

Are ethnic minorities synonymous for genetic isolates? Comparing Walser and Romance populations in the Upper Lys Valley (Western Alps)

Alessio Boattini, Clio Griso & Davide Pettener

Dipartimento di Biologia Evoluzionistica Sperimentale, Area di Antropologia, Università di Bologna, Via Selmi 3, 40126 Bologna, Italy

e-mail: davide.pettener@unibo.it

Summary - Cultural differences between neighbouring populations are often said to give rise to reproductive barriers. For ethnic minorities, these barriers can easily result in genetic isolation. In this study, we analyse the surname structure of the Walser of the upper Lys Valley, a German-speaking ethnic minority in the Italian Western Alps, to better understand the relationships between linguistic and genetic isolation. Analyses were based on 1713 marriages registered from 1838 to 1938 in four villages of the valley: three Walser communities (Issime, Gressoney-Saint-Jean, Gressoney-La-Trinité) and the Romance community of Gaby. The results show that endogamy and inbreeding are lower than in other Italian linguistic minorities, with the exception of Gaby, whose values rank among the highest ever found in Italy. Compared to the Walser communities, Gaby behaves as an outgroup and has an almost exclusively autochthonous surname set. The latter aspect is also true, but to a lesser extent, for the Walser villages, in particular for Issime on the one hand and Gressoney-Saint-Jean and Gressoney-La-Trinité on the other. These findings strongly suggest that the Walser communities' ethnic minority status is not associated with genetic isolation, whereas genetic isolation was found in the linguistically non-isolated Gaby. Finally, our results are consistent with two independent late medieval migration events at the origin of these Walser settlements.

Keywords - Linguistic isolates, Mating structure, Inbreeding, Surnames, Italian Alps.

Introduction

One of the most widely investigated topics in human genetics is the study of geographically and/or culturally marginal human populations (currently or in the recent past). These characteristics, generally combined with small population size, are summarised by the term 'isolate'. Albeit with some debate (Abbott, 2000; Eaves *et al.*, 2000; Jorde *et al.*, 2000; Shifman *et al.*, 2003), medical genetics research considers isolates to be "model" populations for the identification of genes for Mendelian diseases and for the mapping of complex (non-Mendelian) traits (Peltonen *et al.*, 2000; Arcos-Burgos & Muenke, 2002; Varilo & Peltonen, 2004). There is also some interest in studying isolates from an "archaeogenetic"

point of view (Shlush *et al.*, 2008; Boattini *et al.*, 2010a). Indeed, isolates are very often the place to find traces of ancient migration/peopling events; it is more difficult to find such traces in non-isolated communities, as their involvement in further historical-demographic events is more probable.

In addition to geography, cultural features, particularly language, represent the most obvious cause of isolation and are often invoked as a source of genetic isolation in ethnic-linguistic minorities. Strong clues in that direction were found for the Aromuns in the Balkans (Bosch *et al.*, 2006), while the Gagauzes, a Turkish-speaking population from Eastern Europe, are a controversial case: according to Nasidze *et al.* (2007), they show clear signs of admixture with their Indo-European-speaking neighbours, but

these findings were partly contradicted by Varzari *et al.* (2009), who detected Balkan ancestral traces in that population.

Italy is an ideal place to explore relationships between genetics and languages, as it hosts twelve officially acknowledged ethnic-linguistic minorities. These minorities have been the subject of a broad research program on the biodemographic and genetic structure of Italian isolates (Destro-Bisol *et al.*, 2008). The Arbereshe, an Albanian-speaking population in southern Italy, is one of the best studied of these isolates: biodemographic and Y-chromosome data showed a close correspondence between cultural and genetic barriers (Fiorini *et al.*, 2007; Boattini *et al.*, 2011). Similar results were obtained for the German-speaking Mocheni (Western Alps, Italy) (Pettener *et al.*, 1994) and for Carloforte, a Ligurian island in southern Sardinia (Vona *et al.*, 1996). In contrast, a mtDNA study of the Croat minorities in Molise (southern Italy) showed that this population presents features lying between Croats and Italians, concluding that “ethnic consciousness has not created reproductive barriers” (Babalini *et al.*, 2005).

The present study aims to provide further knowledge of Italian isolates within the framework described by Destro-Bisol *et al.* (2008) and, in particular, to investigate the correspondence between ethnic minority status and genetic isolation. As the case-study, we used the Walser people of the Lys Valley, a previously unexplored German-speaking community in Aosta Valley, Italian north-western Alps.

Materials and Methods

The population

The Walser are a German-speaking people living in the Alps of Switzerland, Italy, Liechtenstein and Austria. The Walser language is a group of dialects included in the Highest Alemannic, a branch of Germanic which, besides the Walser people, is mainly spoken in Switzerland. The Walser are believed to have their origins in south-western Germany, from

ancient Alemannic tribes. During the 9th century, they settled in the upper part of the Wallis canton (Oberwallis, south-western Switzerland), from which their name derives. The 13th century was characterised by various migrations of German-speaking people towards northern Italy (e.g. Walser, Mocheni, Cimbri, Saurans). The Walser took part in those movements and settled in the Aosta Valley and Piedmont, where they gave rise to linguistic minorities which were isolated within the Romance regions (Zinsli, 1968; Sibilla, 1980).

The Walser of the upper Lys Valley (Aosta Valley), one of the most important and best conserved German-speaking minorities in Italy, currently live in three communities: Issime, Gressoney-Saint-Jean (hereafter GSJ) and Gressoney-La-Trinité (hereafter GLT). As is typical for Walser settlements, they consist mainly of scattered buildings, usually coinciding with farms; other shared characteristics are the high altitude and geographic isolation. Interestingly, the Lys Valley Walser settlements partly superseded autochthonous Romance-speaking populations, a remnant of which is the Franco-Provençal ‘enclave’ of Gaby (Fig. 1). Furthermore, the Walser dialect spoken in the Lys Valley is not homogeneous, with Northern (GSJ and GLT) and Southern (Issime) variants (Cerutti, 1977). On the whole, these peculiar features make the upper Lys Valley an interesting case-study for the mechanisms of isolation in ethnic-linguistic minorities.

