The surname structure of Trentino (Italy) and its relationship with dialects and genes

The surname structure of Trentino (Italy) and its relationship with dialects and genes

Alessio Boattini, Eugenio Bortolini, Roland Bauer, Marta Ottone, Rossella Miglio, Paola Gueresi & Davide Pettener

The surname structure of Trentino (Italy) and its relationship with dialects and genes

Alessio Boattini[a]*, Eugenio Bortolini[b]*, Roland Bauer[c], Marta Ottone[d], Rossella Miglio[e], Paola Gueresi[e]† and Davide Pettener[a]

[a]Department of Biological, Geological and Environmental Sciences (BIGEA), University of Bologna, Bologna, Italy; [b]Department of Cultural Heritage, University of Bologna, Ravenna, Italy; [c]Fachbereich Romanistik, Universität Salzburg, Austria; [d]Epidemiology Unit, Azienda USL-IRCCS di Reggio Emilia, Reggio Emilia, Italy; [e]Department of Statistical Sciences, University of Bologna, Italy

ABSTRACT

Background: Thanks to the availability of rich surname, linguistic and genetic information, together with its geographic and cultural complexity, Trentino (North-Eastern Italy) is an ideal place to test the relationships between genetic and cultural traits.

Aim: We provide a comprehensive study of population structures based on surname and dialect vari-ability and evaluate their relationships with genetic diversity in Trentino.

Subjects and methods: Surname data were collected for 363 parishes, linguistic data for 57 dialects and genetic data for different sets of molecular markers (Y-chromosome, mtDNA, autosomal) in 10 populations. Analyses relied on different multivariate methods and correlation tests.

Results: Besides the expected isolation-by-distance-like patterns (with few local exceptions, likely related to sociocultural instances), we detected a significant and geography-independent association between dialects and surnames. As for molecular markers, only Y-chromosomal STRs seem to be associated with the dialects, although no significant result was obtained. No evidence for correlation between molecular markers and surnames was observed.

Conclusion: Sumames act as cultural markers as do other words, although in this context they cannot be used as reliable proxies for genetic variability at a local scale.

Introduction

The relationships between genes, language, cultural markers and geography are at the core of a vast field of study (identi-fied by the general name of Dual Inheritance Theory, Cultural Evolutionary Theory) broadly aimed at investigating the co-evolutionary processes underlying biological and cul-tural change over time in human populations (Cavalli-Sforza and Feldman 1981; Boyd and Richerson 1985). A conspicuous section of this unified field of study is interested in under-standing the mechanisms responsible for the diffusion of human culture at different geographic scales (e.g. migratory events, cultural diffusion, etc.; Jordan et al. 2016; Bortolini et al. 2017 among others). Language, in particular, being one of the most apparent and ubiquitous human cultural markers, represents the focus of many pivotal studies (Cavalli-Sforza et al. 1988; Gray and Atkinson 2003; Gray et al. 2011). Exploring the distribution of linguistic features and genetic variants, however, may greatly benefit from evidence on the distribution of surname variability (Cavalli-Sforza et al. 2004). Since surnames are words, a significant association between their distribution and that of dialects/languages could be expected (Manni et al. 2005; Manni et al. 2008). At the same time, in the vast majority of cases, surnames are characterised by patrilineal heritability, and can therefore be considered as proxies for Y-chromosomal genetic markers (King and Jobling 2009a, 2009b). For this reason, due to their remarkable variability, they have been widely used to infer the genetic structure of human populations – especially in case studies entailing a smaller geographical and temporal scale (Pettener 1990; Blanco Villegas et al. 2004; Boattini et al. 2007; Fiorini et al. 2007; Boattini et al. 201 0). In add-ition, thanks to the availability of archival historical sources, changes over time in surnames themselves can be effectively studied as privileged support to infer local popula-tion histories.

Nevertheless, the relationship between surnames and Y-chromosomes is not always characterised by a perfect match, since surnames may be polyphyletic and their patrilineal transmission may be interrupted by non-paternity events, just to mention the most important deviations (King and Jobling 2009a; Larmuseau et al. 2017; Larmuseau et al. 2019).

In particular, a high degree of surname/Y-chromosome ancestry is observed for rare surnames, while frequent ones are more prone to deviations from the expected uniparental genetic pattern (King and Jobling 2009b; McEvoy and Bradley 2006; Solé-Morata et al. 2015; Martinez-Cadenas et al. 2016; Claerhout et al. 2020).

It is interesting to note that, while the demographic implications of surname distributions are well explored, the relationships between surnames and languages/dialects are comparatively less understood. To date, this issue was properly addressed only in the Netherlands, where "surnames cannot be taken as a proxy for dialect variation, even though they can be safely used as a proxy for Y-chromosome genetic variation" (Manni et al. 2008). However, the fact that Dutch surnames have a more recent origin than those recorded in other countries (~200 years ago), together with the peculiar geographical features of that region, suggests that these conclusions cannot be extended to other contexts.

In the present study, we test the hypothesis of congruence between genetic variability, linguistic variability, and surname distribution in Trentino, a historic region of North-Eastern Italy. Trentino offers an ideal perspective on this issue, due to the presence of the most important cultural/environmental variables which usually affect genetic variabil-ities in wider human populations, such as environmental bar-riers (presence of mountains and valleys), social stratification (urban vs. rural populations) and linguistic heterogeneity, including the presence of substantial ethnolinguistic minor-ities (i.e. Ladins, Mocheni, Cimbri).

To date, studies involving individual human groups of Trentino focus on just one of the above mentioned variables. In more detail, there have been investigations on surname variability (Pettener et al. 1994; Gueresi et al. 2000, 2001; Boattini et al. 2006) and, more recently, on linguistic (Goebl et al. 1998; Bauer 2012; Goebl 2012) and genetic variability (Coia et al. 2012; Montinaro et al. 2012; Coia et al. 2013). These efforts, together with the long-standing activity of the Archdiocese of Trento in documenting its historic-demographic heritage, produced a series of datasets that yield detailed information on surname, dialect and genetic variation in Trentino. However, none of the cited studies explored the interaction between all the mentioned variables. In the present study we aim to explain the observed distribution of surnames (intended as an 'intermediate' cultural/genetic trait) in Trentino by comparing it against the distribution of fully cultural/linguistic traits (dialects), and – at the same time – against the distribution of molecular/genetic markers (Y-chromosome, mitochondrial DNA, and autosomal DNA). For the first time, we provide a comprehensive study of population structures based on surnames and dia-lect variability. We then formally compare these structures with the genetic variability and geographic distribution of the sampled populations. Our final goal is to ascertain: a) the relative impact of genetic and cultural processes on surname distribution; b) the specific co-evolutionary processes between linguistic and genetic diversity in Trentino; and c) the possible role played by spatial segregation and geographic distance.

