler et al., 2006
Sappada	SAP	59	GL	1307	Capocasa et al., 2013
Sauris	SAU	48	GL	429	Capocasa et al., 2013
Sole valley	SOL	63	G	15235	Coia et al., 2012
Timau	TIM	46	GL	500	Capocasa et al., 2013
Upper Venosta valley	VVA	50	G	8533	Thomas et al., 2008
Lower Venosta valley	VVB	52	G	5144	Thomas et al., 2008
Central Italy					
Filettino	FIL	43	G	554	Messina et al., 2015
Cappadocia	CAP	89	G	537	Messina et al., 2015
Piglio	PIG	96	G	4775	Messina et al., 2015
Saracinesco	SAR	35	G	164	Messina et al., 2015
Trevi	TRV	58	G	8447	Messina et al., 2015
Vallepietra	VAP	46	G	208	Messina et al., 2015
Vagli	VAG	22	G	995	Capocasa et al., 2014
Jenne	JEN	103	G	407	Messina et al., 2010
Tocco da Casasuria	TOC	50	G	2782	Verginelli et *al.,* 2003
South Italy					
Molise Croats	CRM	41	L	1884	Babalini et al., 2005
Circello	CIR	27	G	2501	Capocasa et al., 2014
Calabria Arberesh	ARC	87	GL	28034	Capocasa et al., 2014
Salento Grecanici	GRC	47	L	40000	Brisighelli et *al.,* 2012
Sardinia					
Benetutti	BEN	50	G	2010	Capocasa et al., 2014
Carloforte	CFT	51	GL	6420	Calò et al., 2012
Sant'Antioco	SNA	42	G	2919	Falchi et al., 2006

* ISTAT (2011) (http://demo.istat.it)

Table II. Genetic diversity, demographic statistics and probability of acceptance of the null hypothesis in the 34 populations subject to cultural and/or geographical isolation. Abbreviations: S.D.=standard deviation.

		HD			FS		A-Fst		
Population	Region	Point estimate	s.d.	Prob. to reject null hypothesis	Value	p-value	Point estimate	s.d.	Prob. to reject null hypothesis
Pattern 1									
Cappadocia	Central Italy	0.920	0.011	99.67%	-3.714	0.125	0.049	0.013	99.97%
Filetino	Central Italy	0.920	0.018	79.50%	-4.011	0.062	0.025	0.008	98.92%
Pusteria valey	North-east Italy	0.890	0.032	88.50%	-3.831	0.065	0.028	0.009	94.62%
Sappada	North-east Italy	0.778	0.049	99.96%	-1.624	0.297	0.083	0.022	99.95%
Sauris	North-east Italy	0.923	0.020	75.27%	-4.293	0.067	0.047	0.009	100.00%
Timau	North-east Italy	0.901	0.026	91.47%	-3.843	0.079	0.028	0.014	89.28%
Vallepietra	Central Italy	0.928	0.019	69.20%	-4.573	0.067	0.026	0.007	99.73%
Pattern 2									
Badia valley Ladins	North-east Italy	0.945	0.017	39.77%	-18.661	0.000	0.009	0.013	51.36%
Calabria Arberesh	South Italy	0.977	0.009	13.64%	-25.791	0.000	0.000	0.000	0.00%
Carloforte	Sardinia	0.979	0.012	0.09%	-25.764	0.000	0.003	0.007	0.34%
Circello	South Italy	0.960	0.023	44.92%	-10.992	0.000	0.003	0.007	2.46%
Fassa valley Ladins	North-east Italy	0.933	0.026	55.67%	-14.142	0.001	0.002	0.005	3.14%
Fersina valley	North-east Italy	0.930	0.030	28.87%	-5.191	0.010	0.000	0.000	0.00%
Fiemme valley	North-east Italy	0.955	0.018	9.97%	-10.280	0.000	0.004	0.010	21.84%
Giudicarie valley	North-east Italy	0.972	0.011	0.17%	-25.586	0.000	0.004	0.010	27.12%
Lessinia Cimbrians	North-east Italy	0.953	0.017	10.79%	-13.817	0.000	0.004	0.011	21.70%
Lower Venosta valley	North-east Italy	0.953	0.018	21.49%	-25.908	0.000	0.001	0.004	1.65%
Molise Croats	South Italy	0.970	0.015	31.10%	-18.745	0.000	0.000	0.000	0.00%
Non valley	North-east Italy	0.957	0.019	14.30%	-25.526	0.000	0.003	0.007	12.25%
Primiero valley	North-east Italy	0.974	0.010	0.00%	-16.647	0.000	0.000	0.000	0.00%
Salento Grecanici	South Italy	0.989	0.007	0.01%	-25.750	0.000	0.000	0.000	0.00%
Sant'Antioco	Sardinia	0.942	0.029	40.74%	-24.478	0.000	0.000	0.000	0.00%
Sole valley	North-east Italy	0.953	0.015	25.21%	-22.433	0.000	0.005	0.009	37.58%
Tocco da Casasuria	Central Italy	0.990	0.008	0.00%	-25.538	0.000	0.003	0.007	31.60%
Upper Venosta valley	North-east Italy	0.955	0.017	15.62%	-16.602	0.000	0.000	0.000	0.00%
Vagli	Central Italy	0.948	0.029	7.99%	-5.690	0.003	0.000	0.000	0.00%
Pattern 3									
Gardena valley Ladins	North-east Italy	0.869	0.036	96.75%	-6.072	0.011	0.023	0.014	79.91%
Jenne	Central Italy	0.833	0.036	99.95%	-22.442	0.000	0.027	0.014	95.03%
Pattern 4									
Isarco valley	North-east Italy	0.961	0.015	0.88%	-9.411	0.000	0.051	0.013	99.68%
Piglio	Central Italy	0.954	0.011	40.64%	-24.040	0.000	0.012	0.011	76.03%
Saracinesco	Central Italy	0.940	0.021	30.19%	-5.101	0.018	0.021	0.014	81.24%
Trevi	Central Italy	0.949	0.013	33.14%	-13.110	0.000	0.030	0.012	98.19%
Pattern 5									
Luserna Cimbrians	North-east Italy	0.919	0.034	32.65%	-1.373	0.265	0.022	0.016	47.40%
Pattern 6									
Benetutti	Sardinia	0.917	0.029	79.17%	-14.519	0.000	0.021	0.012	43.53%