The data

This research is based on 1713 civil matrimonial records from the three Walser communities of Issime (409), GSJ (588) and GLT (148), as well as the Romance community of Gaby (546). The data cover all upper Lys Valley marriages from 1838 to 1938 (due to privacy policies, more recent data are not available). To explore temporal variations in the genetic structure of the valley, we grouped the data into four 25-year intervals (1838-1863, 1864-1888, 1889-1913, 1914-1938). GLT is represented by a particularly low number of marriages. Nevertheless, we

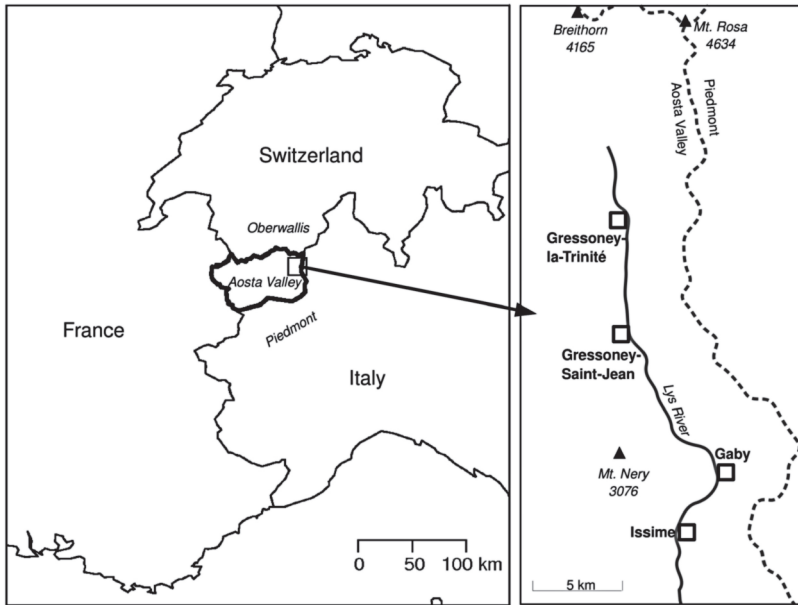


Fig. 1 - Geographic location and altitude of the Upper Lys Valley and of its four communities: Issime (956 m.), Gaby (1047 m.), Gressoney-Saint-Jean (1385 m.) and Gressoney-La-Trinité (1635 m.).

decided to maintain our research scheme also for this village. As a consequence, some caution is needed when considering the GLT results, in particular those which refer to 25-year intervals.

Methods

Endogamy rates were calculated in each community according to the birthplace of the spouses, with reference to a) the single communities (Village Endogamy), b) the whole Lys Valley, i.e. the four studied communities plus Fontainemore, Lillianes and Perloz, Romance villages from the lower part of the valley (Valley Endogamy).

Inbreeding coefficients (F_t , F_r , and F_n) were estimated according to the isonymy method (Crow & Mange, 1965; Crow, 1980); additionally, bootstrapped confidence intervals for F_t were calculated by means of 1000 random resamples of the data. Inbreeding measures were also assessed using the Repeated Pairs (RP) method (Lasker & Kaplan, 1985). The random expectation of repetition (RPr) was calculated with the equation proposed by Chakraborty (1985, 1986).

A distance matrix among the four communities was calculated according to Relethford's (1988) procedure, and it was used to perform a cluster analysis using Ward's algorithm. P-values were calculated via multiscale bootstrap resampling (1000 replications), a procedure that allows a better approximation to unbiased p-values when sample sizes are small (Shimodaira, 2004).

An analysis of the geographic distribution of surnames was performed using Self-Organizing Maps (SOMs). A full description of the method is provided by Manni *et al.* (2005), who first applied it in the case-study of the Netherlands, while its effectiveness even in microgeographic contexts was demonstrated by Boattini *et al.* (2010) and Rodriguez Diaz & Blanco Villegas (2010). Briefly, the SOMs method is a clustering technique through neural networks based on "competitive learning", an adaptive process in which the cells ("neurons") simulating a neural network ("map") gradually become sensitive to different input categories (Kohonen, 1982, 1984). The main idea is that different neurons

specialize to represent different types of input vectors. In doing so, they interact with the neighbouring neurons by means of a “neighbourhood function”. This procedure will result in the differentiation of the whole map-space: a) identical vectors will be mapped at the same neuron, b) slightly different ones at close neurons, while c) very different vectors will be mapped at far neurons. The shape (rectangular or squared) and size (number of cells) of the SOMs are defined by the user. The size of the map determines the maximum number of different clusters. Therefore, larger maps will classify items (surnames, in this study) more accurately than smaller ones. Nevertheless, it may happen that some cells remain empty, while others collect many items. Manni *et al.* (2005) suggested that the SOM method can be considered a “blind” automated approach to identify the geographic origin of surnames. Several trials were necessary before adopting a 3x3 lattice (9 possible clusters). In fact, this cell grid is characterised by the emergence of a few empty cells that are not found in smaller maps; in contrast, empty cells become more numerous in the maps with higher dimensions. In short, 9 different cells proved sufficient to represent the geographic variability of surnames in the upper Lys Valley. Since the method infers the geographic origin of surnames from their geographic distribution, we decided to exclude surnames whose frequency in the whole database was lower than or equal to five; given their rareness, these surnames do not allow the detection of reliable geographic gradients. As a result, SOMs were only applied to the 95 most frequent surnames of the upper Lys Valley.

