Human Biology

Volume 84 | Issue 3 Article 2

2012

General Method to Unravel Ancient Population Structures through Surnames, Final Validation on Italian Data

Alessio Boattini
Dipartimento di Biologia ES, University of Bologna, Italy

Antonella Lisa
Istituto di Genetica Molecolare, CNR, Pavia, Italy

Ornella Fiorani
Istituto di Genetica Molecolare, CNR, Pavia, Italy

Gianna Zei
Istituto di Genetica Molecolare, CNR, Pavia, Italy

Davide Pettener
Dipartimento di Biologia ES, University of Bologna, Italy

Franz Manni
Unité d'Eco-Anthropologie et Ethnobiologie, National Museum of Natural History, Paris, manni@mnhn.fr

General Method to Unravel Ancient Population Structures through Surnames, Final Validation on Italian Data

Abstract

We analyze the geographic location of 77,451 different Italian surnames (17,579,891 individuals) obtained from the lists of telephone subscribers of the year 1993.

By using a specific neural network analysis (Self-Organizing Maps, SOMs), we automatically identify the geographic origin of 49,117 different surnames. To validate the methodology, we compare the results to a study, previously conducted, on the same database, with accurate supervised methods. By comparing the results, we find an overlap of 97%, meaning that the SOMs methodology is highly reliable and well traces back the geographic origin of surnames at the time of their introduction (Late Middle Ages/Renaissance in Italy).

SOMs results enables one to distinguish monophyletic surnames from polyphyletic ones, that is surnames having had a single geographic and historic origin from those that started to be in use, with an identical spelling, in different locations (respectively, 76.06% and 21.05% of the total). As we are interested in geographic origins, polyphyletic surnames are excluded from further analyses.

By comparing the present location of each monophyletic surname to its inferred geographic origin in late Middle Ages/Renaissance, we measure the extent of the migrations having occurred in Italy since that time. We find that the percentage of individuals presently living in the very area where their surname started to be in use centuries ago is extremely variable (ranging from 22.77% to 77.86% according to the province), thus meaning that self-assessed regional identities seldom correspond to the "autochthony" they imply. For example the upper part of the Thyrennian coast (Northern Latium, Tuscany) has a strong identity but few "autochthonous" inhabitants (28%) having been a passageway from the North to the South of Italy.

Keywords

Surnames, Family Names, Y-Chromosome, Italy, Migrations, Population Genetics, Demography

Cover Page Footnote

The lists of the surnames autochthonous of each province can be made available to third parties, for scientific purposes, upon request to the corresponding author. F. Manni would like to acknowledge the Wenner Gren Foundation (grant n° 7247) that financed a seminal and similar research and dedicates this article to Professor Italo Barrai (University of Ferrara, Italy). A. Lisa, O. Fiorani, and G. Zei were supported by a grant CARIPLO 2010–0253. A. Boattini and D. Pettener were supported by grants PRIN 2007 and 2009 to D. P. To avoid any possible conflict of interests, the editorial process has been independently secured by Dr. Pierre Darlu (CNRS, France).

General Method to Unravel Ancient Population Structures through Surnames, Final Validation on Italian Data

ALESSIO BOATTINI,[1+] ANTONELLA LISA,[2+] ORNELLA FIORANI,[2] GIANNA ZEI,[2] DAVIDE PETTENER,[1] AND FRANZ MANNI[3*]

Abstract We analyze the geographic location of 77,451 different Italian surnames (17,579,891 individuals) obtained from the lists of telephone subscribers of the year 1993.

By using a specific neural network analysis (Self-Organizing Maps, SOMs), we automatically identify the geographic origin of 49,117 different surnames. To validate the methodology, we compare the results to a study, previously conducted, on the same database, with accurate supervised methods. By comparing the results, we find an overlap of 97%, meaning that the SOMs methodology is highly reliable and well traces back the geographic origin of surnames at the time of their introduction (Late Middle Ages/Renaissance in Italy).

SOMs results enables one to distinguish monophyletic surnames from polyphyletic ones, that is surnames having had a single geographic and historic origin from those that started to be in use, with an identical spelling, in different locations (respectively, 76.06% and 21.05% of the total). As we are interested in geographic origins, polyphyletic surnames are excluded from further analyses.

By comparing the present location of each monophyletic surname to its inferred geographic origin in late Middle Ages/Renaissance, we measure the extent of the migrations having occurred in Italy since that time. We find that the percentage of individuals presently living in the very area where their surname started to be in use centuries ago is extremely variable (ranging from 22.77% to 77.86% according to the province), thus meaning that self-assessed regional identities seldom correspond to the "autochthony" they imply. For example the upper part of the Thyrennian coast (Northern Latium, Tuscany) has a strong identity but few "autochthonous" inhabitants (~28%) having been a passageway from the North to the South of Italy.

[1] Dipartimento di Biologia ES, University of Bologna, Italy.
[2] Istituto di Genetica Molecolare, CNR, Pavia, Italy.
[3] Unité d'Eco-Anthropologie et Ethnobiologie, UMR 7206 CNRS, National Museum of Natural History, University Paris Diderot, Sorbonne Paris Cité, F-75005 Paris, France.
*Corresponding author: Dr. Franz Manni, UMR 7206 National Museum of Natural History-Musée de l'Homme, CP 139, 57 rue Cuvier, 75231 Paris Cedex 05 – France. E-mail: manni@mnhn.fr.
+These authors contributed equally to the study.

Human Biology, June 2012, v. 84, no. 3, pp. 235–270.

KEY WORDS: SURNAMES, FAMILY NAMES, Y-CHROMOSOME, ITALY, MIGRATIONS, POPULATION GENETICS, DEMOGRAPHY.

Our methodology can be generalized to countries where family names are patrilineally inherited, enabling the fast design of surname samples representative of the population living in given areas at the time of family names introduction. By typing the corresponding Y-chromosomes, it is possible to better depict past anthropological variation and to identify ancient population structures otherwise hidden by migratory flows.

Ho visto un lampo illuminare scene del futuro.
Gli anni mi dividono in sparse parti.
Il numero sapessi lascia tracce.

I have seen a flash enlightening the future.
The years split me in scattered fragments.
The number, you know, leaves its imprint.

—Franco Battiato, "Scherzo in Minore" (2001)

For over 50 years, surname studies have had a long-lasting tradition both in anthropology and population genetics, as they provide a costless and efficient method to describe some aspects of the genetic and demographic variability of populations (for a review of such studies, see Cavalli-Sforza et al. 2004 and Darlu et al. 2012). More recently and in the frame of molecular studies focused on the variability of the Y-chromosome, the study of surnames gained new prominence and has been advocated to design more careful sampling strategies (Boattini et al. 2010; Manni et al. 2005). Nowadays the technological constraints related to DNA typing are becoming negligible, thus allowing more energy for a meticulous collection of the samples.

In the last years, several British scholars have been particularly active in exploring the link existing between surnames of patrilineal descent and the variability of the non-recombining portion of the Y-chromosome (Graf et al. 2010; Jobling 2001; Jobling and Tyler-Smith 2003; King et al. 2006; King and Jobling 2009a, 2009b; Martínez-González et al. 2012; McEvoy and Bradley 2006; Sykes and Irven 2000; Winney et al. 2012). In particular, King and Jobling (2009a) examined the Y-chromosome diversity among the bearers of forty specific British surnames and observed that the degree of co-ancestry increases with the rareness of the surname.

More generally, surnames collected from historical records and pedigrees have been used to increase the "archaeogenetic" power of anthropological studies (Bedoya et al. 2006; Boattini et al. 2011; Bowden et al. 2008; Darlu et al. 2012; Hill et al. 2000; McEvoy et al. 2006; Moore et al. 2006; Shlush et al. 2008; Zei et al. 2003), similarly to the investigations addressing the transmission of Mendelian traits and/or complex diseases in isolated populations (Angius et al. 2001; Colonna et al. 2007; Mocci et al. 2009; Traglia et al. 2009). Unfortunately, the analysis of historical records and pedigrees is time-demanding and, besides specific research projects, discourages a wider effort to depict the genetic variability of the past.

To overcome this limitation, Manni et al. (2005) introduced a general surname method (General Method) providing a fast and efficient identification of those geographical areas that remained genetically closer to their past genetic makeup: in other words where immigration has been less intense. When studies about past DNA variability are undertaken, this method enables the design of a more efficient Y-chromosome sampling, because donors can be selected according to the geographic origin of their patrilineal ancestor (being such inference limited to the interval spanning from the present to the time of surname introduction). By using Self-Organizing Maps (SOMs), a clustering technique based on neural networks (Kaski 1997; Kohonen 1982, 1984), Manni et al. (2005) examined the whole surname body of The Netherlands all at once and without the need of any historical or genealogical insight. This study enabled the identification of groups of surnames sharing the same geographic origin and an identical migration history, allowing the targeting of individuals still living in the area where their surname originated (in other words "autochthonous"—for an application see Manni and Toupance 2010). Manni et al. (2005) suggested that the degree of autochthony can considerably vary from zone to zone, being influenced by geographical and historical factors that result in differential patterns of migration.

If the General Method improves the accuracy of biodemographical inference, it also enables the estimation of more reliable population structures, as polyphyletic surnames (that are spelled in the same way though having multiple independent origins) can be identified and their confounding effect controlled. To say it differently, Y-chromosome studies can benefit from the definition of lists of donors representative of the area considered, thus virtually sampling a population as it was at the time of surname introduction.

Recently, Boattini et al. (2010) and Rodriguez-Diaz and Blanco Villegas (2010) have applied the General Method to isolated Italian and Spanish populations to unravel otherwise totally hidden ancient genetic structures. These two studies demonstrate the effectiveness of the General Method in dissecting population structures even at a micro-geographic level. With respect to the methodology of Manni et al. (2005), a final validation is required as the databases analyzed so far were of a low complexity, either because they concerned small geographic areas (Boattini et al. 2010, Rodriguez-Diaz and Blanco Villegas 2010) or because they implied a limited temporal depth, as in The Netherlands where surnames were officially introduced by Napoleon. Finally, the geographical complexity (hydrography and orography) of the areas embraced by these three studies was negligible, thus leaving the performance of the General Method untested in more challenging geographic regions.