Materials

Geographic and demographic context

The region of Trentino is located in North-Eastern Italy and takes its name from the city of Trento, its secular and religious capital. It extends for 6212 km^2 and is mainly occupied by steep mountains, most notably the Dolomites mountain range, whose highest peak is Marmolada, at 3343 m. Trentino is crossed from north to south by the Adige Valley, which also hosts the principal communication route as well as the main cities (Trento, Rovereto). Other important Trentino valleys are Val di Non and Val di Sole (north-west), Giudicarie (south-west), Fiemme and Fassa (north-east) and Sugana (south-east), the latter hosting another ancient and important communication route ("Via Claudia Augusta"). Based on its geographic structure, Trentino is administratively subdivided into 16 "Comunità di Valle" (Valley Communities), to which we will refer as units of analysis for summary statistics (Figure 1, Supplementary Figure 1, Supplementary Table 1).

The current Trentino population size is around 505,000. Historical censuses from 19th and early 20th centuries show that this region experienced an appreciable demographic growth, at least from 1869 (335,591) to 1921 (387,809), after which followed a modest decline (372,084 in 1931), reportedly due to strong emigration and depopulation (Ascolani 2010; seriestoriche.istat.it). A new and still ongoing phase of population growth started after the Second World War (seriestoriche.istat.it).

The data

Surnames

The data used in this study were extracted from the "Nati in Trentino" database compiled by the Archdiocese of Trento in collaboration with the Province of Trento. The database version to which we were granted access includes 1,254,623 baptismal registrations recorded from 1815 to 1923. Each record comprises birth date, name of the parish, name, surname and sex of the newborn, names and surnames of the parents. Here, we focussed on a generation-long period (25 years) from 1897 to 1923. This interval was selected in order to: a) avoid the superimposition of different generations; and b) exploit the highest number of active parishes, given its increase with time, accordingly to the increase of the population size. In total, we used 312,314 complete baptism records from 1897 to 1923. As for the units of analysis, we relied on the 432 Trentino parishes included in the dataset, which we grouped in 363 units (to which we still refer as parishes) after merging those parishes with too few records and/or too few years of coverage. The merging criteria and a complete list of the parishes/populations are detailed in Supplementary Text and Supplementary Table 2.

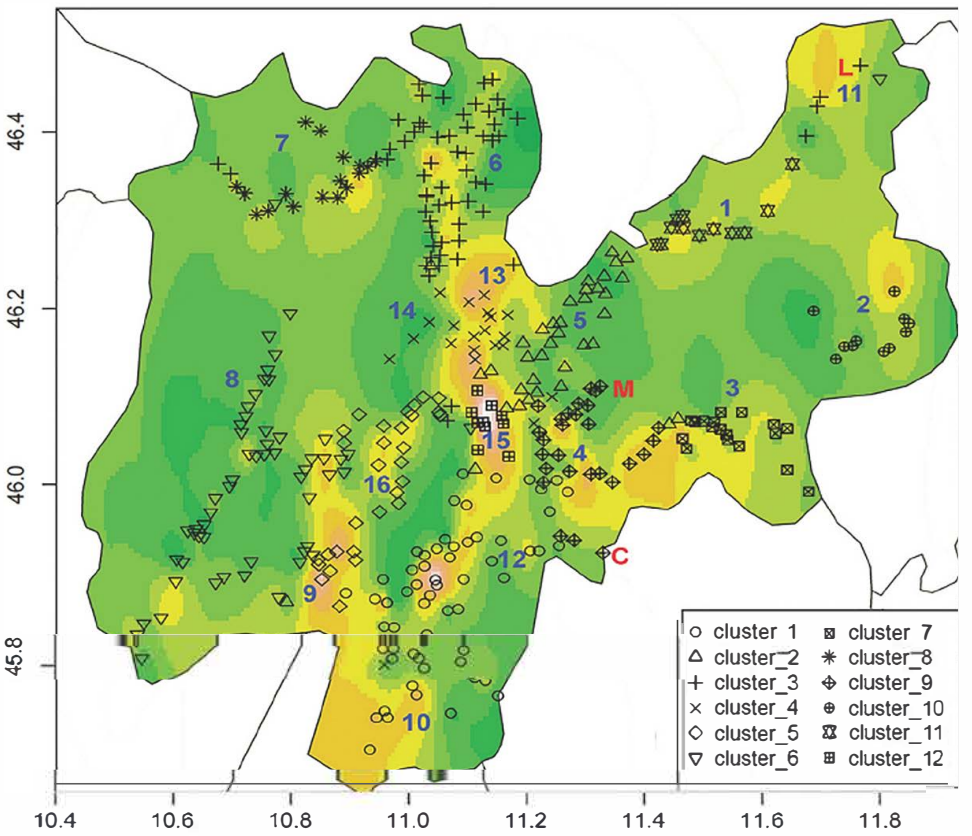

Figure 1. Surname diversity and distribution of surname Clusters in Trentino. Surname diversity (Shannon index) is represented by colour gradients. Points represent the 363 considered parishes. Point characters are associated with the 12 observed surname Clusters, as in the legend. Blue numbers identify the 16 "Comunità di Valle" administrative units (for details see Supplementary Table 1). Red letters identify the position of ethnolinguistic minorities (Abbreviations: C: Cimbrians; M: Mocheni; L: Ladins).

Dialects

Dialect data were obtained from the Linguistic Atlas of Dolomitic Ladin and Neighbouring Dialects (ALD). The latter consists of a linguistic atlas published in two issues (Issue I: Goebl et al. 1998; Issue II: Goebl 2012) and comprising nine cartographic volumes and 1950 thematic maps on dialects. The whole of the investigated area amounts to about 24,500 km^2 and includes 217 inquiry points, 60 of which are located in Trentino. Each thematic map focuses on a specific linguistic question (involving 57 phonetic, lexical, morpho-logical, or syntactic elements), e.g. the local variants for terms such as "sun", "church", "dog" and so on. As far as Trentino is concerned, in sampling areas where German-Bavarian dialects are the dominant ones (e.g. Cymbric of Luserna, point 118), ALD does not document Germanic basi-lect and reports instead Romance mesolect spoken by local populations in their exchange with non-local speakers. The first issue of ALD (Goebl et al. 1998), which is mostly focused on phonetic traits, also comprised an intensive, systematic taxonomic description, while dialectometric analyses are cur-rently being performed at the Department of Romance Philology of Salzburg University (Bauer 2003, 2009, 2014, 2016, Bauer and Casalicchio 2017). For this study, we retained the 57 dialects matching with at least one of the considered 363 parishes. When more than one parish matched with a given dialect, we selected the parish with the highest number of records. A full list of the considered dialects with the corresponding parishes is reported in Supplementary Table 3 (where ID indicates the location-number used in the ALD dataset).

Genetics

Genetic information about Trentino populations was retrieved from a series of studies (Coia et al. 2012; Montinaro et al. 2012; Coia et al. 2013) that shared the same sampling scheme while yielding data on different markers, namely Y-chromosome (50 SNPs, 17 STRs), mtDNA (17 coding region SNPs, HVR I and II sequences), and autosomal DNA (15 STRs). All these studies comprise ten Trentino populations, which were matched to their corresponding parishes and dialects (Supplementary Table 4). The criteria used for genes/dialects/ surnames matching are described in the Supplementary Text.