Software

All calculations were performed using the R software (R Development Core Team, 2010) with the following libraries: Biodem (Boattini & Calboli, 2009) to calculate inbreeding and RP coefficients; pvclust (Suzuki & Shimodaira, 2009) to perform cluster analysis with p-values; kohonen (Weherens & Buydens, 2007) to perform SOMs analysis. Confidence intervals for inbreeding coefficients were calculated by custom R scripting.

Results

Endogamy

Both Village and Valley Endogamy (Tab. 1) show a clear descending trend in all the villages for the first three periods, while they tend to stabilize in the fourth period (1914-1938), with the exception of the very small village of GLT. For Village Endogamy, the most endogamous villages are Issime and Gaby, with very similar values (64.24% and 61.34%, respectively). For Valley Endogamy, the Romance village, Gaby, is distinct from the Walser villages not only for its particularly high value (86.30%) but also for its tendency to maintain high endogamy rates over time.

Inbreeding and Repeated Pairs (RP)

The highest total inbreeding values (Ft) are in Gaby, while the three Walser communities generally show lower values (the only exception being GLT in the second period) (Tab. 1). In agreement with random (Fr) and non-random (Fn) inbreeding components, bootstrapped Confidence Intervals (Tab. 1) show that Ft values are significantly higher than those expected (under random matings) (Fr) only in Gaby (1864-1888, 1889-1913, 1914-1938). In other words, Gaby shows a culturally driven tendency towards consanguineous marriages. The Repeated Pairs index (RP) confirms the isonymy results: the highest values (for the whole 1838-1938 period) are found in Gaby (0.0077), followed by the Walser communities: GLT (0.0053), Issime (0.0038), GSJ (0.0014). These values are similar to those expected under random mating (RPr) for GSJ (0.0011) and Issime (0.0038), while they are higher in Gaby (0.0058) and GLT (0.0036).

Cluster analysis

Surname distances among the four upper Lys Valley communities in the four periods were used to perform a cluster analysis (Fig. 2). The results show that: a) the highest probability values ($p > 95\%$) were found for the clusters grouping the same population for different periods, suggesting that the surname structure of the four

Tab. 1 - Population size, Village and Valley Endogamy rates, Inbreeding coefficients (Fr, Fn, Ft) with bootstrapped confidence intervals for the Upper Lys Valley populations.

COMMUNITY	PERIOD	POP. SIZE ¹	ENDO GAMY (%)		INBREEDING* 10 ⁻²			
			VILLAGE	VALLEY	FR	FN	FT	(CI 0.025, 0.0975)
Gressoney-Saint-Jean (GSJ)	1838-1863	882	70.80	86.73	1.280	0.270	1.550	(0.400, 2.457)
	1864-1888	909	54.68	64.75	0.760	0.290	1.050	(0.159, 1.581)
	1889-1913	949	38.89	46.91	0.460	0.160	0.620	(-0.005, 1.089)
	1914-1938	1010	41.42	49.70	0.540	-0.100	0.440	(-0.005, 1.194)
	total		51.45	62.02	0.760	0.160	0.910	(0.136, 1.580)
Gressoney-La-Trinité (GLT)	1838-1863	224	57.14	82.14	1.500	0.310	1.800	(-0.092, 3.675)
	1864-1888	214	40.54	70.27	1.810	2.420	4.190	(-0.085, 4.187)
	1889-1913	167	16.67	47.92	1.410	-0.950	0.470	(-0.049, 3.136)
	1914-1938	158	6.06	24.24	0.410	-0.420	-0.010	(-0.016, 1.487)
	total		30.10	56.14	1.280	0.340	1.610	(-0.060, 3.121)
Issime	1838-1863	790	88.10	94.44	1.580	0.000	1.590	(0.550, 2.835)
	1864-1888	869	71.58	87.37	1.510	-1.070	0.460	(0.454, 2.868)
	1889-1913	776	46.08	53.92	1.190	-0.230	0.970	(0.217, 2.483)
	1914-1938	730	51.22	57.32	1.150	1.010	2.140	(0.257, 2.455)
	total		64.24	73.26	1.360	-0.070	1.290	(0.370, 2.660)
Gaby	1838-1863	682	81.25	92.86	1.680	0.360	2.030	(0.609, 2.967)
	1864-1888	751	61.07	92.37	2.190	1.370	3.530	(1.041, 3.526)
	1889-1913	841	51.75	81.12	1.820	1.430	3.230	(0.817, 3.039)
	1914-1938	842	51.28	78.85	1.610	1.540	3.120	(0.620, 2.616)
	total		61.34	86.30	1.820	1.170	2.970	(0.772, 3.037)

¹ Italian ISTAT census data, years 1861, 1881, 1901, 1921

communities generally remained unchanged in time; b) the three Walser communities form a highly supported cluster ($p = 89\%$), while the Romance village Gaby behaves as an outgroup.

SOMs

The spatial organisation of the 3x3 SOM lattice is summarised in Table 2, while figure 3 gives a synthetic overview of the geographic distribution of SOM surname clusters in the four villages. The distribution within the map of the 95 most frequent surnames in the upper Lys Valley is very heterogeneous. For instance, cells from column two – coordinates: (1, 2), (2, 2), (3, 2) – are

nearly empty, while those from column one and three collect 94 of the 95 considered surnames. In particular, cells from column one group surnames that are mostly frequent in Gaby and Issime, while cells from column three group surnames from GSJ and GLT. More precisely, cells at the corners of the map collect the surnames with the highest frequencies in each of the four communities; hence their geographic origin can be attributed to the same populations: Gaby (1, 1), Issime (3, 1), GSJ (3, 3) and GLT (1, 3). Among them, Issime and GSJ show the highest rates of clusters of surnames that have their geographic origin in the same villages (70.37% and 72.99%,