In this paper we apply the general surname method of Manni et al. (2005) to one of the most awkward data set available, that is the surnames of the Italian population. Italy has one of the highest amounts of different surnames (~330,000) of any country (De Felice 1980, 1982), is geographically very diverse with two major chains of mountains (the Alps and the Appennines), and has very extended coastlines (~7460 km) (see Figure 1). To add to this complexity, many dialects are spoken in Italy, and their diversity, between and within varieties, is probably the highest of all

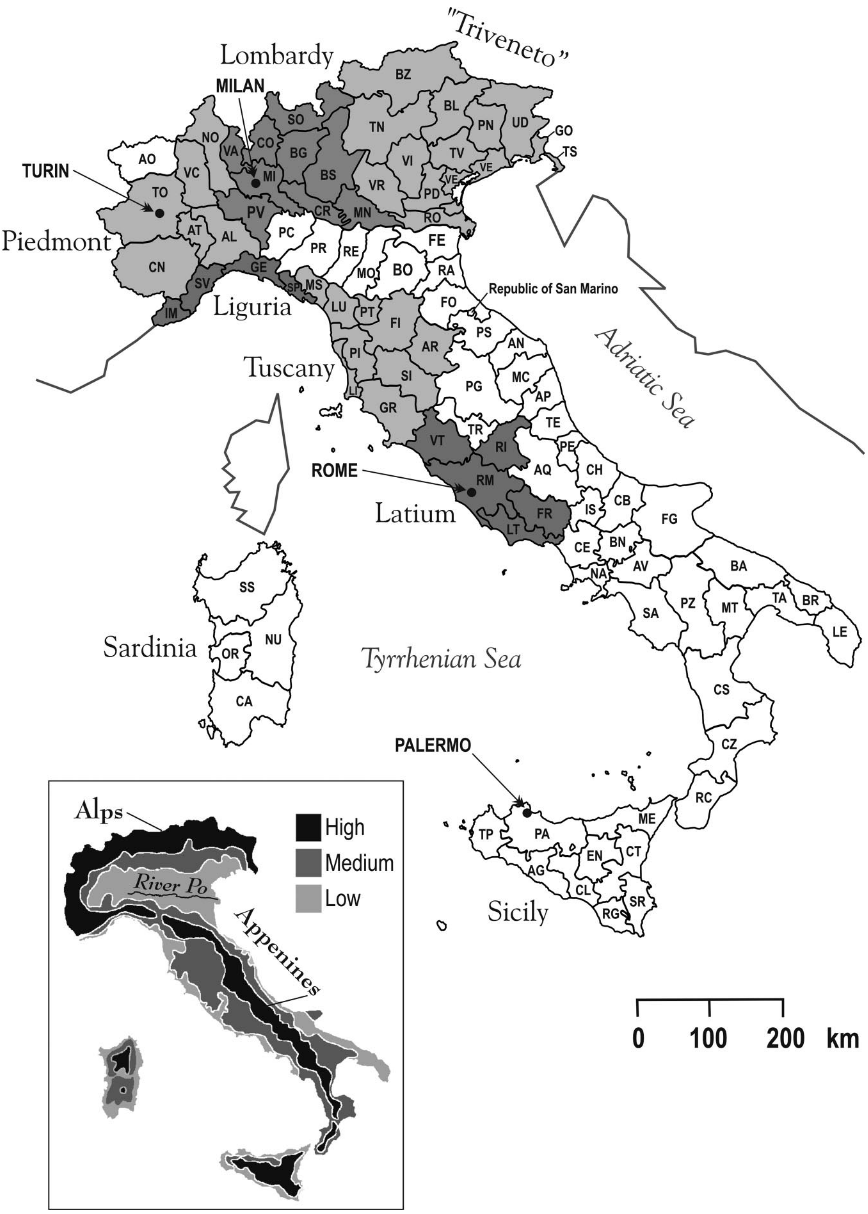

Figure 1. Map of Italy showing the 95 provinces (and the Republic of San Marino) according to the political asset of the year 1993. Fully spelled names of provinces are reported in Table 1. Major Italian cities and some regions mentioned in the text are labeled. A small map about the main geographic features of the country is reported at the bottom (with average altitudes above the sea level: "High," "Medium," "Low").

Europe. It may have had an influence on surname variability as family names are also words (Manni et al. 2006, 2008).

By studying Italy after The Netherlands, we adopt an approach by country that may be questionable as it does not correspond to the outcomes of more comprehensive population studies pointing to the weakness of geographical operational units defined according to political borders, as it is the case with nations. Nevertheless and concerning Western Europe, Cheshire et al. (2011) demonstrate that country borders and cultural zones overlap almost exactly: a consequence of the strong nationalist policies pursued for centuries. In this context, Italy offers several layers of diversity, its identity being a blend composed by long-lasting remnants of regionalism. In fact, the modern Italian state was created only 150 years ago from a large number of independent states that all had a rather long history.

The purpose of this paper is 4-fold: (1) to automatically identify, through the SOMs clustering technique, those Italian surnames that unambiguously originated in each of the 95 Italian provinces (plus the Republic of San Marino) at the time of surname introduction (13th–17th centuries); (2) to crosscheck this SOMs-based inference with a data bank of monophyletic surnames previously assembled with a more elaborate and supervised approach (see Methods) by some of us (AL, OF, GZ); (3) to compare our retrospective inference to the present-day distribution of surnames in order to describe the migrations that took place in the last centuries; and (4) to identify the provinces that, today, are more “autochthonous” with respect to the Medieval and Renaissance period, that is when surnames spread.

By validating the General Method we advocate its use to safely and rapidly reconstruct, through migration matrices, the population mobility of any area or country where surnames of patrilineal descent are available. As we will discuss, this methodology may constitute the backbone for a trustworthy depiction of ancient and remote Y-chromosome variability.

Materials and Methods

Surnames in Italy. The origin and diffusion of most Italian surnames are related to the Ecumenical Council of the Roman Catholic Church held in Trento (Italy) from 1545 to 1563. After the Council of Trento, all parishes had to keep exhaustive birth and marriage records (death and census records became compulsory in 1614). Even if the late 16th century can be considered the beginning of Italian surnames, temporal differences exist over the country. While in rural or mountainous areas (e.g., the Central Apennines) their use started in the 16th/17th century, birth and marriage records are documented in several urban areas (e.g., Venice and Florence) as early as the 12/13th century. Temporal differences also relate to social status because, as in other European countries, prominent social groups generally had family names long before lower social classes. Anyway, the majority of Italian surnames can be traced back to the beginning of the 17th century and have a time depth of four centuries at least.

The Database. The database analyzed here was extracted from the complete national telephone directory of the year 1993 (SEAT—Società Elenchi Abbonati al Telefono) accounting for 18,554,688 subscribers (~33% of the whole population of that year). To infer the geographical origin of surnames, a frequency-gradient based on their current geographical distribution (see below) was necessary; therefore those family names having a frequency lower than 20 occurrences were excluded from the study as the corresponding frequency-gradient could have been misleading as stochastic phenomena were likely to weaken computed gradients. Such cut-off value is based on empiric experience (Manni et al. 2005), but slightly different ones (18, 19, 21, 22 occurrences) would have yielded similar results. Furthermore, we note that the surnames excluded from the analysis correspond to 5.25% of the whole total with a similar percentage region-by-region, meaning that the results of the analysis are not geographically biased.

The database finally analyzed consisted of 77,451 different surnames corresponding to 17,579,891 individuals (31% of the Italian population of that time). Data were processed according to 96 operational units (Figure 1 and Table 1), that is the 95 administrative subdivisions (provinces) existing in 1993 plus the Most Serene Republic of San Marino, the oldest surviving constitutional Republic of the world located in the northeastern part of the Italian peninsula. Small islands could not be addressed as they never constitute autonomous provinces.

Automatic Identification of the Geographical Origin of Surnames by SOMs

Data Inputs. The automatic geographical clustering of Italian surnames has been obtained with SOMs, a technique derived from neural networks and based on "competitive learning" (Kohonen, 1982, 1984). The methodology, briefly described in the next section, is an adaptive process in which the cells (neurons) of a network (map) gradually become sensitive to different input-vectors that are finally mapped to the neurons that best describe them. In our case the input-vectors are the 77,451 different Italian surnames of the database whose 96 components correspond to their weighted frequency in the 95 Italian provinces plus the Republic of San Marino. The weighted frequencies were obtained as follows.

(1) The absolute surname frequencies in a given province p (F_{ip}) were weighted by the natural logarithm of the corresponding population size as estimated from the data (N_p):

$$f_{ip} = F_{ip}/\mathrm{Ln}(N_p). \tag{1}$$

By this step we were able to scale surname frequencies according to population sizes. The logarithmic procedure is related to the fact that the population sizes of various provinces vary by different orders of magnitude; a simple division by *Np* would have led to an increased, and erroneous, attribution of many surnames to those provinces that have a small population size.