Methods

Surname analysis

After removing rare surnames (frequency lower than 10) from the dataset, we evaluated within-population surname variability by calculating the following parameters: a) a total number of surnames, b) number of different surnames, c) surname diversity (Shannon's index: Lewontin 1972). A surface plot of surname diversity in Trentino was obtained using

Table 1. Distance correlation, bias corrected distance correlation (bcdCor) or partial distance correlation (pdCor) among surnames, dialects, molecular markers and geography.

Matrix1	Matrix2	Matrix3 (partial test)	bcDcor / pdCor	p value	Bonferroni
Surnames/Geography (362 points)					
Surnames	Geography (least cost)		0.25	<0.001	<0.001
Surnames	Geography (resistance)		0.30	<0.001	<0.001
Surnames	Geography (great circle)		0.26	<0.001	<0.001
Surnames/Dialects (57 points) Dialects	Surnames		0.33	<0.001	<0.001
Dialects	Geography (great circle)		0.38	<0.001	<0.001
Surnames	Geography (great circle)		0.38	<0.001	<0.001
Dialects	Surnames	Geography (great circle)	0.22	0.001	0.02
Surnames/Dialects/Genetics (10 points)					
Surnames	Y STR		0.08	0.3200	1
Surnames	Y Hgs		0.06	0.6400	1
Surnames	mtDNA seq		0.16	0.8200	1
Surnames	mtDNA SNPs		0.03	0.5700	1
Surnames	Autosomal STRs		0.03	0.5800	1
Dialects	Y STR		0.27	0.0550	1
Dialects	Y Hgs		0.16	0.1700	1
Dialects	mtDNA seq		0.02	0.5400	1
Dialects	mtDNA SNPs Autosomal		0.05	0.6200	1
Dialects	STRs Geography (great		0.04	0.5900	1
Y STR	circle)		0.14	0.2100	1

the kriging algorithm as implemented in the R package *kriging* (Olmedo 2011).

In order to classify and evaluate the surname structure of Trentino, we designed the following procedure.

1. We estimated a matrix of the isonymic relationships (H) among the 363 considered parishes using the standardised coefficient H_{ij} proposed by Hedrick (1971), as implemented in the R package *Biodem* (Boattini and Calboli 2009).
2. The H matrix was transformed in a distance matrix (*d*) according to the formula $d_{ij} = 1 - H_{ij}$.
3. We reduced data complexity by applying a Non-Metric Multidimensional Scaling (NM-MDS) algorithm to the *d* matrix. We used the *R* function *isoMDS* from the *MASS* package (Venables and Ripley 2002; R Development Core Team 2020), which implements the Kruskal method. We explored configurations from $n = 10$ to $n = 20$ dimensions.
4. We then applied a non-hierarchical Model-Based Clustering algorithm, as implemented in the *Mclust* function (*mclust* R package: Fraley and Raftery 2002, 2006), to the calculated NM-MDS configurations. Mclust explores a set of ten different models for Expectation-Maximization (EM), each one of them for different numbers of clusters (K). The model and K that maximise the Bayesian Information Criterion (BIC) are retained. We ran Mclust for NM-MDS configurations from $n = 10$ to $n = 20$ dimensions. Then, we represented the relationship between *n* and *K* values with a scatterplot. As represented in Supplementary Figure 2, *K* shows a plateau-like pattern; accordingly, we selected the "best" *n* value as the one for which *K* reaches the plateau.
5. We re-ran Mclust with the most appropriate *n* and *K* values (as determined in step 4) and for each parish, we calculated its affiliation to one of the *K* inferred Clusters and the corresponding uncertainty of the classification. In order to get a synthetic representation of the identi-fied surname structure, we plotted the *K* Clusters on a geographic map using the R packages *PBSmapping*, *maps* and *mapdata* (Becker and Wilks 1993; Schnute et al. 2014).

Classification and evaluation of the dialectal structure

The linguistic description and analysis used in this work draws on methods developed for dialectometry (DM), whose tenets are derived both from numerical taxonomy and linguistic geography. Individual qualitative data are explored and inductively grouped based on the recognition of higher-ranking structural patterns. In the present case, objects to be classified consist of local dialects (*locolects*) represented in a linguistic atlas, while the term *characters* refers to individual maps. Objects are described in a binary data matrix which consists of *N* objects and *p* characters. Each character can present a different number of *coinages* and *qualities*. In the present case, the latter refers to *taxates*, i.e. individual basilectal characters reported in a map as divergent onomasiological, phonetic, or morphosyntactic types. The analysis of each original map aims to generate a data matrix based on the attribution of characters expressed at a nominal measurement scale. The matrix contains a vector for each of the *N* objects (= 57 inquiry points, Trentino dialects) containing *p* characters, where *p* equals the number of completed intralinguistic analyses (3094 working maps). The first step (Q-analysis) consists of ordinating objects by measuring pairwise similarity among them. The preferred similarity index in DM is the *Relative Identity Value* (RIV_{jk}), calculated as the relative number of shared characters between each pair of dialects (Supplementary Figure 6: Dialectometrical Flow Chart; see also Bauer 2009:88). The resulting squared similarity matrix (with dimensions *N* x *N*) is then used to generate groups via NM-MDS configurations from $n = 1$ to $n = 10$.

Comparisons among surnames, geography, dialects and DNA

1. Geography and surnames (363 points). We calculated three matrices of geographic distances between the 363 considered parishes using the R package *gdistance* (van Etten 2014): 1) great circle distances (distance measured along the surface of the earth); 2) least-cost distances (based on the presence of obstacles and accounting for a local friction index in the landscape); and 3) resistance distances (the average travel cost/commute time during a random walk from a starting point to a given destination and return). The relationship between these matrices and surname-based distances was quantitatively assessed using bias-corrected distance correlation (Székely and Rizzo 2013) as implemented in the function *dcor.ttest* in the package *energy* in R (Rizzo and Székely 2016).
2. Surnames, dialects and geography (57 points). Distance matrices based on surnames, dialects and geography (great circle distances) were compared with the above-mentioned bias-corrected distance correlation and with partial distance correlation using the function *pdcor.test* in the package *energy*. A consensus representation of the considered three matrices was obtained with the Distatis procedure (Beaton et al. 2019). Finally, affiliation to dialect-based Clusters is compared with those obtained from surnames with a chi-square test.
3. DNA, surnames and dialects (10 points). We obtained five distance matrices from the considered molecular datasets, e.g. Y-HG, Y-STR, mtDNA-SNPs, mtDNA-HVR, aut-STR. For consistency, the same distance measure was used for all datasets, i.e. Reynolds distance (Reynolds et al. 1983) as implemented in the *adegenet* R package (Jombart 2008). In addition, in the Y-STRs dataset DYS385a/b loci were excluded from calculations and DYS398I was subtracted from DYS398II. Finally, surname- and dialect-based distance matrices were extracted from the full matrices in order to match the ten genetic sampling locations (see Supplementary Table 4 and Supplementary Text for details). Comparisons among the above matrices were again performed with bias-cor-rected distance correlation and partial distance correlation.