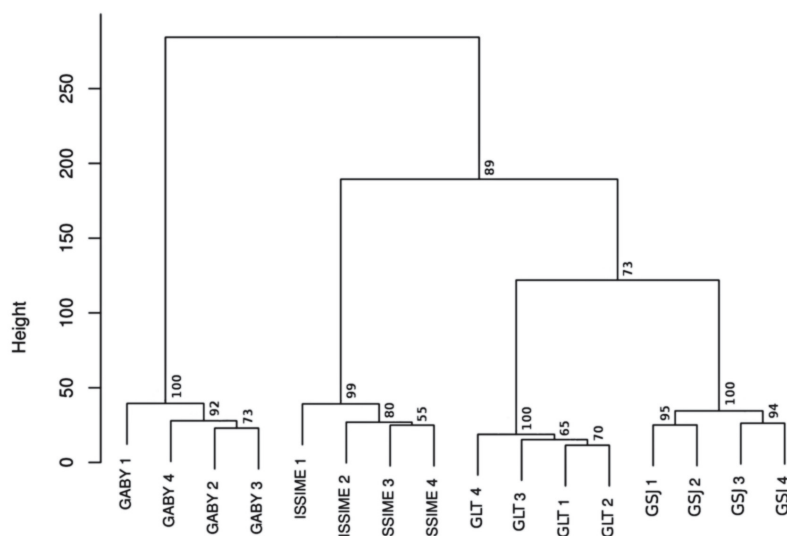


Fig. 2 - Cluster analysis from surname distances between the Upper Lys Valley communities in four historic periods (1: 1838-1863, 2: 1864-1888, 3: 1889-1913, 4: 1914-1938); approximately unbiased multiscale bootstrap resampling p-values (%) are reported.

respectively). Cell (2, 3) shows the broad sharing of surnames between GSJ and GLT, as expected given that these communities are strongly linked both linguistically and geographically.

Intriguingly, cell (2, 1) does the same – albeit with lower percentages – for the Walser Issime and the Romance Gaby. This interesting case is related to the fact that the municipality of Gaby included some small Walser settlements and consequently surnames (see Discussion). It must be noted that the upper Lys Valley generally lacks surnames with wide geographic diffusion (with the possible exception of the one surname in cell (3, 2)).

Discussion

There have been few studies of Italian ethnic-linguistic minorities. The best known case is that of the Arbereshe of the Pollino area (southern Italy). Analyses based on surname and Y-chromosome data showed that Arbereshe villages “conserved much of their ancestral genetic diversity along with their founder surnames and

cultural features” (Fiorini *et al.*, 2007; Boattini *et al.*, 2011). Cultural – and more specifically linguistic – barriers working also as genetic barriers have been identified for geographically and linguistically diverse populations such as the German-speaking Mocheni in the north-eastern Alps (Pettener *et al.*, 1994) and Carloforte, a Ligurian island in southern Sardinia (Vona *et al.*, 1996). However, a mtDNA study of Croat minorities of southern Italy showed significant admixture with their autochthonous neighbours (Babalini *et al.*, 2005), while Ladins from South Tyrol (northern Italy) exhibited contrasting behaviour between Y-chromosome and mtDNA data, the former characterised by reduced diversity within the Ladins and high differentiation with the neighbouring populations, the latter showing higher diversity and lineages shared with the rest of the region (Pichler *et al.*, 2006). Differential behaviour between patrilineal and matrilineal markers – with higher heterogeneity for mtDNA – has also been found for the Arbereshe of Pollino (Boattini, unpublished data). These observations suggest that we need to be particularly cautious

when comparing results from studies based on different markers and sampling strategies.

Unfortunately, endogamy and inbreeding values are available for only a few of these populations (Arbereshe, Mocheni). Therefore, we extend our comparisons to include other Italian populations which, although not ethnic minorities, are genetically isolated due to environmental factors. In recent years, several studies based on genealogical reconstructions have been conducted on isolated populations such as Talana (Sardinia) (Angius *et al.*, 2001; Mocci *et al.*, 2009), Val Borbera (Piedmont, northern Italy) (Traglia *et al.*, 2009), Carlantino (Puglia, southern Italy) and Campora (Campania, southern Italy) (Colonna *et al.*, 2007). All these populations showed village endogamy values between 80% (Campora) and 99.5% (Carlantino) for the 19th century (Tepper *et al.*, 2008). Since these values are based on marriage records, they can be directly compared with those for the Walser. Endogamy in our Walser populations, ranging from 30.10% to 64.24%, is substantially lower than in those isolates. It is also lower than the values found in other linguistic isolates, and in particular in the geographically and linguistically similar Mocheni, in which endogamy ranges around 70% (Pettener *et al.*, 1994). In Valsesia, another Walser community in Piedmont (Italy), values as high as 72.1% were found for the same historical period (1850-1899) (Malacarne *et al.*, 2005). Finally, the studied population shows a clear descending trend that is absent, or barely recognizable, in the other populations for the same historical period, suggesting that the Walser may have experienced a largely anticipated breakdown of isolates.