(2) The surname frequencies f_{ip} were weighted a second time by their absolute frequencies in the whole set of P Italian provinces by subdividing them by their per-surname sum:

Table 1. Migration Statistics for Each Italian Province According to the Results of the General Method[a]

Cod.	*Province*	*Region*	*Tel*	*Mono*	*Avg Mono*	*Poly* (%)	*dF* (%)	*dB* (%)	*MB* (%)	d_i (*km*)	d_j (*km*)
AG	Agrigento	Sicily	135,423	45191.0	107.85	62.60	48.87	48.11	0.76	494.16	190.44
AL	Alessandria	Piedmont	164,758	44383.0	104.92	51.67	58.01	71.65	13.64	131.94	314.10
AN	Ancona	Marche	146,044	47741.0	93.79	60.21	56.70	58.42	1.72	203.97	204.02
AO	Aosta	Aosta Valley	52,187	15251.5	51.88	34.11	25.08	56.31	31.23	223.06	424.93
AP	Ascoli Piceno	Marche	103,345	63331.0	118.38	53.90	59.68	38.99	20.70	187.01	197.69
AQ	L'Aquila	Abruzzo	109,891	36330.0	74.45	60.45	57.90	57.59	0.32	159.48	202.15
AR	Arezzo	Tuscany	102,079	78135.0	133.00	49.08	69.82	47.91	21.91	160.26	174.43
AT	Asti	Piedmont	74,990	49949.0	125.97	47.02	73.19	59.73	−13.46	143.00	351.42
AV	Avellino	Campania	113,992	45374.0	103.24	61.53	60.54	51.47	−9.07	319.14	164.29
BA	Bari	Apulia	441,242	287009.0	167.06	47.79	47.32	26.84	−20.49	488.95	231.68
BG	Bergamo	Lombardy	299,858	142522.0	206.55	50.77	42.21	39.28	−2.93	124.13	266.63
BL	Belluno	Veneto	79,116	54738.0	97.57	36.38	48.06	35.99	−12.07	211.44	204.57
BN	Benevento	Campania	75,842	48554.0	116.86	53.81	65.13	44.37	−20.76	307.83	136.10
BO	Bologna	Emilia-Romagna	350,980	100804.0	168.01	58.33	43.53	55.94	12.41	151.22	245.30
BR	Brindisi	Apulia	115,026	69144.0	158.41	48.49	59.87	48.48	−11.39	428.43	169.51
BS	Brescia	Lombardy	332,484	141711.0	119.84	49.95	42.27	44.58	2.31	142.99	231.64
BZ	Bolzano	Trentino Alto Adige	121,268	68299.0	70.48	21.92	22.14	30.14	8.00	241.00	247.90
CA	Cagliari	Sardinia	219,697	58194.0	150.76	68.21	31.39	35.86	4.47	439.63	369.65
CB	Campobasso	Molise	69,999	36693.0	89.06	52.22	60.37	47.10	−13.27	292.15	174.85
CE	Caserta	Campania	216,051	103931.5	141.69	55.80	58.09	46.59	−11.49	245.50	142.79
CH	Chieti	Abruzzo	117,959	75205.0	116.06	51.40	57.56	36.35	−21.21	244.88	201.56
CL	Caltanissetta	Sicily	82,231	56323.0	125.16	53.48	68.58	48.42	−20.16	525.64	168.74
CN	Cuneo	Piedmont	178,919	116861.5	138.96	43.05	53.00	38.47	−14.53	166.83	340.68
CO	Como	Lombardy	273,058	89402.0	213.37	54.12	52.42	61.83	9.41	160.44	338.97
CR	Cremona	Lombardy	104,074	70157.0	129.32	51.25	72.48	57.05	−15.43	125.73	184.84
CS	Cosenza	Calabria	200,562	82897.5	122.81	58.11	47.29	40.11	−7.18	527.66	246.46
CT	Catania	Sicily	321,408	167339.5	239.57	51.46	44.88	35.18	−9.70	505.77	193.48

Table 1. (*continued*)

Cod.	*Province*	*Region*	*Tel*	*Mono*	*Avg Mono*	*Poly* (%)	*dF* (%)	*dB* (%)	*MB* (%)	d_i (*km*)	d_j (*km*)
CZ	Catanzaro	Calabria	200,316	128285.5	132.73	50.69	55.61	34.14	−21.47	619.60	256.79
EN	Enna	Sicily	54,223	25017.0	80.96	57.37	68.78	60.43	−8.35	461.90	143.82
FE	Ferrara	Emilia-Romagna	125,215	100894.0	208.89	52.46	68.84	42.37	−26.47	176.88	173.61
FG	Foggia	Apulia	189,099	85283.0	99.80	50.04	58.71	48.64	−10.07	429.60	192.19
FI	Firenze	Tuscany	407,891	120414.7	149.27	55.46	50.38	61.94	11.56	137.34	241.83
FO	Forlì	Emilia-Romagna	197,749	58391.5	142.77	62.45	44.71	51.16	6.46	151.44	213.47
FR	Frosinone	Lazio	141,273	81544.2	116.83	53.84	57.24	39.19	−18.05	165.17	212.87
GE	Genova	Liguria	397,743	117424.0	180.10	47.28	43.48	62.91	19.43	180.88	381.19
GO	Gorizia	Friuli Venezia Giulia	49,951	25060.0	63.93	29.74	60.35	62.46	2.11	172.59	224.95
GR	Grosseto	Tuscany	89,471	25314.0	92.39	55.63	72.94	79.86	6.92	171.59	212.22
IM	Imperia	Liguria	101,288	20792.0	113.00	46.92	50.94	77.00	26.05	232.66	404.11
IS	Isernia	Molise	27,122	15443.0	85.79	56.65	67.29	49.31	−17.99	226.85	155.75
LE	Lecce	Apulia	237,312	194142.5	188.76	43.49	48.83	19.00	−29.83	538.19	301.46
LI	Livorno	Tuscany	127,234	54972.7	127.65	53.50	77.23	75.69	−1.54	129.60	221.36
LT	Latina	Lazio	160,912	37021.7	88.78	52.93	49.94	70.95	21.01	196.52	254.36
LU	Lucca	Tuscany	135,369	71956.5	167.54	53.32	58.14	46.49	−11.65	170.35	198.25
MC	Macerata	Marche	88,246	54604.0	89.30	51.58	64.29	46.77	−17.51	179.02	175.76
ME	Messina	Sicily	215,158	138130.5	159.32	48.13	55.45	38.19	−17.26	498.33	205.92
MI	Milan	Lombardy	1,432,141	227674.5	194.10	52.76	38.26	75.07	36.81	126.47	409.85
MN	Mantova	Lombardy	107,475	51517.0	106.66	54.04	69.57	63.53	−6.04	143.31	162.28
MO	Modena	Emilia-Romagna	203,534	85904.5	193.26	54.13	51.60	50.91	−0.68	130.14	207.37
MS	Massa Carrara	Tuscany	71,924	30487.5	102.48	55.06	58.19	55.08	−3.11	171.14	204.43
MT	Matera	Basilicata	56,293	32024.0	95.31	51.02	62.18	45.01	−17.17	420.23	192.13
NA	Napoli	Campania	828,923	398208.5	216.65	52.39	44.04	33.24	−10.81	297.37	206.66
NO	Novara	Piedmont	177,326	56760.5	93.43	48.02	56.96	67.40	10.45	217.46	364.35
NU	Nuoro	Sardinia	75,087	33894.5	105.26	63.94	56.19	38.14	−18.05	279.62	221.04
OR	Oristano	Sardinia	44,436	16273.0	113.01	70.94	63.87	49.23	−14.64	300.90	207.17

Table 1. (*continued*)

Cod.	*Province*	*Region*	*Tel*	*Mono*	*Avg Mono*	*Poly* (%)	*dF* (%)	*dB* (%)	*MB* (%)	d_i (*km*)	d_j (*km*)
PA	Palermo	Sicily	388,343	184323.0	143.16	54.02	46.81	37.39	−9.42	505.18	227.85
PC	Piacenza	Emilia-Romagna	94,288	59767.0	140.96	53.10	65.96	45.72	−20.24	154.05	238.66
PD	Padova	Veneto	243,839	144020.0	161.19	43.45	38.40	44.09	5.68	168.88	192.31
PE	Pescara	Abruzzo	90,754	51194.5	115.30	50.74	65.63	53.87	−11.76	235.34	177.15
PG	Perugia	Umbria	187,586	105339.0	86.63	50.08	59.29	44.01	−15.28	191.62	216.03
PI	Pisa	Tuscany	127,987	47794.7	146.76	50.93	67.33	72.28	4.95	124.48	194.77
PN	Pordenone	Friuli Venezia Giulia	87,542	58945.5	100.68	33.91	59.90	53.12	−6.78	185.20	165.20
PR	Parma	Emilia-Romagna	140,349	43613.0	101.90	57.34	49.02	56.98	7.96	141.22	215.85
PS	Pesaro	Marche	107,464	51617.0	100.91	59.14	62.33	48.88	−13.45	198.13	212.71
PT	Pistoia	Tuscany	86,218	30927.5	174.24	55.66	66.61	69.70	3.10	130.84	190.64
PV	Pavia	Lombardy	175,777	64951.5	109.44	52.43	66.43	68.36	1.92	150.87	281.74
PZ	Potenza	Basilicata	110,685	44097.5	99.10	49.98	53.02	44.29	−8.73	455.50	205.91
RA	Ravenna	Emilia-Romagna	122,200	37317.5	160.85	59.79	53.42	60.57	7.15	132.86	176.64
RC	Reggio Calabria	Calabria	158,954	159617.5	169.90	41.62	59.57	23.59	−35.97	624.31	235.84
RE	Reggio Emilia	Emilia-Romagna	134,346	99251.5	223.29	49.00	63.05	41.73	−21.33	136.33	199.82
RG	Ragusa	Sicily	94,081	96915.5	231.03	40.00	59.39	25.79	−33.60	476.12	192.15
RI	Rieti	Lazio	58,988	17360.0	78.55	60.81	61.11	65.58	4.47	129.38	191.48
RM	Rome	Lazio	1,378,925	139189.2	88.85	55.41	35.99	82.02	46.03	208.41	286.81
RO	Rovigo	Veneto	73,954	32221.0	100.69	56.67	68.00	58.52	−9.48	206.07	163.27
SA	Salerno	Campania	288,574	123784.0	149.41	59.16	55.44	45.70	−9.74	290.42	163.49
SI	Siena	Tuscany	89,584	38324.0	116.13	49.87	67.23	68.69	1.46	139.05	159.51
SM	S. Marino	S. Marino	7,797	1609.0	73.14	70.90	50.71	61.22	10.51	158.65	126.50
SO	Sondrio	Lombardy	58,046	29285.0	71.08	45.37	46.57	42.90	−3.68	187.23	259.90
SP	La Spezia	Liguria	89,028	19039.0	77.39	53.71	48.02	71.64	23.62	174.52	266.06
SR	Siracusa	Sicily	125,094	82144.5	145.65	46.66	59.55	45.16	−14.39	509.52	189.51
SS	Sassari	Sardinia	143,853	58974.5	117.95	60.46	48.55	37.72	−10.83	365.61	340.10
SV	Savona	Liguria	138,509	32205.0	115.02	46.97	54.86	76.33	21.46	222.48	288.95

Table 1. (*continued*)