Results

Surname diversity and surname-based classification

Within-population surname diversity (as measured by Shannon's index) ranges between 1.50 (Agrone, 281), and 6.63 (Trento – SS. Pietro e Paolo, 446), while the mean value is 3.38 and the median is 3.25 (for details see Supplementary Table 2). The geographic distribution of surname diversity in Trentino is represented in Figure 1. Areas with low surname diversity are mostly found in Giudicarie, Sole, Primiero and Cembra valleys, while relatively high values are observed along the most important communication routes (Adige, Lagarina, Garda and Sugana valleys) with peaks in the most important cities (Trento, Rovereto).

As for the surname-based classification, we preliminarily fitted our surname distance matrix to a given number of dimensions (n) by means of an NM-MDS algorithm, with $10 \leq n \leq 20$. The corresponding stress values decrease from 12.45% ($n = 10$) to 7.46% ($n = 20$) (Supplementary Figure 2). All these values are lower than the 30.5% threshold established by Sturrock and Rocha (2000) for 100 objects and 3 dimensions. After applying Mclust to the generated NM-MDS configurations, we observe that the number of inferred Clusters reaches its plateau ($K = 12$) at $n = 15$ (Supplementary Figure 2). Accordingly, we retained the configuration with $n = 15$ and $K = 12$ as the most appropriate.

As shown by the map reported in Figure 1, the 12 retained Clusters are clearly associated with geography, and in particular seem to correspond to the most important Trentino valleys, as summarised by the 16 "Comunità di Valle" (Supplementary Figures 1 and 3, Supplementary text). The significant association between surname-based distances and geographic distances (great circle and least coast distan-ces) is further demonstrated by bias-corrected distance corre-lations (Table 1). An interesting exception to this pattern is introduced by Cluster 3, which pools together the Non-Valley (north-west) with the geographically and linguistically dis-jointed Fassa Valley (north-east), home of the Ladin-speaking ethnic-linguistic minority. As for the Germanic-speaking eth-nic-linguistic minorities, both Mocheni and Cimbri are com-prised in the Alta Valsugana-specific Cluster 9. This fact suggests that their position in the surname space seems more related to geography than to their language. Instead, Cluster 12 coincides with the city of Trento, the most important urban centre of the region.

Segregation among the twelve Clusters, however, should not be overstated. In fact, in most cases, there are no clear demarcation lines between them, which on the contrary show some partial overlapping. This fact is particularly evident for Clusters involving the most important communication routes, such as Adige and Sugana Valleys (Clusters 1, 2, 4 and 9). As expected, when parishes with membership uncertainty (mu) higher than 0.01 are removed from the map, segregation becomes more clear-cut (Supplementary Figure 4). Therefore, we speculate that the percentage of parishes with mu > 0.01 is a measure of the "openness" of a Cluster. Our results (Supplementary Table 5) suggest that the most open Clusters are 1 (~Vallagarina) and 2 (~Cembra), while the most isolated are 11 (~Fiemme) and, surprisingly, 12 (~City of Trento).

Dialect-based classification

Following the procedure previously described, we fitted the dialect-based distance matrix to a given number of dimensions ($1 \leq n \leq 10$) with an NM-MDS algorithm; the corresponding stress values are lower than the 28.2% threshold established by Sturrock and Rocha (2000) for 57 objects and 3 dimensions. After applying Mclust to the NM-MDS configurations we retained the one with $n = 5$, corresponding to

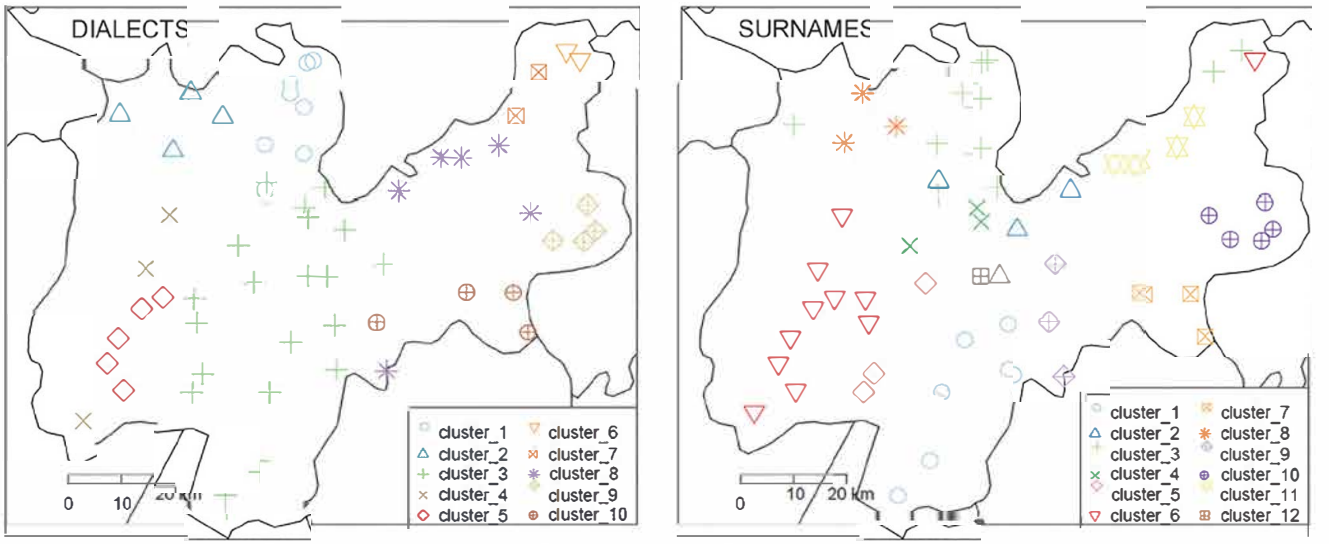

Figure 2. Distribution of dialect based (left) and surname based (right) Clusters in 57 Trentino parishes/dialects.

$K=10$ Clusters (Supplementary Figure 2). Similarly to surnames, the distribution of dialect Clusters is clearly related to geography (Figure 2b), as further assessed by means of a bias-corrected distance correlation comparing geographic (great-circle) and dialects-based distances ($p < .001$). The main feature of the dialect-based structure is a wide Cluster that occupies a great part of Central Trentino (Cluster 3). As far as ethnolinguistic minorities are concerned, Ladin groups are clearly identified (Clusters 6 and 7), while Germanic ones are included in other groups (Mocheni in Cluster 3 and Cimbrians in Cluster 8). This was anticipated since the present data refer to their Romance mesolects. The geographic distribution of dialect-based Clusters is fully described in the Supplementary Text.