Inbreeding estimates from isonymy, despite being quite accurate in describing historical trends, may overestimate actual inbreeding values within a population: many studies have reported isonymy values 2-3 times higher than those from genealogies (Gagnon & Toupance, 2002). For the above-mentioned Italian isolates, genealogy-based studies showed inbreeding values between $0.6 \cdot 10^{-2}$ (Campora) and $1.8 \cdot 10^{-2}$ (Talana). Isonymy-based studies of the Pollino Arbereshe

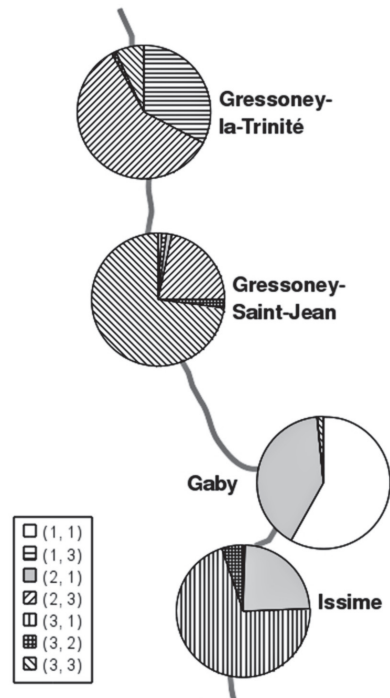


Fig. 3 - Frequencies of SOM surname clusters in the communities of the Upper Lys Valley. Villages are represented by pie charts, each one of them showing their surname composition according to SOM cells.

revealed mean Ft values around $1.73 \cdot 10^{-2}$ for the period 1820-1884 (Fiorini *et al.*, 2007), while in the majority of Mocheni communities they ranged from $1.393 \cdot 10^{-2}$ to $2.439 \cdot 10^{-2}$ (Pettener *et al.*, 1994). Similar values were found for the Walser of Valsesia ($1.458 \cdot 10^{-2}$, Malacarne *et al.*, 2005). Higher Ft values were recorded in geographic isolates from the Italian Alps such as the upper Sole and Non Valleys (north-eastern Alps) – ranging around $2.199 \cdot 10^{-2}$ (1775-1923) (Martuzzi Veronesi *et al.*, 1996; Gueresi *et al.*, 2001) – and Acceglio (north-western Alps) – $4.32 \cdot 10^{-2}$ (1889-1908) (North *et al.*, 1996) – as well as in the Val di Scalve (Italian central Pre-Alps), where Ft reached $3.806 \cdot 10^{-2}$ (1866-1935) (Boattini *et al.*, unpublished data). Therefore, if we consider that isonymy overestimates inbreeding compared to genealogies, we can conclude

that the Walser villages have modest kinship levels (Tab. 1). The RP index confirms these results, showing that the Walser communities (with the exception of GLT) are characterised by lower values than those found in ethnic minorities such as the Pollino Arbereshe (0.0056, 1870-1884; Fiorini *et al.*, 2007) and geographic isolates such as Acceglio (0.0050, 1889-1908; North & Crawford, 1996) and Val di Scalve (0.0123, 1866-1935, Boattini *et al.*, unpublished data).

The Romance village of Gaby behaves differently from its Walser neighbours: it shows not only the highest Valley Endogamy rate (86.30%), but also the highest inbreeding (F_t) value (2.974×10^{-2}). In particular, F_t values in Gaby are in the same range as those found in the most isolated Arbereshe (Fiorini *et al.*, 2007) and Mocheni villages (Pettener *et al.*, 1994), as well as in some geographically isolated upper Sole Valley parishes (Martuzzi Veronesi *et al.*, 1996). Furthermore, Gaby is the only upper Lys Valley village in which the observed F_t values are significantly higher than the expected ones (Tab. 1), meaning that marriages between relatives were deliberately favoured. Another – and not necessarily alternative – explanation is that the municipality of Gaby included some small Walser settlements, thus generating not only a linguistic boundary but also a barrier to random mating within the community itself. Unfortunately, our database does not include information on the mates' places of residence within the municipalities, and thus we could not directly test this hypothesis. Nevertheless, further evidence in this direction comes from the results of cluster analysis and especially SOMs.

Cluster analysis (Fig. 2) highlights the outlier position of Gaby in the Walser domain; the same representation suggests the strong similarity – nearly identical – between GSJ and GLT, and, more in general, among the Walser villages. SOMs analysis adds further evidence to these observations. SOMs cells collect clusters of surnames that share the same geographic distribution and origin (Manni *et al.*, 2005); thus spatial patterns of SOMs cells (fig. 3) can be interpreted with a phylogeographic approach. For instance,

the identity between GJS and GLT is underlined by the existence of a SOM cell (2, 3) that is shared between these villages, while a single, albeit common, surname, the sole member of cell (3, 2), represents the link between GJS, GLT and the southernmost Issime, the other Walser village. Nevertheless, and despite the low endogamy values, the four villages have maintained their surname identity over time. In fact, each of them has a 'private' surname cluster in the SOM map, e.g. Gaby (1, 1); Issime (3, 1); GSJ (3, 3); GLT (1, 3). In other words, these cells comprise the surnames that have their geographic origin in each of the considered villages. The spatial organization of the SOMs (Tab. 2) reflects the geography of the valley and, partly, its dialectal differentiation: surnames from the southern Issime and Gaby are located in the first column, while those from the northern GSJ and GLT are in the third column, and they are separated by an almost empty column (with the exception of the above-mentioned (3, 2)). One of the most interesting features of the SOMs is cell (2, 1): a cluster of surnames whose distribution overlaps the Walser Issime and the Romance Gaby. This is in apparent contradiction with the outlier status of Gaby suggested by cluster analysis. However, we must observe that cluster analysis uses populations/periods as units of analysis (by means of a distance matrix), while SOMs apply directly to surname frequencies. Thus, SOMs may reveal population structures that do not coincide with units of analysis. This is exactly the case of cell (2, 1), whose frequency peak is observed in Gaby (40.34%), although it is also very common in Issime (23.88%). In fact, the municipality of Gaby included some small Walser settlements (the most important was called Niel), and these settlements shared surnames with the neighbouring Issime. In contrast, virtually no 'private' Gaby surname (1, 1) is present in the Walser villages. Therefore, Gaby appears to be composed of two different subpopulations: a numerically dominant Romance community and a minor Walser one, each with different surname pools. This subdivision helps to explain the higher-than-expected F_t values observed in Gaby (Wahlund effect).