Cod.	*Province*	*Region*	*Tel*	*Mono*	*Avg Mono*	*Poly* (%)	*dF* (%)	*dB* (%)	*MB* (%)	d_i (*km*)	d_j (*km*)
TA	Taranto	Apulia	169,970	100096.5	155.55	48.90	60.20	48.55	−11.65	409.09	197.09
TE	Teramo	Abruzzo	82,066	51502.5	96.63	47.63	63.60	48.07	−15.53	198.23	171.08
TN	Trento	Trentino Alto Adige	163,252	84465.0	92.06	44.06	38.40	35.07	−3.33	178.82	237.05
TO	Turin	Piedmont	818,426	136119.5	85.39	45.93	31.70	73.51	41.81	183.26	509.48
TP	Trapani	Sicily	138,041	132939.0	216.69	46.94	63.24	28.29	−34.95	502.81	239.30
TR	Terni	Umbria	75,932	26588.5	67.66	54.17	63.12	65.44	2.31	164.77	190.24
TS	Trieste	Friuli Venezia Giulia	111,422	40878.5	58.90	30.92	48.25	57.70	9.45	294.31	291.42
TV	Treviso	Veneto	222,668	128315.5	134.43	43.26	51.88	44.64	−7.24	170.26	151.42
UD	Udine	Friuli Venezia Giulia	170,732	134036.0	93.83	31.44	49.93	32.42	−17.50	279.57	280.34
VA	Varese	Lombardy	268,694	88919.5	182.59	48.15	60.86	70.78	9.91	154.39	375.26
VC	Vercelli	Piedmont	133,291	33979.5	68.16	48.98	51.62	69.60	17.98	169.75	368.61
VE	Venezia	Veneto	268,681	149336.5	160.06	36.75	54.21	49.08	−5.13	158.74	174.04
VI	Vicenza	Veneto	235,484	127490.5	138.20	47.78	46.01	35.06	−10.95	162.24	193.83
VR	Verona	Veneto	245,730	108191.5	105.24	49.68	45.05	42.72	−2.33	169.17	228.87
VT	Viterbo	Lazio	102,340	33436.0	66.81	55.17	60.55	65.69	5.14	164.89	220.36

a. *Tel*, Number of telephone users in the database; *Mono*, number of individuals (telephone users) living in the province and having a monophyletic surname (decimal numbers result from a correction related to ambiguous cells—see methodology); *Avg Mono*, average number of individuals sharing each; *Poly*, rate of polyphyletic surnames (computed as the percentage of individuals bearing a polyphyletic surname in each province); *dF*: emigration rate of individuals having a surname autochthonous of each province (forward matrix—see methodology); *dB*: immigration rate of individuals having a surname whose origin is outside the province (backward matrix—see methodology); *MB*, migratory balance; d_i, mean emigration distance; d_j mean immigration distance.

$$W_{ip} = f_{ip}/\sum_{p=1}^{p} f_{ip} \tag{2}$$

The latter step is meant to normalize the relative surname frequency as if all surnames had the same absolute frequencies. W_{ip} values were subsequently arranged in vectors (one per surname) and analyzed with the SOMs algorithm.

The Algorithm and the Software. In a neural network, different neurons specialize to represent different types of input-vectors. In other words, the neural network grid (map visible in Figure 2 A/B/C/D) is composed of cells that correspond to reference-vectors, which, through a learning process, mathematically adapt their components to the input-vectors that, progressively, are linked (mapped) to the most similar reference vectors encountered during the analysis. Several inputs can be associated with the same reference vector, thus giving rise to clusters. The key process, the *adaptation* of the map to the data, is governed by a neighborhood function. It means that when a reference-vector is "hit" by a similar input-vector, its vector components *adapt* to the input and so do neighboring reference-vectors. The adaptation process occurs with an intensity and a radius governed by the neighborhood function. The overall procedure results in the differentiation of the map-space: (*a*) identical input-vectors are mapped to the same neuron, (*b*) slightly different ones to close neurons, while (*c*) very different input-vectors are mapped far from each other, and (*d*) extremely different input-vectors are mapped at the highest distance allowed by the map, that is opposite corners. A full description of the method is provided by the author (Kohonen 1982, 1984) and, concerning another application to surnames, by Manni et al. (2005). All the analyses were performed using the software R (R Development Core Team, 2008). SOMs have been computed by using the software library "kohonen" (Wehrens and Buydens 2007).

Setting of the SOMs Map. The size of the map is defined by the user and determines the maximum number of different clusters (for example a 6 × 6 map yields 36 possible clusters). It can happen that only few inputs are mapped to given cells, that is, to the reference-vectors that define the vectorial properties of

Figure 2. Spatial patterns of surname distribution obtained by plotting the corrected geographic frequency (see Methods for details about the correction) of each of the clusters of surnames corresponding to the 400 cells of the 20 × 20 Self-Organizing Map (SOM) on a geographic map of Italy. The majority of the maps display a frequency peak within a single province. Detailed results are reported in Tables 1 and 2. The intensity of the frequency peak is variable (see the grayscale at the bottom). Four maps are enlarged in Figure 3. As the figure is quite big, we divided it into four parts (A, B, C, D). Cells to which the text refers can be identified by their X and Y coordinates. To increase readability, the X axes run downward. By carefully analyzing the figure, a pattern of geographical distribution of surnames emerges; it is typical of SOMs. Roughly, close clusters correspond to surnames having similar distributions and vice versa.

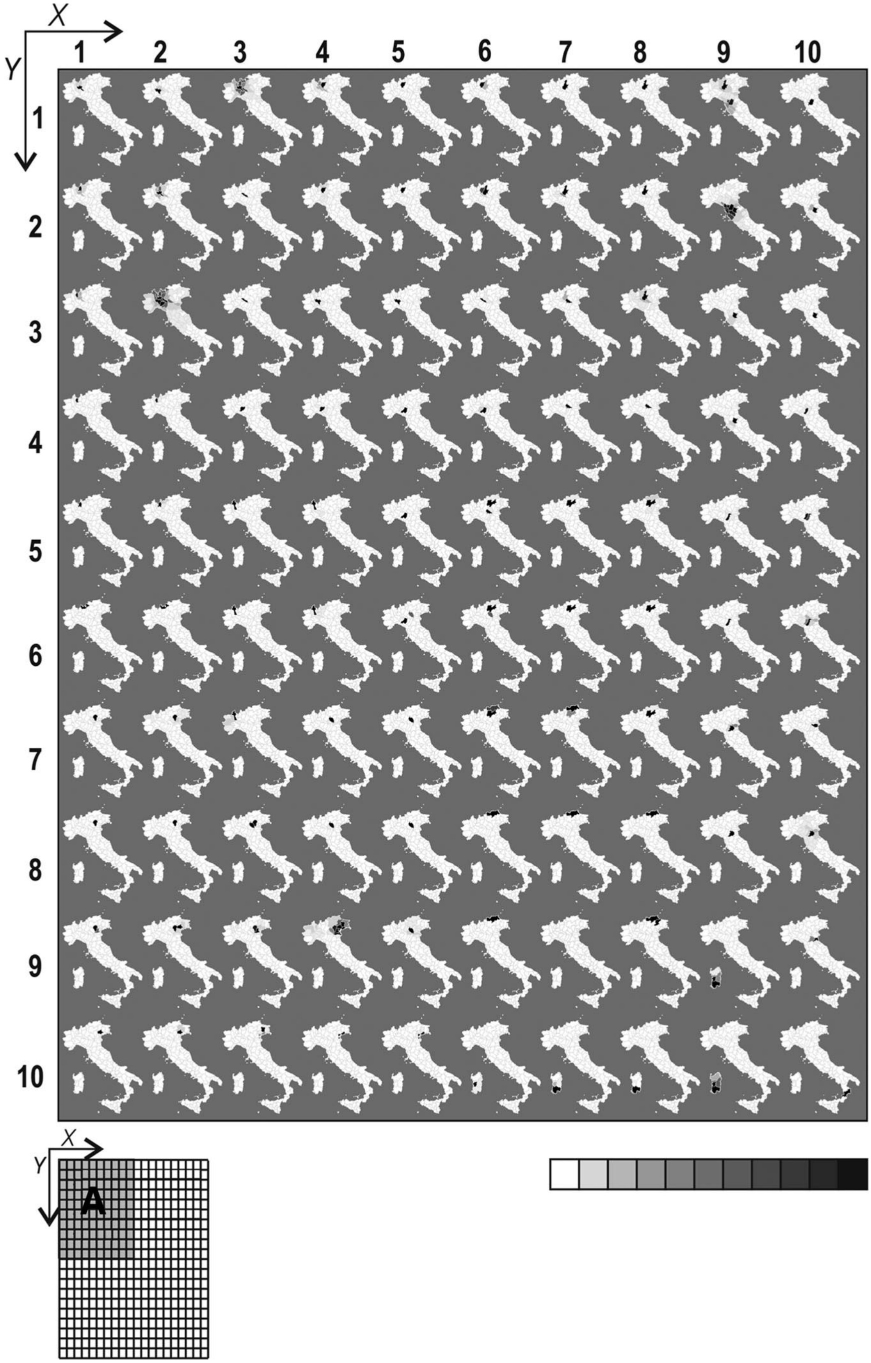

Figure 2. (A)