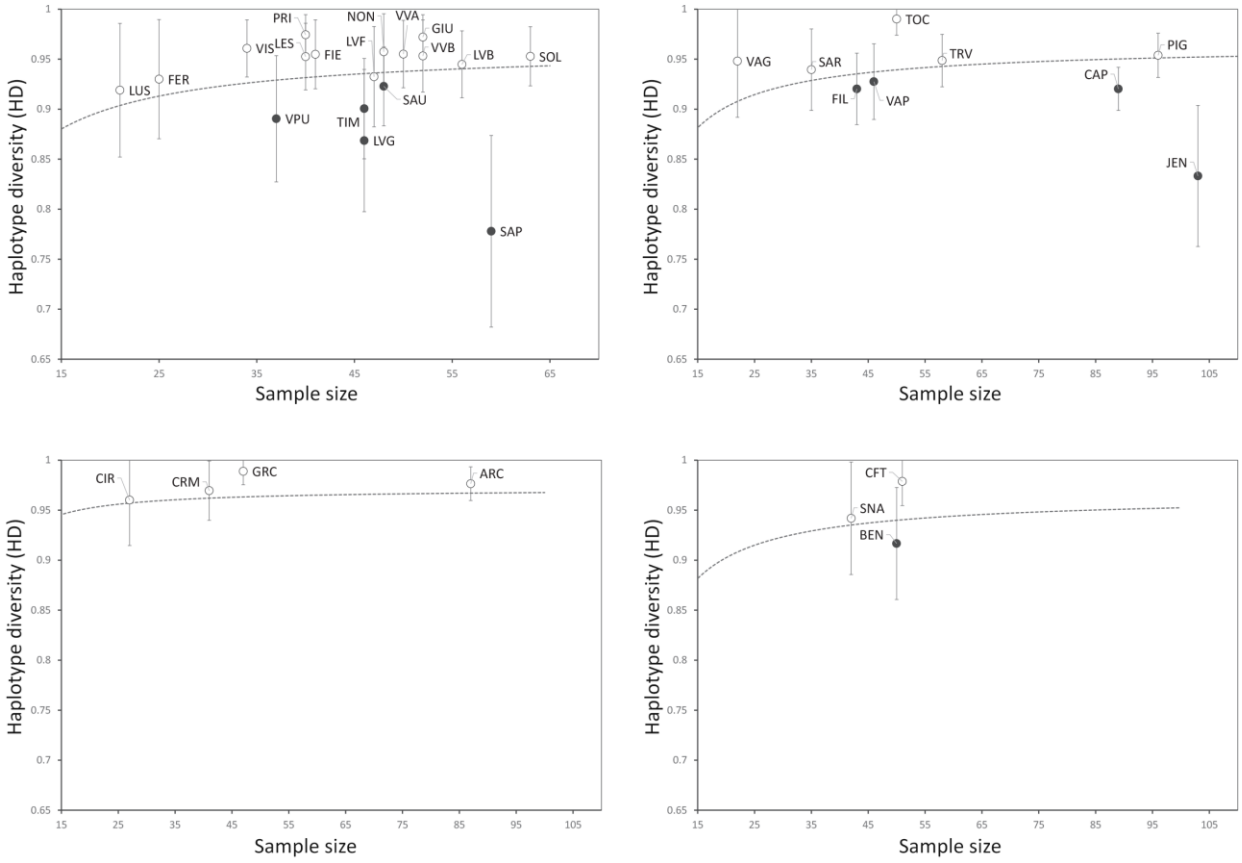

Figure 1. Comparison of HD values in respect to sample size between populations subject to isolation factors and the lower observable value (continuous line) in (A) North-east Italy, (B) Central Italy, (C) South Italy and (D) Sardinia. Vertical bars indicate 95% confidence intervals and filled circles represent the populations for which the null hypothesis could be rejected. Population labels as in Table I.