These results allow us to draw some conclusions about the relationships between linguistic and genetic isolation in the Walser of the upper Lys Valley. This population has conserved its cultural features and languages (Töitschu in Issime, Titsch in GSJ and GLT), as acknowledged by regional (law 47/1998) and national (law 482/1999) regulations. Thus, from a linguistic point of view, the Walser represent a clear discontinuity within the surrounding Franco-Provençal continuum. Nevertheless, their endogamy and inbreeding values are lower than those observed in other linguistic isolates such as the Arbereshe and Mochenì. As a consequence, although the last two populations can be considered as both linguistic and genetic isolates, this does not seem to be true for the Walser given they are a linguistic isolate but not a genetic isolate. Paradoxically, the Romance Gaby is the only upper Lys Valley population to show clear marks of genetic isolation. In other words, the linguistic barrier between Germans and Romances worked ‘against’ the autochthonous population.

These findings agree with information on Walser history and traditional society. Along with more usual activities such as cattle breeding, the Walser developed a strong talent for commercial trading. Walser merchants, especially those from the Gressoney area, were so well known on both sides of the Alps that the term “Grischeneyer” (i.e. “from Gressoney”) became synonymous of merchant. Many of them settled in the main cities of southern Germany and Switzerland, while still maintaining contacts with their homeland. Thus, the Walser of the upper Lys Valley were characterised by high population mobility, which fits well with the low endogamy and inbreeding values.

Finally, these data help shed light on the still obscure history of the Walser settlement of the valley. The populations of Issime and GSJ-GLT, besides being characterised by different variants of the Walser language, also show different surname sets and limited surname sharing (Fig. 3). This suggests that the Walser of the upper Lys Valley are the result of two independent late medieval migration events, each characterised by a different dialect and thus sets of surnames. The

Tab. 2 - Self-Organizing Map of the 95 most frequent surnames. The number of surnames per cell and their frequencies in each community are reported.

CELL	(1, 1)	(1, 2)	(1, 3)
<i>N. Surnames</i>	8	0	5
Gressoney-SJ	0.83%	-	1.18%
Gressoney-LT	0.00%	-	32.62%
Issime	0.51%	-	0.00%
Gaby	58.08%	-	0.00%

CELL	(2, 1)	(2, 2)	(2, 3)
<i>N. Surnames</i>	5	0	7
Gressoney-SJ	1.30%	-	21.56%
Gressoney-LT	0.00%	-	59.36%
Issime	23.88%	-	0.00%
Gaby	40.34%	-	0.00%

CELL	(3, 1)	(3, 2)	(3, 3)
<i>N. Surnames</i>	16	1	30
Gressoney-SJ	0.12%	2.01%	72.99%
Gressoney-LT	0.00%	1.07%	6.95%
Issime	70.37%	5.24%	0.00%
Gaby	0.11%	0.00%	1.47%

two Walser migrations settled, respectively, in Issime and Gressoney. The newborn communities literally encircled the pre-existing Romance village of Gaby (but conserved the original Romance toponyms that still characterise the whole valley). These events led to the rise of a linguistic and – as shown here – a ‘reverse’ genetic barrier that lasted to the 20th century.

In summary, the results of our biodemographic analysis have allowed us to make reliable inferences on the population structure of the upper Lys Valley. As such, they can be considered a useful preliminary step for molecular anthropology studies. Researchers agree that knowledge of the “underlying population structure” (Colonna *et al.*, 2007) is extremely useful for sampling strategies in population-based association studies and in phylogenetic studies (Shlush *et al.*, 2008; Helgason *et al.*, 2005). The most widely

used sampling method is the so-called “grandfather criterion”, in which individuals who share grandparents born in the study area are sampled (all four, if the targets are autosomes; the paternal grandfather or the maternal grandmother if the target is Y-chromosome or mtDNA). In this way, it is possible to quickly exclude the recent immigrants, but it is not possible to identify immigrants beyond the second generation and, above all, to make reliable inferences concerning the genetic structure of the population. To overcome these problems, extensive reconstructions of genealogies, with the identification of patrilineal and matrilineal genealogical lines, have been proposed and successfully used for some Italian villages. For example, in Talana (Sardinia), researchers were able to establish that more than 80% of the present-day population descends from eight paternal and eleven maternal ancestral lineages (Angius *et al.*, 2001), while in Campora (southern Italy) 69% of the living individuals belong to 46 and 70 maternal and paternal lineages whose presence in the village predates 1890 (Colonna *et al.*, 2007). Surname analysis, particularly as performed with the SOM method, can be of invaluable help. Given the shared patrilineal transmission of surnames and the Y-chromosome, the method is particularly suited for sampling in Y-chromosome DNA studies. In our case, SOM cells detected five main groups of surnames that can be considered founders of each of the studied villages, namely: 5 for GLT (cell 1, 3), 30 for GSJ (cell 3, 3), 16 for Issime (cell 3, 1), 8 for the Romance Gaby (cell 1, 1) and 5 for its Walser minor settlements (cell 2, 1). An ideal upper Lys Valley sample for Y-chromosome studies should be based on such surname groups, with two important precautions: a) the number of sampled individuals per community should be proportional to the number of ‘founder’ surnames; b) avoid the over-representation of some surnames. The surname-based sampling strategy allows one to maximize the number of different Y-chromosome lineages for a given sample size and to get as near as possible to the period in which surnames spread, excluding recent immigrants by default. It should be noted that the

proposed method is not designed to detect the current allele frequencies. In this case, as in population-based association studies, random sampling (and a higher sample size) inside each unit of analysis seems more useful.