Comparison between surnames and dialects

We tested the association between dialects and surnames in 57 parishes/dialects (Figure 2) with different tools. A chi-squared test shows that surname- and dialect-based Clusters are significantly associated ($p = 1.11 \times 10^{-14}$) and bias-corrected distance correlation confirmed the positive and significant relationship between surnames and dialects ($p < .001$, Table 1). Both datasets, however, exhibit significant correlations with geography (Table 1). In order to check if the relationship between surnames and dialects is mediated by geography, we performed a partial distance correlation, which confirms (Table 1) that dialects and surnames are significantly correlated ($p = .001$, $p = .02$ with Bonferroni correction).

Finally, Distatis was used to inspect the relationship between surnames, geography and dialects in the considered 57 parishes. Results are represented with a scatter-plot, where each point corresponds to the consensus position for each parish within a bi-dimensional space (Figure 3). Besides confirming the role of geography in shaping both the surname and the dialect variability of Trentino, Distatis results clearly show that Ladin-speaking communities (Fassa Valley) are the most important outliers. Accordingly, the highest Distatis residuals (i.e. the total displacement between the consensus configuration and the composing datasets) were observed for Ladin parishes/dialects (Delba/Alba di Canazei e Penia, Ciampedel/Campitello di Fassa, Vich/Vigo di Fassa, Moena/Moena; Supplementary Figure 5). In addition, when looking at single datasets, the highest contributions to such residuals come from dialects and geography. High residuals were also found in Primiero (particularly in S. Martino di Castrozza/Fiera di Primiero, Transacqua/Transacqua, Canal S. Bovo/Canal S. Bovo, Caoria/Caoria), where the three datasets contribute almost equally to the total value. The case of Trento is of particular interest, exhibiting a high surname-related residual component.

Comparison with genetic data

We performed pairwise bias-corrected distance correlations between each of the five considered genetic datasets (Y-HGs, Y-STRs, mt-SNPs, mt-HV, AUT) and surnames, dialects, and geography (Table 1). While no significant correlation ($p < .05$) was observed, the relationship between dialects and Y-STRs is clearly stronger than all other cases, including those between surnames and Y-chromosome markers.

Discussion

In this study, we compared the geographic structure of surnames, dialects and molecular markers in a well-documented and relatively heterogeneous context, Trentino. Our research relies on the exceptional availability of rich datasets for this Italian region.

Our first aim was to investigate the surname structure of Trentino using a wide baptismal record dataset, which covers almost all of the Trentino parishes for the period 1897–1923. As expected, our results clearly show that the distribution of surnames in Trentino is related to geography, thus suggesting that isolation-by-distance patterns had a primary role in shaping the surname structure of Trentino. At the same time, surname diversity was higher in parishes located along the

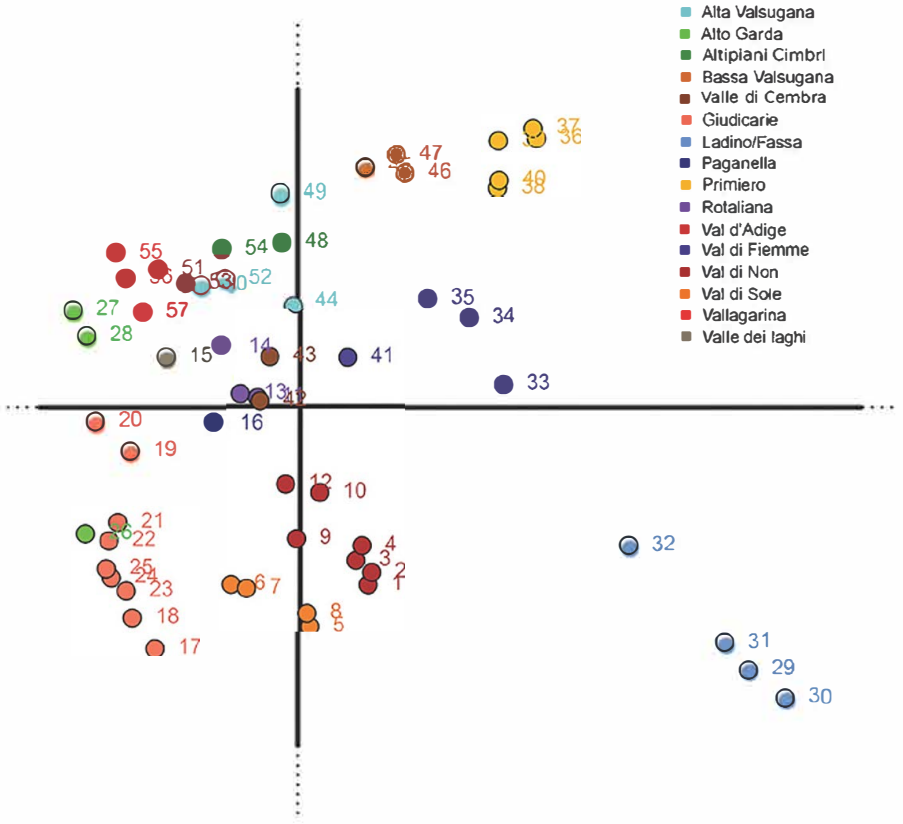

Figure 3. Distatis consensus representation of surname , dialect and geography based distances in 57 Trentino parishes/dialects.

most important valleys (Lagarina, Adige, Sugana), which in turn coincide with the principal communication routes (Figure 1). Of course, these findings reflect well-known pat-terns of distribution of human genetic variability, both at a local and at a global scale (Slatkin 1993; Ramachandran et al. 2005; Novembre et al. 2008; Salmela et al. 2008; Capocasa et al. 2014).

Accordingly, surname-based Clusters of parishes are clearly associated with the geographic distribution of the main valleys of Trentino (Figures 1 and 2, Table 1, Supplementary Text, Supplementary Figures 3 and 4). Nevertheless, interesting patterns appear when considering specific valleys. For instance, we expected the Adige Valley (and in particular the city of Trento) to be markedly homogeneous, considering its high population size and its role as an important communication route. Instead, our results revealed the presence of a remarkable sub-structure, with the simultaneous presence of a number of Clusters. Interestingly, one of them (12) is tightly associated with the city of Trento and exhibits no parishes with membership uncertainty (mu) higher than 0.01 (Supplementary Table 5). At the same time, the neighbourhood of the capital city appears as an "admixture belt", given that different Clusters (1, 2, 4 and 9) converge and partially intermingle with each other. These apparently contrasting results may be easily explained if we consider that Trento – by far the most important city of Trentino – hosted a wide range of professions and social strata, as it is typical of an urban environment. On the contrary, the other centres of the Adige Valley were mainly settled in a rural environment. In conclusion, the apparent segregation of Trento from the rest of the Adige Valley may be interpreted as the result of the peculiar status and social complexity of the city. It is worth mentioning that socioeconomic stratification at a micro-geographic scale has already been shown to generate surname and/or genetic differences in other Italian populations (Darlu et al. 2012; Boattini et al. 2015).

Importantly, this fact does not contradict the postulated role of the Adige Valley as one of the most important communication routes of the region, which is in fact exemplified by the moderate differentiation and partial overlapping of different Clusters around the city of Trento.