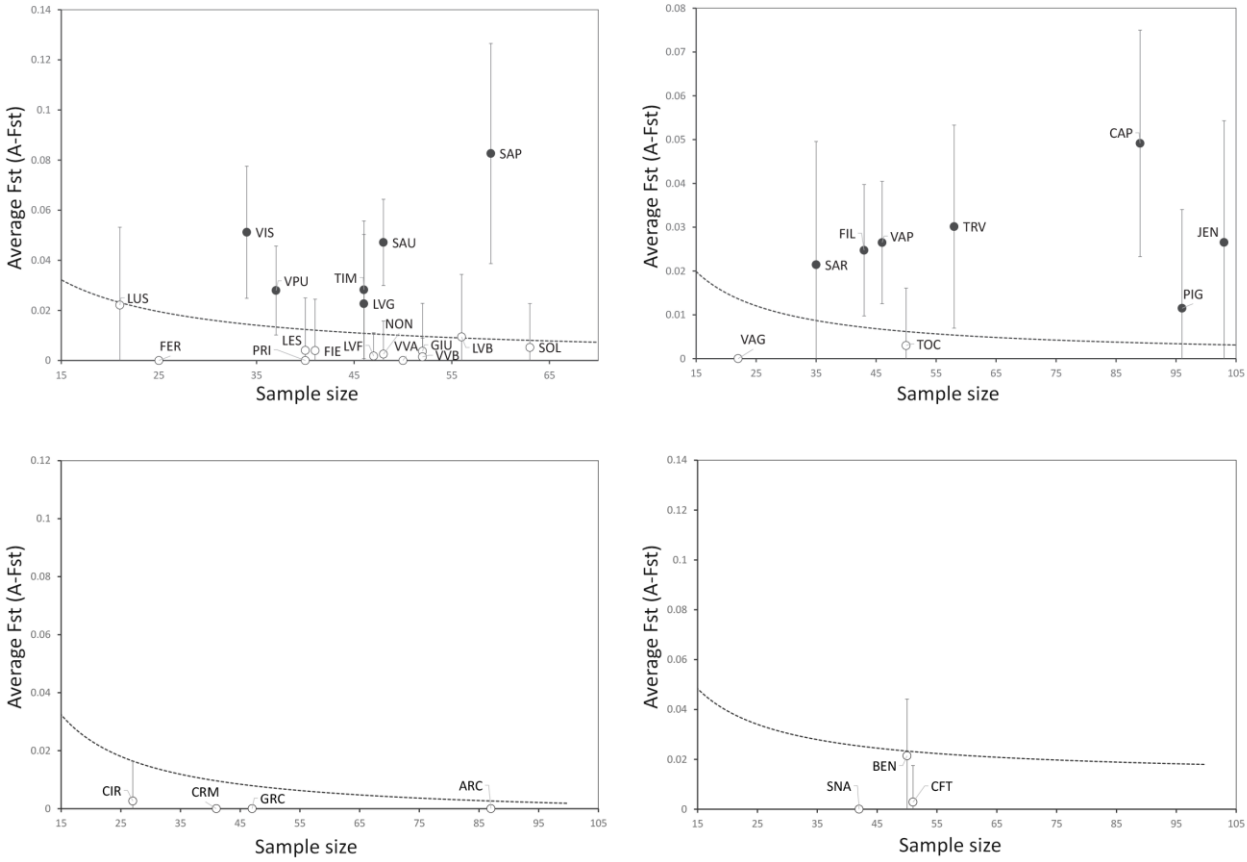

Figure 2. Comparison of A-Fst values in respect to sample size between populations subject to isolation factors and higher observable value (continuous line) in (A) North-east Italy, (B) Central Italy, (C) South Italy and (D) Sardinia. Vertical bars indicate 95% confidence intervals and filled circles represent the populations for which the null hypothesis could be rejected. Population labels as in Table I.