Analyses based on molecular markers are needed to extend our study beyond the geographic limits of the upper Lys Valley and the temporal depth covered by surnames (4-5 centuries ago). In particular, research on German-speaking ethnic minorities of northern Italy should take into consideration both biodemographic results and molecular analyses in order to unravel their genetic history and to assess similarities and differences in their population structure.

In conclusion, the case of the Walser demonstrates that a population may well conserve its cultural and linguistic features while losing much of its biological ancestry. The Walser, due to their activities as merchants, were a markedly mobile population that, for centuries, maintained a large network of contacts on both sides of the Alps and favoured marital exchange with other populations. As a consequence, they were not as isolated as their ethnic minority status would suggest. Indeed, the linguistic barrier between the Walser and their Romance neighbours was not completely lacking in biological effects on the population: it managed to isolate the autochthonous village of Gaby. Further investigations based on molecular markers (in particular the Y-chromosome) are needed to unravel the genetic history of German-speaking communities of the Italian Alps.

Acknowledgements

We would like to thank the municipalities of all the investigated villages for their collaboration with archival data collection. Special thanks are due to the Chairman of the “Comunità Montana Walser -Alta Valle del Lys”, Mr. Christian Linty, and to the Director of the “Centre for Walser Culture”, Mrs. Irene Alby, who kindly supported our work. The study was partially supported by Italian MIUR-PRIN 2007 and by Strategic Project 2006-09 University of Bologna grants to DP.

References

- Abbott A. 2000. Manhattan versus Reykjavik. *Nature*, 406: 340-342.
- Angius A., Melis P.M., Morelli L., Petretto E., Casu G., Maestrale G.B., Fraumene C., Bebbere D., Forabosco P. & Pirastu M. 2001. Archival, demographic and genetic studies define a Sardinian sub-isolate as a suitable model for mapping complex traits. *Hum Genet.*, 109: 198-209.
- Arcos-Burgos M. & Muenke M. 2002. Genetics of population isolates. *Clin. Genet.*, 61: 233-247.
- Babalini C., Martinez-Labarga C., Helle-Viivi T., Kisivild T., Giampaolo R., Tarsi T., Contini I., Barac L., Janicijevic B., Martinovic Klaric I., Pericic M., Sujoldzic A., Vilems 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.
- Bartlett J.E., Kotrlik J.W. & Higgins C.C. 2001. Organizational Research: Determining Appropriate Sample Size in Survey Research. *Information Technology, Learning, and Performance Journal*, 19: 43-50.
- Boattini A. & Calboli FCF. 2009. *Biodem: Biodemography functions. R package version 0.2* (VC Cassola and M Maechler authored the function `mtx.exp`), <http://CRAN.R-project.org/package=Biodem>
- Boattini A., Luiselli D., Sazzini M., Useli A., Tagarelli G. & Pettener D. 2011. Linking Italy and the Balkans. A Y-chromosome perspective from the Arbereshe of Calabria. *Ann. Hum. Biol.*, 38: 59-68.
- Boattini A., Pedrosi M.E., 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: 604-609.
- Bosch E., Calafell F., Gonzalez-Neira A., Flaiz C., Mateu E., Scheil H.G., Huckenbeck W., Efremovska L., Mikerezi I., Xirrotiris 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.
- Cerutti A.V. 1977. *La storia del clima e delle genti del Monte Rosa*. In Augusta, Aosta.
- Chakraborty R. 1985. A note on the calculation of random RP and its sampling variance. *Hum. Biol.*, 57: 713-717 (Erratum, 1986, *Hum. Biol.*, 58: 991).
- Colonna V., Natile T., Astore M., Guardiola O., Antoniol G., Ciullo M. & Persico M.G. 2007. Campora: a young genetic isolate in South Italy. *Hum. Hered.*, 64:123-35.
- Crow J.F. & Mange A.P. 1965. Measurement of inbreeding from the frequency of marriages between persons of the same surname. *Eugen. Q.*, 12: 199-203.
- Crow J.F. 1980. The estimation of inbreeding from isonymy. *Hum. Biol.*, 52:1-14.
- Destro Bisol G., Anagnostou P., Batini C., Battaggia C., Bertoncini S., Boattini A., Caciagli L., Caló M.C., Capelli C., Capocasa M., Castri L., Ciani G., Coia V., Corrias L., Crivellaro F., Ghiani M.E., 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: 179-88.
- Eaves I.A., Merriman T.R., Barber R.A., Nutland S., Tuomilehto-Wolf E., Tuomilehto J., Cucca F. & Todd J.A. 2000. The genetically isolated populations of Finland and Sardinia may not be a panacea for linkage disequilibrium mapping of common disease genes. *Nat. Genet.*, 25: 320-323.
- 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. Anthropol.*, 109: 735-746.
- Gagnon A. & Toupance B. 2002. Testing isonymy with paternal and maternal lineages in the early Quebec population: the impact of polyphyly and demographic differentials. *Am. J. Phys. Anthropol.*, 117: 334-341.