As far as ethnolinguistic minorities are concerned, our results show that none of the three considered groups (Ladins, Cimbrians, Mocheni) are represented by a specific Cluster (Figure 1, Supplementary Figure 4). Both Germanic-speaking groups (Cimbrians and Mocheni) are in fact affiliated to Cluster 9, which occupies most of the neighbouring Alta Valsugana. We conclude that surname-wise, the Germanic-speaking groups of Trentino are not significantly different from the Romance-speaking nearest villages.

This result may be interpreted according to the following, non-mutually exclusive hypotheses: a) significant gene flow between Mocheni, Cimbrians and Romance-speakers as well as language shift may have deleted possible between-group differences; b) Mocheni and Cimbrians may have adopted Romance-sounding surnames (or vice-versa). The case of Ladins from Fassa Valley is the most peculiar one. In fact, Ladins do not group either in a Ladin-specific Cluster or with the geographically closer clusters (with the exception of 222-

Moena, Cluster 11). Instead, they mostly group with parishes from the Non-Valley (Cluster 2), which is geographically separated from Fassa. A tentative explanation may be related to the fact that *Nonès*, the dialect spoken in Non-Valley, according to some scholars, is characterised by a Ladin substratum (Mastrelli Anzilotti 1997), which in the past may have generated similar or even identical surnames. However, it must be mentioned that such opinion is not confirmed by dialectometric analysis (see Bauer 2012, maps 3–5).

Our second aim was to explore the dialect structure of Trentino. Similar to surnames, Clusters based on 57 dialectal varieties are clearly related to geography and in particular to the most important valleys of Trentino (Figure 2), as confirmed by bias-corrected distance correlation tests (Table 1). Compared to surnames, dialects revealed some interesting features. First, a large part of central Trentino (including the city of Trento and the Valleys of the Adige, Lagarina and Laghi) is affiliated to a single, wide Cluster (3), thus suggesting a substantial dialectal homogeneity in this area, whereas surnames showed a considerable degree of segregation. Second, Ladin-speaking communities of the Fassa Valley are clearly grouped in two exclusive clusters. This was anticipated since local Ladin and Italian dialects are classified under different branches of the Romance family (Rhaeto-Romance and Italo-Romance/Gallo-Italic, respectively). Hence, dialects do not show the connection between Fassa and Non-Valleys which emerged for surnames. We cannot exclude that these differences may be at least in part influenced by the temporal discrepancy between dialectal data, which were collected recently, and surname data, which refer to more than one century ago.

Finally, Germanic-speaking groups cluster with other Romance dialects (Mocheni with Cluster 3 and Cimbrians with Cluster 8), but this was also anticipated considering that Germanic speakers are represented here by their Romance mesolects (i.e. the dialect they use to communicate with their Romance neighbours) and not by their own Germanic language.

We then measured the correlation between surnames and dialects by taking into account the influence of geography and using different methods (chi-square test, bias-corrected distance correlation and Distatis). Our results (Table 1, Figure 3, Supplementary Figure 5) show that a significant correlation between surnames and dialects does exist, even after controlling for geography. In the Netherlands the same correlation is instead completely independent from geographic distances (Manni et al. 2005, 2008), suggesting that the orographic differences between Italy and Netherlands, as well as the different temporal depth of their surname systems, may affect dialect and surname distributions.

Distatis (Figure 3, Supplementary Figure 5) confirmed that Ladin-speaking groups are the most prominent outliers due to high contributions of dialects and geography to their residuals.

We finally compared the variability of molecular markers (Y-chromosome, mtDNA, autosomal DNA) with the distribution of surnames and dialects in ten populations of Trentino. Surname studies often assume that they approximate Y-chromosome genetic markers (Cavalli-Sforza et al. 2004; King and Jobling 2009b). King and Jobling (2009a) observed that the probability of finding the same Y-chromosomal hap-logroup/haplotype is higher for English individuals bearing the same surname and such probability increases with the rarity of the surname. Similar results were obtained for Catalunia (Solé-Morata et al. 2015) and Ireland (McEvoy and Bradley 2006). These studies, however, did not consider the spatial structure and the distribution of surnames within the same geographic area. The correlation between surname structures and molecular markers was firstly tested in Sicily and Southern Italy (Boattini et al. 2018). Results of that study showed that unexpectedly "surnames are not good predic-tors of Y-chromosomal genetic structures", at least at the regional scale explored in that study. Instead, a significant correlation with haplotype-based autosomal data emerged, particularly when considering the longest – hence most recent – class of identical-by-descent shared tracts (>5 cM). Accordingly, the authors concluded: "the observed significant correlation is independent of the transmission modalities of autosomal markers and surnames, while seemingly related to their specific time depth".

Our results (Table 1) confirm the lack of association between surname structures and patterns of molecular markers, Y-chromosome included, even at a local geographic scale. Of course, we cannot exclude that a higher number of DNA-sampled populations would yield higher and more significant correlations. Similarly, more ancient surname data than those used here would probably help to reduce the confounding effect due to internal migration at the turn of the century. However, in this specific context, our results may genuinely discourage using surnames as reliable proxies for Y-chromosomal markers such as SNPs and STRs both at a regional and at a local scale. This does not mean that surnames are deprived of any biological meaning; rather they focus on different time scales – very recent for surnames, more ancient for molecular markers – making their study complementary to population genetics research and of great help to design sampling campaigns (Boattini et al. 2012).

As far as the gene-language co-inheritance hypothesis is concerned, we detected no significant relationship between molecular markers and dialects. However, the strongest effect size is observed between the distribution of dialects and that of Y-chromosomal STRs, which are fast-evolving markers specifically associated with the paternal line. In other words, our results hint at a possible relationship between patrilocality (Y-STRs) and cultural transmission (dialects). At the same time, patterns of diversity and cultural links between non-contiguous areas could also be interpreted in the light of higher male mobility and the presence of other socioeconomic barriers. These hypotheses, however, would require a higher number of sampling points for being properly tested. In this scenario, surnames seem to act like words, i.e. as pieces of cultural information whose distribution is more often affected by geography, human interaction, cultural admixture, and linguistic shift than by paternal lineages alone, as suggested by significant and positive correlations independently obtained with dialects.

Conclusions

Our research, which relies on rich and detailed surname, dialect and genetic information about Trentino, for the first time clarifies the relationships between these variables at a local scale. As expected, geography has a great impact both on the surname and dialect structures of Trentino. In fact, clusters of dialects and surnames are associated with particular valleys ("Comunità di Valle") of Trentino. Some parishes/ dialects showed outlier-like behaviour within these geog-raphy-related structures. Most notably, from the surnames point of view, the Ladin-speaking Fassa valley showed some affinity with the geographically distant Non-Valley. This fact may suggest traces of an ancient Ladin continuum that involved extended portions of Northern Trentino. Ladin-speaking communities, being surrounded by a wide Italo-Romance-speaking area, are the most obvious outliers also from the dialectal point of view. The city of Trento, instead, revealed its socio-economic differentiation from the rest of Trentino by forming a peculiar and exclusive sur-name cluster.