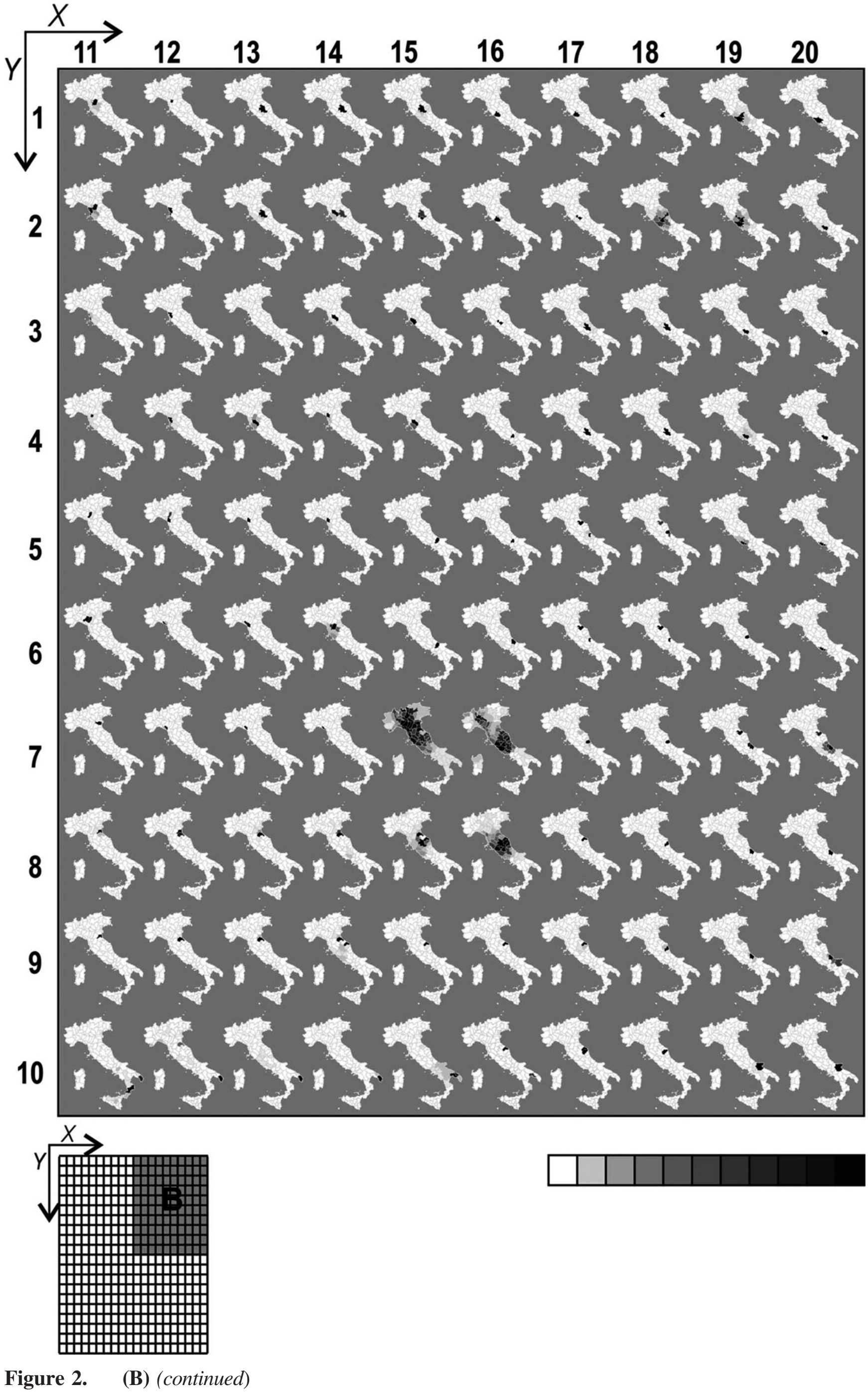

Figure 2. **(B)** *(continued)*

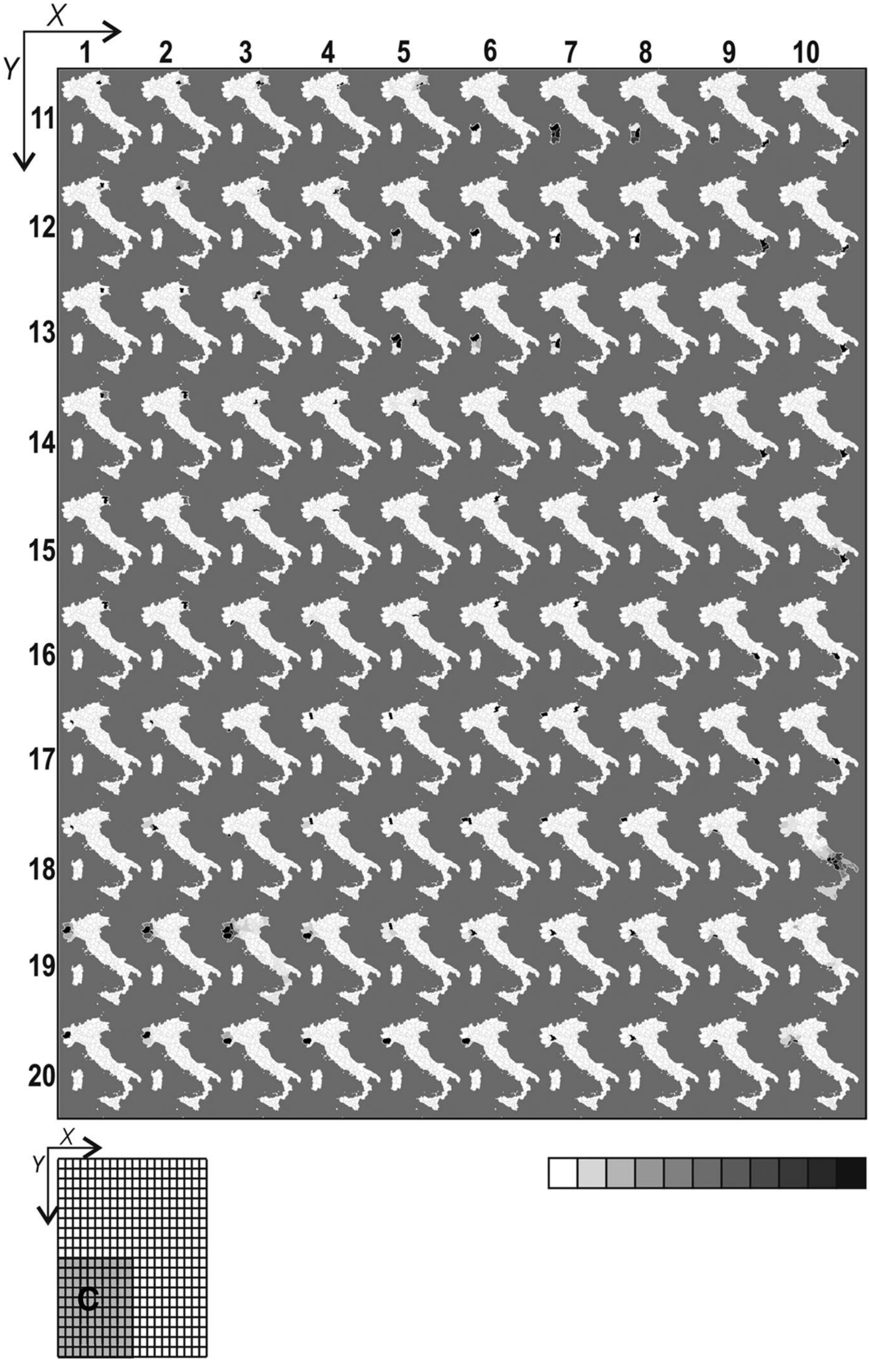

Figure 2. **(C)** (*continued*)

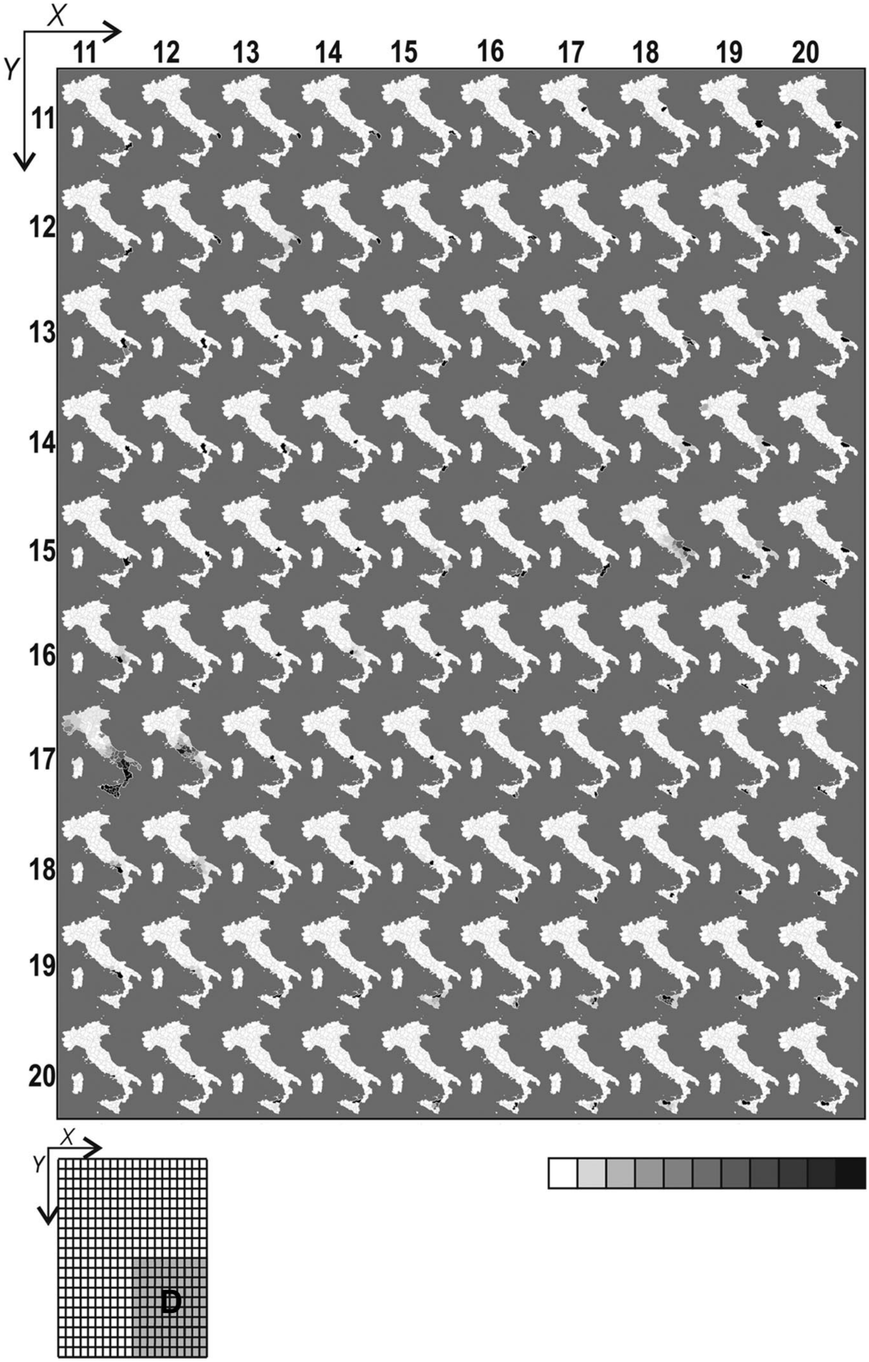

Figure 2. **(D)** (*continued*)

cells. It can also happen that some cells remain empty, meaning that there are no data input-vectors corresponding to the reference vectors defining such cells. Before adopting the 400 clusters SOM (20 × 20 cells map) of this article (Figure 2 A/B/C/D), several trials were necessary in order to achieve a good compromise between detail and synthesis. A map too small yields clusters grouping together surnames whose origin is in several neighboring provinces without a distinction between them, whereas a map too big yields many uninformative empty clusters and separates surnames belonging to a given province in too many unnecessary clusters. We wanted each province represented by *at least* a cluster (cell), and the optimal map had a size of 20 × 20 cells. The parameters of the analysis (radius, neighborhood function, learning rate function, etc.) were chosen according to the recommendations of Dr. Samuel Kaski (personal communication). The radius of the neighborhood function was set at 11 cells (half of the map plus one); the learning rate (α) was set to 0.05 and 0.01 for the first and second training procedure, respectively. The final clustering was obtained by inputting 1000 times the entire corpus of input vectors to the map.