- Helgason A., Yngvadóttir B., Hrafnkelsson B., Gulcher J. & Stefánsson K. 2005. An Icelandic example of the impact of population structure on association studies. *Nat. Genet.*, 37: 90-95.
- Gueresi P., Pettener D. & Martuzzi Veronesi F. 2001. Isolation and patterns in marriage behaviour in an Alpine valley (Non Valley, 1825-1923). *Ann. Hum. Biol.*, 28: 157-171.
- Jorde L.B., Watkins W.S., Kere J., Nyman D. & Eriksson A.W. 2000. Gene mapping in isolated populations: new roles for old friends? *Hum. Hered.*, 50: 57-65.
- Kohonen T. 1982. Self-organized formation of topologically correct feature maps. *Biol. Cybern.*, 43: 59-69.
- Kohonen T. 1984. *Self-organization and associative memory*. Springer, Berlin.
- Lasker G.W. & Kaplan B.A. 1985. Surnames and genetic structure: Repetition of the same pairs of names of married couples, a measure of subdivision of the population. *Hum. Biol.*, 57: 431-440.
- Malacarne E., Danubio M.E. & Gruppioni G. 2005. The effects of geographical and prolonged cultural isolation on the marital behaviour of an Alpine community (Valsesia-Italy, 1618-1899). *Hum. Evol.*, 20: 167-179.
- Manni F., Toupance B., Sabbagh A. & Heyer E. 2005. New method for surname studies of ancient patrilineal population structures, and possible application to improvement of Y-chromosome sampling. *Am. J. Phys. Anthropol.*, 126: 214-228.
- Martuzzi Veronesi F., Gueresi P. & Pettener D. 1996. Biodemographic analysis of Italian Alpine communities (Upper Sole Valley, 1725-1923). *Rivista di Antropologia*, 74: 55-75.
- Mocci E., Concas M.P., Fanciulli M., Pirastu N., Adamo M., Cabras V., Fraumene C., Persico I., Sassu A., Picciau A., Prodi D.A., Serra D., Biino G., Pirastu M. & Angius A. 2009. Microsatellites and SNPs linkage analysis in a Sardinian genetic isolate confirms several essential hypertension loci previously identified in different populations. *BMC Med. Genet.*, 10: 81.
- Nasidze I., Quinque D., Udina I., Kunizheva S. & Stoneking M. 2007. The Gagauz, a linguistic enclave, are not a genetic isolate. *Ann. Hum. Genet.*, 71: 379-89.
- North K.E. & Crawford M.H. 1996. Isonymy and repeated pairs analysis: The mating structure of Acceglio, Italy, 1889-1968. *Rivista di Antropologia*, 74: 83-103.
- Peltonen L., Palotie A. & Lange K. 2000. Use of population isolates for mapping complex traits. *Nat. Rev. Genet.* 1:182-190.
- Pettener D., Gueresi P. & Martuzzi-Veronesi F. 1994. Struttura biodemografica della valle del Fersina (Valle dei Mocheni). *Bollettino di Demografia Storica*, 20: 131-156.
- Pichler I., Mueller J.C., Stefanov S.A., De Grandi A., Beu Volpato C., Pinggera G., Mayr A., Ogriseg M., Ploner F., Meitinger T. & Pramstaller P.P. 2006. *Hum. Biol.*, 78: 441-464.
- R Development Core Team. 2010. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org>.
- Relethford J.H. 1988. Estimation of kinship and genetic distance from surnames. *Hum. Biol.*, 60: 475-492.
- Rodriguez Diaz R. & Blanco-Villegas M.J. 2010. The genetic structure of a rural region in Spain: distribution of surnames and genetic flow. *Hum. Biol.*, 82: 301-314.
- Shifman S., Kuypers J., Kokoris M., Yakir B. & Darvasi A. 2003. Linkage disequilibrium patterns of the human genome across populations. *Hum. Mol. Genet.*, 12: 771-776.
- Shimodaira, H. 2004. Approximately unbiased tests of regions using multistep-multiscale bootstrap resampling. *Ann. Stat.*, 32: 2616-2641.
- Shlush L.I., Behar D.M., Yudkovsky G., Templeton A., Hadid Y., Basis F., Hammer M., Itzkovitz S. & Skorecki K. 2008. The Druze: a population genetic refugium of the Near East. *PLoS ONE*, 3:e2105. doi:10.1371/journal.pone.0002105
- Sibilla P. 1980. Una comunità Walser delle Alpi. Strutture tradizionali e processi culturali. Olschki editore, Firenze.
- Suzuki R. & Shimodaira H. 2009. *pvcust: Hierarchical Clustering with P-Values via Multiscale Bootstrap Resampling. R package version*

- 1.2-1. <http://www.is.titech.ac.jp/~shimo/prog/pvclust/>
- Tepper B.J., Koelliker Y., Zhao L., Ulrich N.V., Lanzara C., d'Adamo P., Ferrara A., Ulivi S., Esposito L. & Gasparini P. 2008. Variation in the bitter-taste receptor gene TAS2R38, and adiposity in a genetically isolated population in Southern Italy. *Obesity (Silver Spring)*, 16: 2289-95.
- Traglia M., Sala C., Masciullo C., Cverhova V., Lori F., Pistis G., Bione S., Gasparini P., Ulivi S., Ciullo M., Natile T., Bosi E., Sirtori M., Mignogna G., Rubinacci A., Buetti I., Camaschella C., Petretto E. & Toniolo D. 2009. Heritability and demographic analyses in the large isolated population of Val Borbera suggest advantages in mapping complex traits genes. *PLoS One*, 4: e7554.
- Varilo T. & Peltonen L. 2004. Isolates and their potential use in complex gene mapping efforts. *Curr. Opin. Genet. Dev.*, 14: 316-323.
- Varzari A., Kharkov V., Stephan W., Dergachev V., Puzyrev V., Weiss E.H. & Stepanov V. 2009. Searching for the origin of Gagauzes: inferences from Y-chromosome analysis. *Am. J. Hum. Biol.*, 21: 326-36.
- Vona G., Calò C.M., Lucia G., Mameli G.E., Succa V., Esteban E. & Moral P. 1996. Genetics, geography, and culture: the population of S. Pietro Island (Sardinia, Italy). *Am. J. Phys. Anthropol.*, 100: 461-71.
- Wehrens R. & Buydens L.M.C. 2007. Self- and Super-organising Maps in R: the kohonen package. *J. Stat. Softw.*, 21(5).
- Zinsli P. 1968. *Walser Volkstum in der Schweiz, in Vorarlberg, Liechtenstein und Piemont*. Huber, Frauenfeld und Stuttgart.

Editor, Giovanni Destro-Bisol