As for molecular markers, our results show that surnames are not significantly related to any of them even at a local scale. The same holds also for dialects and molecular markers, but a strong positive association is detectable between dialects and Y-STRs suggesting a possible relationship between patrilocality and cultural transmission.

Acknowledgements

The authors would like to thank Katia Pizzini (of the Archivio Diocesano di Trento) and Chiara San Giuseppe (of the Provincia Autonoma di Trento) for their kind collaboration and for allowing us using the "Nati in Trentino" database.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

Ascolani A. 2010. Il contesto demografico trentino nello scorcio del lungo Ottocento. Atti Acc. Rov. Agiati Ser. VIIIA, X:9 38.

Bauer R. 2003. Sguardo dialettometrico su alcune zone di transizione dell'Italia nord orientale (lombardo vs. trentino vs. veneto). In: Bombi R, Fusco F, editors. Parallela X. Sguardi reciproci. Vicende linguistiche e culturali dell'area italofona e germanofona. Udine: Forum Editrice; p. 93 119.

Bauer R. 2009. Dialektometrische Einsichten. Sprachklassifikatorische Oberflachenmuster und Tiefenstrukturen im lombardo venedischen Dialektraum und in der Ratoromania. San Martin de Tor: Istitut Ladin Micurá de Ru.

Bauer R. 2012. Zur inneren Arealgliederung des Trentino. Eine dialekto metrische Nachschau. In: Kohler C, Tosques F (eds.), Das diskrete Tatenbuch. Digitale Festschrift fur Dieter Kattenbusch zum 60. Geburtstag, Berlin: Humboldt Universitat, Institut fur Romanistik, (CD ROM); http://www.festschrift kattenbusch.de/bauer arealgliederung trentino.html

Bauer R. 2014. Zur Dialektometrisierung des ALD (I und II): Ein Arbeits und Erfahrungsbericht 2000 2012. In: Tosques F., editor. 20 Jahre dig itale Sprachgeographie. Berlin: Humboldt Universitat, Institut fur Romanistik; p. 95 120.

Bauer R. 2016. Analisi qualitativa e classificazione quantitativa dei dialetti altoitaliani e ladini/retoromanzi: dalla fonetica al lessico. In: Vicario F, editor. Ad Limina Alpium. VI Colloquium retoromanistich. Udine: Società Filologica Friulana; p. 11 38.

Bauer R, Casalicchio J. 2017. Morphologie und Syntax im Projekt ALD DM. Ladinia. XLI:81 108.

Beaton D, Fatt CC, Abdi H. 2019. DistatisR: DiSTATIS Three Way Metric Multidimensional Scaling. R package version 1.0.1. https://CRAN.R pro ject.org/package=DistatisR

Becker RA, Wilks AR. 1993. Maps in S, AT&T Bell Laboratories Statistics Research Report [93.2]. http://public.research.att.com/areas/ stat/doc/ 93.2.ps

Blanco Villegas MJ, Boattini A, Rodriguez Otero H, Pettener D. 2004. Inbreeding patterns in La Cabrera, Spain: dispensations, multiple con sanguinity analysis, and isonymy. Hum Biol. 76(2):191 210.

Boattini A, Blanco Villegas MJ, Pettener D. 2007. Genetic structure of la Cabrera, Spain, from surnames and migration matrices. Hum Biol. 79: 649 666.

Boattini A, Calboli FCF. 2009. Biodem: biodemography functions. R pack age version 0.2 (Cassola VC and Maechler M authored the function mtx.exp). http://CRAN.R project.org/package=Biodem

Boattini A, Calboli F, Blanco Villegas MJ, Gueresi P, Franceschi M, Paoli G, Cavicchi S, Pettener D. 2006. Migration matrices and surnames in populations with different isolation patterns: Val di Lima (Italian Apennines), Val di Sole (Italian Alps), and La Cabrera (Spain)). Am J Hum Biol. 18(5):676 690.

Boattini A, Lisa A, Fiorani O, Zei G, Pettener D, Manni F. 2012. General method to unravel ancient population structures through surnames, final validation on Italian data. Hum Biol. 84(3):235 270.

Boattini A, Pedrosi ME, Luiselli D, Pettener D. 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(4):604 609.

Boattini A, Sarno S, Fiorani O, Lisa A, Luiselli D, Pettener D. 2018. Ripples on the surface. Surnames and genes in Sicily and Southern Italy. Ann Hum Biol. 45(1):57 65.

Boattini A, Sarno S, Pedrini P, Medoro C, Carta M, Tucci S, Ferri G, et al. 2015. Traces of medieval migrations in a socially stratified population from Northern Italy. Evidence from uniparental markers and deep rooted pedigrees. Heredity. 114(2):155 162.

Bortolini E, Pagani L, Crema ER, Sarno S, Barbieri C, Boattini A, Sazzini M, et al. 2017. Inferring patterns of folktale diffusion using genomic data. Proc Natl Acad Sci U S A. 114(34):9140 9145.

Boyd R, Richerson PJ. 1985. Culture and the evolutionary process. Chicago: University of Chicago Press.

Capocasa M, Anagnostou P, Bachis V, Battaggia C, Bertoncini S, Biondi G, Boattini A, et al. 2014. Linguistic, geographic and genetic isolation: a collaborative study of Italian populations. J Anthropol Sci. 92:201 231.

Cavalli Sforza LL, Feldman MW. 1981. Cultural transmission and evolu tion: a quantitative approach. Princeton: Princeton University Press.

Cavalli Sforza LL, Moroni A, Zei G. 2004. Consanguinity, inbreeding, and genetic drift in Italy. Princeton, Oxford: Princeton University Press. Cavalli Sforza LL, Piazza A, Menozzi P, Mountain J. 1988. Reconstruction of human evolution: bringing together genetic, archaeological, and linguistic data. PNAS. 85(16):6002 6006.

Claerhout S, Roelens J, Van der Haegen M, Verstraete P, Larmuseau MHD, Decorte R. 2020. Ysurnames? The patrilineal Y chromosome and surname correlation for DNA kinship research. Forensic Sci Int Genet. 44:102204.

Coia V, Boschi I, Trombetta F, Cavulli F, Montinaro F, Destro Bisol G, Grimaldi S, et al. 2012. Evidence of high genetic variation among lin guistically diverse populations on a micro geographic scale: a case study of the Italian Alps. J Hum Genet. 57(4):254 260.

Coia V, Capocasa M, Anagnostou P, Pascali V, Scarnicci F, Boschi I, Battaggia C, et al. 2013. Demographic histories, isolation and social factors as determinants of the genetic structure of alpine linguistic groups. PLOS One. 8(12):e81704 2013.

Darlu P, Bloothooft G, Boattini A, Brouwer L, Brouwer M, Brunet G, Chareille P, et al. 2012. The family name as socio cultural feature and genetic metaphor: from concepts to methods. Hum Biol. 84(2): 169 214.