Specific clusters (cells) of the 20 × 20 SOM of Figure 2 will be referred to by their row and column coordinates ($x;y$). As the 20 × 20 map is too big to be printed on a single page, it has been split into four sections (Figure 2 A/B/C/D). In Table 3 we provide evidence for the number of surnames clustered in the 400 cells. The small maps of Italy visible in Figure 2 correspond to a geographical frequency plot of the whole group of surnames clustered together in each cell. Once the plot was obtained (some maps are enlarged in Figure 3), it was very easy, by a simple and straightforward visual inspection, to distinguish the clusters corresponding to surnames whose geographic origin was unambiguous and located within a well-defined province (monophyletic surnames) from those not compatible with a single geographic origin (polyphiletic surnames). The 400 maps of Figure 2 were visually analyzed in this way and the results are reported in Table 3.

Migration Matrices. To estimate the migratory flows, we compared the SOMs-inferred geographic origin of surnames to their present-day distribution by a migration matrix model (Bodmer and Cavalli-Sforza 1968). A migration matrix is a formal representation of population mobility (here the population of the 95 Italian provinces plus the Republic of San Marino) across one or more generations (approximately 16 generations by considering a generation every 25 years in the 400 years time depth of Italian surnames). Data were organized in a 96 × 96 migration matrix (M), where rows represent the provinces of origin of surnames, and columns represent their current location. In this way, the Mij element of the matrix M corresponds to individuals bearing a surname that historically originated in the province i and currently residing in the province j. Such M matrix was then transformed in a forward matrix F, by subdividing each element of M by the row sum (row-stochastic matrix), and in a backward matrix B by subdividing each element of M by the column sum (column-stochastic matrix). The F matrix tells to which provinces emigration took place. The B

Table 2. Patterns of *dF* and *dB* Values and Interpretation of Some Extreme Scenarios (see Figure 6)

	dB	*dF*	*Diagnosis*
A	**High** *Many immigrants came from outside.*	**High** *The original population and their descendants massively emigrated.*	The genetic background is highly mixed. The area is a corridor with a high turnover of the population.
B	**Low** *Few immigrants came from outside.*	**High** *The original population and their descendants massively emigrated.*	The genetic background is autochthonous even though the dispersal to other areas was high. Situation typical of unattractive areas.
C	**High** *Many immigrants came from outside.*	**Low** *The original population and their descendants did not move.*	The genetic background is highly mixed. The area is highly attractive.
D	**Low** *Few immigrants came from outside.*	**Low** *The original population and their descendants did not move.*	The genetic background is autochthonous and the dispersal to other areas was limited. Situation typical of self-sufficient isolated areas.

Table 3. Geographic Origin of Surnames and Summary Statistics for the 400 Cells of the Self-Organizing Map (SOM)[a]

Y↓X→	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	MI	PV, MI	*PF*	*PF*	BG	BG, BS	BS	BS	*PF*	FI	FI	PT	PG	PG	PG	VT	VT, RM	RI	RM	RM
	371	199	423	136	299	83	392	295	39	244	317	104	380	271	193	208	153	221	460	431
	45621	25019	118930	57910	43855	18932	43498	23623	22786	23737	45726	10245	26364	12722	15849	9388	11882	17360	33076	25367
2	*PF*	MI	CR, MI	*PF*	BG	BS, BG	BS, CR	BS	*PF*	AR, FI	FI, PI, LI	LI, PI	PG	PG	PG	VT	TR	*PF*	RM	FR, RM
	156	519	154	235	311	77	161	302	380	137	128	107	303	48	90	216	210	614	461	208
	149699	134827	17492	97532	77717	22968	15704	41786	304495	19942	30008	11820	45478	4389	5463	18107	17614	164669	58580	25165
3	VA, MI	*PF*	CR	PV	PV	CR	MN, BS	*PF*	AR	AR	LI	PI	LI	SI	GR	TR	*PF*	AQ	FR	FR
	122	508	151	299	195	234	66	320	202	130	144	118	79	149	178	138	197	163	189	201
	42119	303199	10682	38056	14386	42877	8004	122975	41285	9263	13353	10449	4274	10949	10429	6243	42373	15585	27989	21610
4	VA	VA	PC	PC	PR	*PF*	MN	MN	AR	RE	PT, FI	PI, LI	SI, FI	LU	SI, GR	IS	AQ	AQ	*PF*	FR
	234	192	156	268	195	215	287	163	187	183	147	223	122	133	192	106	187	138	199	170
	24219	43641	11137	48630	21490	59701	36273	11242	17616	54573	41365	42866	20591	39092	29770	11708	12929	7816	100897	15720
5	CO	CO	NO	NO	PR	MN, TN	TN	*PF*	MO	RE, MO	RE	LU, RE	LU	LU	CB	IS	PE	TE, PE	LT, FR, RM	LT
	200	219	232	169	204	15	156	124	209	89	183	10	127	152	198	74	138	72	102	210
	31663	57739	20526	21570	15509	2285	22387	37484	52837	18749	27910	1560	11057	19710	11749	3735	26601	7549	10928	22932
6	SO	SO	NO	NO, MI	*??*	TN, VR	TN	TN	MO	*PF*	RE, PR	SP	MS, LU	*PF*	CB	CH	PE	PE	TE	LT
	230	182	161	91	9	32	212	177	191	225	58	141	25	277	214	56	112	128	166	173
	12884	16401	9753	9823	574	2802	13695	15717	23693	131098	13228	11721	2635	70664	24944	7830	9427	8001	8645	10447
7	VI	*PF*	*??*	VR	VR	TN, BZ	BZ, TN	TN	BO	FE	FE	MS	MS	SP	*PF*	*??*	TE	TE	PE, CH	*PF*
	320	283	47	248	202	104	39	285	314	296	187	154	131	105	876	260	145	186	60	164
	60047	84519	6406	34194	15903	7515	2661	26177	66150	80379	20515	17735	11435	7318	1431304	39120	25018	14065	6782	96678
8	VI	VI	VI, VR	VR	VR	BZ	BZ	BZ	BO	*PF*	*PF*	RA, FO	*PF*	SM	PS	*PF*	AN	AP	CH	CH
	203	307	86	256	263	74	561	163	252	321	156	110	243	22	287	705	179	246	226	155
	15548	38083	10143	26853	24769	4766	39402	12152	28998	170856	56578	21661	75236	1609	33901	640821	19037	21368	21680	10763
9	VI, PD	*PF*	*PF*	*PF*	*PF*	BZ	—	BZ, BL	*PF*	RA, BO	RA	FO	FO	PS	AN, PS	AN	*PF*	AP	CH	*PF*
	99	97	160	187	282	89	0	21	213	68	143	139	215	181	87	240	276	289	181	31
	17482	28775	47910	196453	103734	6363	0	1056	93130	11312	20831	12236	35325	11722	11988	17466	53609	41963	31541	8127
10	TV	TV	VE, PN	VE	VE	OR	CA	CA	*PF*	CZ	*??*	*??*	LE	LE	*PF*	*??*	AN, MC	MC	FG	FG
	177	246	136	208	256	144	233	153	96	90	28	1	54	167	158	2	93	180	388	250
	11277	36780	14305	39878	34139	16273	44961	13233	53654	6850	12759	77	14626	47687	130107	105	10488	11318	59375	22165
11	TV	TV, VE	TV, VE	VE	*PF*	SS	*PF*	*PF*	*??*	CZ	CZ	LE	LE	LE, TA	TA	TA	MC	MC	FG	FG
	282	122	186	159	135	220	121	112	4	252	84	297	152	54	178	183	263	122	130	330
	35045	24547	37111	12728	29657	24117	144065	70380	155	20598	9095	48864	11902	9361	15873	26442	29950	8092	9570	30001

Table 3. (*continued*)

Y ↓ X→	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
12	PN, TV 124 19603	PF 202 51959	*??* 182 29593	VE, PD 176 49220	SS 98 19024	SS 152 9851	NU 126 10614	NU 93 5415	CS, CZ 104 23036	CZ 431 65781	*PF* 276 79739	BR, LE 75 12470	*PF* 198 188960	LE 294 60148	TA 155 24192	BR, TA 66 18730	BR 207 23525	BR 159 30019	BA, FG 141 15650	FG, BA 148 31444
13	PN 153 11951	PN 253 23414	*PF* 121 44001	PD 109 8799	SS, NU 60 11965	*PF* 173 37888	NU 73 11883	GO, TS 75 3786	TS 162 12144	CZ 365 47377	PZ, CZ 44 5899	PZ 151 14104	BN 199 20534	BN 122 9557	*PF* 147 51798	RC 262 44824	RC 101 9489	BA, TA 135 39088	BA 536 120415	BA 532 81232
14	PN, UD 99 13253	UD 241 39234	PD 363 67227	PD 284 34643	*??* 288 57525	GO 118 6735	TS 211 12534	TS 195 9461	*??* 3 143	CZ 236 21053	MT 215 24460	PZ 272 27044	PZ, MT 301 53179	BN, NA 189 36926	RC 248 56348	RC 152 15993	RC 91 7529	*PF* 322 78488	BA 106 12753	BA 332 29518
15	UD 291 16369	GO, UD 114 7481	*PF* 287 69543	RO 244 28052	TS, GO 177 9693	BL, TV 67 9166	GO 91 7845	BL 161 16730	TS 4 119	*PF* 56 18554	*PF* 282 57220	MT 121 7564	AV 239 24564	AV 146 10175	*PF* 38 6451	RC, ME 56 21982	CZ, RC 115 28887	*PF* 386 265500	*??* 41 14869	*??* 1 113
16	UD 443 36922	UD 347 31144	SV 124 9033	SV 156 23172	RO 76 4169	BL 233 23237	BL 123 9660	— 0 0	SA 140 8370	SA 250 37961	*PF* 266 85737	EN 208 19915	AV, NA 109 21270	*PF* 226 93889	*PF* 240 52548	RG 202 34225	RG 182 55987	CL 182 13059	AG 261 33586	AG 145 9351
17	AT 121 11783	AT 191 29633	IM 98 7668	VC 109 8634	VC 160 10743	*PF* 118 20690	*??* 5 236	— 0 0	SA 293 34356	SA, NA 138 28402	*PF* 558 1114381	*??* 167 22116	NA, CE 133 34401	CE 155 11101	CE 208 36602	RG, SR 71 13407	SR 186 39817	CL 268 43264	*PF* 311 91926	TP, AG 26 4508
18	AT, TO 169 17066	*PF* 148 15447	IM 86 13124	VC, TO 127 10458	VC 164 9264	VC, AO 4 219	AO 223 10915	AO 69 4227	GE 165 33100	*PF* 503 552248	*PF* 181 106103	*PF* 511 235331	CE, NA 106 19880	CE 192 25519	CE 59 3569	SR 131 9930	SR 172 19947	EN 101 5102	TP 195 30831	TP 176 25553
19	TO 391 37070	*PF* 357 49465	*PF* 355 115226	*PF* 251 92441	*PF* 145 25557	*PF* 227 50504	AL 206 21900	AL, GE 107 20898	GE 193 49245	NA 72 2721	NA, SA 153 57792	NA 447 142648	ME 34 3185	ME 292 42411	*PF* 251 145906	SR, CT 79 11494	*PF* 356 179852	*PF* 453 222590	TP 171 63786	TP, PA 117 21030
20	TO 443 25940	TO 495 39104	CN, TO 234 40487	CN 301 44675	CN 235 24089	CN 188 27854	AL 118 6795	AL, GE 91 10478	GE 195 19391	*PF* 190 68168	NA 521 119752	NA 384 33752	ME 266 53401	ME 182 16676	ME, CT 130 22933	CT 269 59904	CT 325 90222	PA 469 85313	PA 457 64931	PA 303 23564

a. In the first line of each cell, we report the probable geographic origin of the corresponding cluster of surnames by province (see Table 1 for the codes) after visual inspection of the corrected frequency gradients plotted in the 400 geographical maps of Figure 2 A/B/C/D. The number of different surnames per cluster is reported in the second line. The third line corresponds to the number of telephone subscribers sharing one of such surnames. When no geographic origin can be inferred, either because the cells correspond to polyphiletic surnames or because the pattern is not clear, the abbreviations "PF" and "??" are respectively reported instead of province codes. Main Italian macroregions are surrounded by a thick solid line (Northern Italy, Central Italy, Southern Italy and Sicily, Sardinia). Cells to which the text refers can be identified by their X and Y coordinates. To increase readability, the X axes run downward.

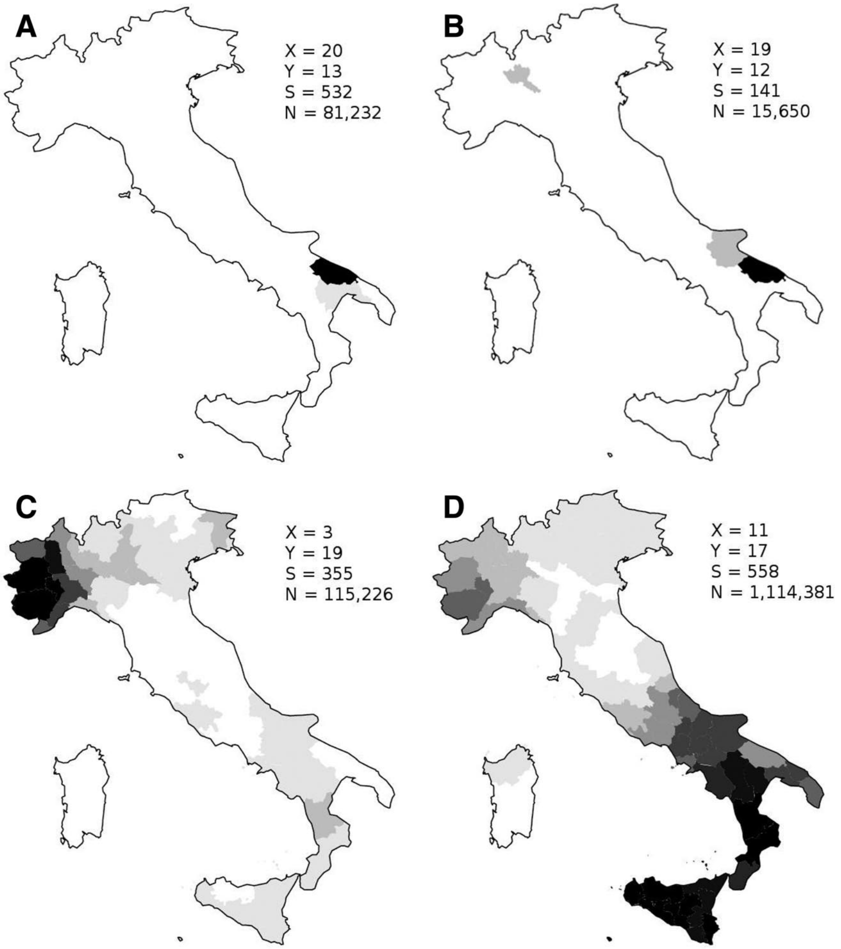

Figure 3. Enlargement of some of the 400 maps displayed in Figure 2 A/B/C/D. In the two maps at the top (corresponding to those visible in the previous figure at the position X = 20, Y = 13 and X = 19, Y = 12) are displayed two clusters of surnames having an identical peak of frequency (province of BA-Bari). While the two patterns are slightly different and concern a different number of surnames types (532 and 141, respectively), no ambiguity exists on their origin around Bari. Surnames with a single origin are defined as monophiletic. In the two bottom maps (C and D at the position X = 3, Y = 19 and X = 11, Y = 17 of Figure 2) we show clusters of surnames whose original geographical origin cannot be assessed because their frequency peak falls in a wide area. Such surnames are called polyphiletic, that is having had multiple independent origins. (C) concerns regional polyphiletic surnames (Piedmont); whereas (D) shows macroregional polyphiletic surnames (all southern Italy besides Sardinia). Regional polyphiletic surnames are quite frequent in Italy and point to long-lasting regional socio-cultural identities and dialects. Polyphiletic cells are marked as "*PF*" in the first line of the cells of Table 3; otherwise the province of origin is reported according to the codes of Table 1. S is the number of surnames clustered in the SOM cell and N is the total number of individuals bearing one of the S surnames.

matrix is somewhat complementary, it tells from which provinces the migratory flow to a given province took place.

Analytically, we define

$$dF = 1 - f_{ii} \tag{3}$$

$$dB = 1 - b_{ii} \tag{4}$$

where f_{ii} and b_{ii} are the diagonal elements of the *F* and *B* matrices, respectively. In this way, *dF* and *dB* coefficients provide information about the dispersal outside the province of origin of given surnames (male lineages) and of the amount of immigrants to that province (Table 1). We invite the reader to pay careful attention to the definition of the *dF* and *dB* indexes as they will be extensively mentioned throughout the article.

Province-specific migration patterns were estimated according to the nondiagonal values of the *M* matrix and geographic distances between couples of Italian provinces were prepared in a square distance matrix (*G*).

Mean emigration distances were calculated as:

$$d_i = \sum_i (m_{ij} \cdot g_{ij}) / \sum m_{ij}, \quad i \neq j, \tag{5}$$

and mean immigration distances were obtained as:

$$d_j = \sum_j (m_{ij} \cdot g_{ij}) / \sum m_{ij}, \quad i \neq j, \tag{6}$$

where m_{ij} are the elements of the *M* matrix and g_{ij} are the elements of the *G* matrix.

To clarify the meaning of *dF* and *dB* parameters and illustrate their possible combinations, we invite the reader to consult Table 2.

Control Method: The Frequency-Based Approach. The ways to analyze Italian family names (processed as genetic markers) and the methods to infer their probable geographic origin have been widely published by some of us (AL, OF, GZ) (Piazza et al. 1987; Zei et al. 1983, 1993, 2003). The first studies, about the geographic structure of Sardinian surnames, have shown that more than two-thirds of them were scattered on very small areas around the highest frequency locations. As a consequence, these surnames were classified as monophyletic. Additional linguistic data confirmed or improved the results obtained by geographical analysis. The experience acquired on Sardinian data allowed one to better address peninsular Italy.

Two databases were available: (1) the surnames of husbands and wives reported in 540,000 consanguineous marriage acts celebrated from 1910 and 1970 in the 280 Italian dioceses and (2) the list of telephone subscribers of the year 1993 (already mentioned) that accounts for 332,525 different surnames whose frequencies were available by commune, province, and region. From their distribution in these administrative subdivisions, their dispersion and attribution to a probable area of origin were achieved.

Surnames whose frequencies were higher than 50% in a region and, within that region, higher than 50% in a given province were copied in a new database meant to include surnames highly likely to be monophyletic and autochthonous of that province. Such inference was validated by taking into account the geographic distribution of surnames provided by the acts about consanguineous marriages.

Further, a third crosscheck of the reliability of the inferred geographic origin of surnames consisted of the analysis of the orographic, historical, and linguistic features of Italian provinces as they were likely to reveal the existence of possible genetic isolates inside the artificial administrative boundaries of the provinces (database copyright of the Italian scientific institution *Consiglio Nazionale delle Ricerche*, CNR).