Fiorini S, Tagarelli G, Boattini A, Luiselli D, Piro A, Tagarelli A, Pettener D. 2007. Ethnicity and evolution of the biodemographic structure of Arbereshe and Italian populations of the Pollino area, Southern Italy (1820 1984). Am Anthropologist. 109(4):735 746.

Fraley C, Raftery AE. 2002. Model based clustering, discriminant analysis and density estimation. J Am Statist Assoc. 97(458):611 631.

Fraley C, Raftery AE. 2006. MCLUST Version 3 for R: normal mixture mod eling and model based clustering. Technical Report No. 504, Department of Statistics, University of Washington (revised 2009). Goebl H, editor. 2012. Sprachatlas des Dolomitenladinischen und angren zender Dialekte II/Atlant linguistic dl ladin dolomitich y di dialec vejins II/Atlante linguistico del ladino dolomitico e dei dialetti limitrofi II. Vol. 7. Strasbourg: Editions de Linguistique et de Philologie.

Goebl H, Bauer R, Haimerl E. 1998. Sprachatlas des Dolomitenladinischen und angrenzender Dialekte I/Atlant linguistic dl ladin dolomitich y di dialec vejins I/Atlante linguistico del ladino dolomitico e dei dialetti limitrofi I. Vol. 7. Wiesbaden: Reichert.

Gray RD, Atkinson QD. 2003. Language tree divergence times support the Anatolian theory of Indo European origin. Nature. 426(6965): 435 439.

Gray RD, Atkinson QD, Greenhill SJ. 2011. Language evolution and human history: what a difference a date makes. Philos Trans R Soc Lond B Biol Sci. 366(1567):1090 1100.

Gueresi P, Martuzzi VF, Pettener D. 2000. Biodemography of populations in the Eastern Alps of Trentino region. Rivista di Antropologia. 78: 169 178.

Gueresi P, Pettener D, Veronesi FM. 2001. Marriage behaviour in the Alpine Non Valley from 1825 to 1923. Ann Hum Biol. 28(2):157 171.

Hedrick PW. 1971. A new approach to measuring genetic similarity. Evolution. 25(2):276 280.

Jombart T. 2008. Adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics. 24(11):1403 1405.

Jordan P, Gibbs K, Hommel P, Piezonka H, Silva F, Steele J. 2016. Modelling the diffusion of pottery technologies across Afro Eurasia: emerging insights and future research. Antiquity. 90(351):590 603. King TE, Jobling MA. 2009a. Founders, drift, and infidelity: the relation ship between Y chromosome diversity and patrilineal surnames. Mol Biol Evol. 26(5):1093 1102.

King TE, Jobling MA. 2009b. What's in a name? Y chromosomes, sur names and the genetic genealogy revolution. Trends Genet. 25(8): 351 360.

Larmuseau MHD, Claerhout S, Gruyters L, Nivelle K, Vandenbosch M, Peeters A, van den Berg P, et al. 2017. Genetic genealogy approach reveals low rate of extrapair paternity in historical Dutch populations. Am J Hum Biol. 29(6):e23046.

Larmuseau MHD, van den Berg P, Claerhout S, Calafell F, Boattini A, Gruyters L, Vandenbosch M, et al. 2019. A historical genetic recon struction of human extra pair paternity. Curr Biol. 29(23):4102 4107.

Lewontin RC. 1972. The apportionment of human diversity. Evol Biol. 6: 381 398.

Manni F, Heeringa W, Toupance B, Nerbonne J. 2008. Do surname differ ences mirror dialect variation? Hum Biol. 80(1):41 64.

Manni F, Toupance B, Sabbagh A, Heyer E. 2005. New method for sur name studies of ancient patrilineal population structures, and possible application to improvement of Y chromosome sampling. Am J Phys Anthropol. 126(2):214 228.

Martinez Cadenas C, Blanco Verea A, Hernando B, Busby GB, Brion M, Carracedo A, Salas A, et al. 2016. The relationship between surname frequency and Y chromosome variation in Spain. Eur J Hum Genet. 24(1):120 128.

Mastrelli Anzilotti G. 1997. I caratteri di tipo ladino nei dialetti dell'Alta Val di Non. Mondo Ladi. 21:491 501.

McEvoy B, Bradley DG. 2006. Y chromosomes and the extent of patrilin eal ancestry in Irish surnames. Hum Genet. 119(1 2):212 219. Montinaro F, Boschi I, Trombetta F, Merigioli S, Anagnostou P, Battaggia C, Capocasa M, et al. 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(6):827 833.

Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, Indap A, et al. 2008. Genes mirror geography within Europe. Nature. 456(7218): 98 101.

Olmedo OE. 2011. Kriging: ordinary Kriging. R package version 1.0.1. http://CRAN.R project.org/package=kriging

Pettener D. 1990. Temporal trends in marital structure and isonymy in S. Paolo Albanese. Italy. Hum Biol. 62:837 851.

Pettener D, Gueresi P, Martuzzi VF. 1994. Struttura biodemografica della valle del Fersina (Valle dei Mocheni) dal 1800 al 1914. Bollettino di Demografia Storica. 20:131 140.

R Core Team. 2020. R: a language and environment for statistical com puting. Vienna (Austria): R Foundation for Statistical Computing. http://www.R project.org/

Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli Sforza LL. 2005. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proc Natl Acad Sci U S A. 102(44): 15942 15947.

Reynolds JB, Weir BS, Cockerham CC. 1983. Estimation of the coancestry coefficient: basis for a short term genetic distance. Genetics. 105(3): 767 779.

Rizzo ML, Székely GJ. 2016. Energy: E Statistics: multivariate inference via the energy of data (R Package), Version 1.7 0. https://CRAN.R project. org/package=energy

Salmela E, Lappalainen T, Fransson I, Andersen PM, Dahlman Wright K, Fiebig A, Sistonen P, et al. 2008. Genome wide analysis of single nucleotide polymorphisms uncovers population structure in Northern Europe. PLOS One. 3(10):e3519.

Schnute JT, Boers N, Haigh R, Grandin C, Johnson A, Wessel P, Antonio F. 2014. PBSmapping: mapping fisheries data and spatial analysis tools. R package version 2.67.60. http://CRAN.R project.org/package= PBSmapping

Slatkin M. 1993. Isolation by distance in equilibrium and non equilibrium populations. Evolution. 47(1):264 279.

Solé Morata N, Bertranpetit J, Comas D, Calafell F. 2015. Y chromosome diversity in Catalan surname samples: insights into surname origin and frequency. Eur J Hum Genet. 23(11):1549 1557.

Sturrock K, Rocha J. 2000. A multidimensional scaling stress evaluation table. Field Methods. 12(1):49 60.

Székely G, Rizzo M. 2013. The distance correlation t test of independence in high dimension. J Multivar Anal. 117:193 213.

van Etten J. 2014. gdistance: Distances and Routes on Geographical Grids (R Package), Version 1.1 5.

Venables WN, Ripley BD. 2002. Modern applied statistics with S. 4th ed. New York: Springer.