In conclusion, 49,117 surnames were consistent with a monophyletic origin and were traced back to a reliable geographic origin. The results of the General Method were compared to such 49,117 surnames.

Results

Reading the Map. As we mentioned in the Methods, the cells of the SOM (Figure 2 A/B/C/D, Table 3) that correspond to monophyletic surnames exhibit a geographic distribution whose maximum frequency is located inside one single province, whereas polyphyletic surnames have wider geographical origins. Given that provinces are geometrically complex areas and not immaterial sample points, it is perfectly conceivable that a given cluster of monophyletic surnames may have its geographical and historical origin in an area located across the border of two or more neighboring provinces. For this reason, we considered clusters of surnames to have had a monophyletic origin even when their peak of frequency concerned two or three neighboring provinces. In such case, we attributed the "shared" cluster to all the involved provinces with equal probability. For example, if the geographic origin of the surnames in a given cell of the map appears to be in two neighboring provinces (let's say Pavia and Milan), we randomly partitioned such surnames in two sets accounting for 50% of them each and attributed, following the example, one to Pavia and the other to Milan. This step may not sound orthodox, but the bias introduced is of little influence as the geographic attribution still remains within a radius of 50 kilometers from the real geographic origin that we could not very precisely assess. This aspect will be further discussed in the last section of the article about the limitations of the General Method. Concerning the map, on average, there are 3.28 clusters (cells) corresponding to the same province, and the inspection of corresponding maps reported in Figure 2 A/B/C/D shows that the geographic distribution of some surnames belonging to a same province can exhibit different patterns (Figure 3). This phenomenon points to the existence of geographical substructures that we are not addressing; they are related to the political nature of the borders of many provinces. Interestingly, the general topology of SOM clearly mirrors the geography of the Italian peninsula as the *x*-axis corresponds to a North-South direction (Figure 2 A/B/C/D).

Monophyletic versus Polyphyletic Surnames. 315 of the 400 cells of the map displayed in Figure 2 A/B/C/D (78.75%) correspond to 58,906 family names (76.06% of the total) that, in turn, correspond to 7,713,027 telephone subscribers (43.87% of the telephone users accounted for by the database). As a consequence, a same monophyletic surname is shared, on average, by 131 ± 82 telephone users. 62 of the 400 cells of the map (15.5%) are clusters of polyphyletic surnames. They correspond to 16,307 family names (21.05% of the total) and identify 9,488,993 telephone users (53.98% of the total), meaning that 582 ± 345 individuals share each of them, on average. Finally 20 cells, corresponding to 2,238 surnames shared by 377,871 individuals (2.15% of the total), show geographic distributions incompatible both with a monophyletic and a polyphyletic surname origin (e.g., they show several peaks of frequency in different regions). Such "dubious" clusters were excluded from following analyses. A posteriori we find this bias of little importance. As examples, we invite the reader to check the cell X = 9, Y = 11 (Figure 2C), where three peaks of frequency occur in the provinces of CA-Cagliari, AT-Asti and CZ-Catanzaro, and the cell X = 7, Y = 17 (Figure 2C), where two peaks of frequency occur in the provinces of AO-Aosta and BL-Belluno. To conclude, three cells (X = 7, Y = 9 and X = 8, Y = 16 in Figure 2A; X = 8, Y = 17 in Figure 2C) are empty, meaning that there are no input-vectors (surnames) linked to them. More details are reported in Table 3.

Differently from The Netherlands (Manni et al. 2005), Italian polyphyletic surnames are rather regional. Some of them cover wide areas like the whole North (X = 15, Y = 7 in Figure 2B), the South (X = 11, Y = 17 in Figure 2D), or the Centre (X = 16, Y = 8 in Figure 2B). Other polyphyletic surnames are distributed within single regions like Lombardy (X = 2, Y = 3 in Figure 2A), Piedmont (X = 3, Y = 19 in Figure 2C), Sicily (X = 18, Y = 19 in Figure 2D), or Sardinia (X = 7, Y = 11 in Figure 2C).

If more than half of the Italians might bear polyphyletic surnames (53.98% according to the database analyzed), this percentage is geographically highly variable (Table 1), ranging from the 21.92% of BZ-Bolzano (Northeast) to the 70.94% of OR-Oristano (Sardinia). Noteworthy, the highest rate of surnames with a polyphyletic origin is found in Sardinia, whereas the lowest percentages are found in provinces located near the border dividing Italy from neighboring countries (AO-Aosta 34.11%, TS-Trieste 30.92%, GO-Gorizia 29.74%).

Validation of the Methodology. The main purpose of this article is to validate the General Method to Unravel Ancient Population Structures Through Surnames (General Method), which, by analyzing the present-day frequency distribution of family names, enables the detection of their historic geographic origin.

When the General Method was first applied to Dutch material (Manni et al. 2005), the results were compared to historical *census* data. The number of surnames having had their origin in the different provinces was found highly correlated to almost contemporary census data (beginning of the 19th century). Such agreement was considered sufficient to validate the Self-Organizing approach. Unfortunately, the same *census*-based validation cannot be applied to Italian surnames as the unification of many preexisting independent Republics and Kingdoms gave birth to the Italian nation two or three centuries later than the spread and adoption of surnames, meaning that older census reports are uneven in time and space because each pre-Italian State had its own way to collect the data. Therefore, to validate the present analyses, we adopted a different approach, that is to measure the overlap existing between (1) the number and geographic origin of the monophyletic surnames obtained by the SOM algorithm and (2) the results independently and previously obtained with the supervised method described in the last section of the Methods and called *The frequency-based approach*. With the General Method, 59,006 probable monophyletic surnames were identified, whereas the frequency-based approach yielded 49,117 of them. This discrepancy can be largely explained, considering that 11,305 surnames were not processed in the frequency-based approach because their frequency, inside a single region, was lower than 50% (see last section of the Methods). With respect to this restriction, we note that one of the advantages of the General Method over the frequency-based approach is that arbitrary frequency thresholds, like the cutoff at 50% just mentioned, are not needed.

To come to the validation, we think that a better way to compare the results of the two methods is to say that 47,761 of 49,117 surnames listed as monophyletic and specific of a given province with the frequency-based approach were equally predicted to be so with the General Method (97% overlap).

Population Mobility

Emigration (dF). By knowing where surnames started to be in use and looking at their present-day distribution, we can measure the emigration rate (*dF*) of the descendants of the Medieval/Renaissance Italian population in subsequent times until the year 1993 (our database of telephone subscribers is updated to that year). In Table 1 we report, province by province, *dF* values that are visually displayed in Figure 4. Low values of *dF* indicate that a large proportion of those surnames that are autochthonous of a province remain today, in their majority, where they were adopted centuries ago, thus implying that the male descendants of the Middle Ages/Renaissance stock of the population did not emigrate much in the following four centuries. This is the case of many provinces located in the northern part of the country (BZ-Bolzano 22.14%, AO-Aosta 25.08%, TN-Trento 38.40%, PD-Padova 38.40%), of big attractive metropolises (TO-Turin 31.70%, RM-Rome 35.99%, MI-Milan 38.26%) and of CA-Cagliari (31.39%) in

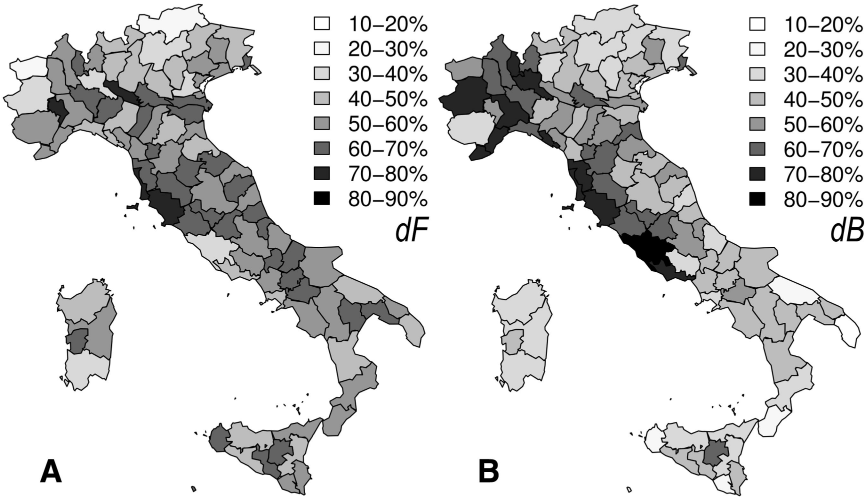

Figure 4. Major results of the General Method plotted by province according to Table 1. *dF* values (A) correspond to the number of telephone subscribers having a surname autochthonous of a province and living in it. Low percentages mean that the autochthonous remained in the area and, vice versa, high percentages correspond to massive emigration of them. *dB* values (B) correspond to the number of individuals bearing, within each province, a surname that comes from other provinces (non-autochthonous). High values correspond to major Italian cities like TO-Turin, MI-Milan, RM-Rome, the cost of Tuscany, and a part of Liguria (see Figure 1 for geographical details). Southern Italy has not been the destination of many immigrants. *dF* and *dB* values correspond to different aspects of the contemporary Italian population, and a synthesis can be obtained by considering them together as is visible in Figure 6. As *dF* values just tell if the descendants of the original population emigrated outside the province of origin but do not say how far they went, we provide additional detail in the third map (C) where mean emigration distances are reported by province. The figure shows that long-distance emigration concerns southern Italy (including Sardinia and Sicily), whereas emigrants from northern and central provinces did not go very far. More detailed evidence is provided in Figure 5. In the last map (D) we provide a migratory balance showing which provinces increased their surname diversity and population and which ones lost population diversity and surnames as a consequence of emigration to other areas. Piedmont, Liguria, TO-Turin, MI-Milan, BO-Bologna, FO-Florence, and RM-Rome attracted many immigrants, the other areas did not. We note that none of the phenomena displayed in (A), (B), (C), and (D) can be dated; they took place between the establishment and spread of surnames, at least four centuries ago, and present times as sandwich of undefined phases that the General Method cannot help to identify. Provinces codes are reported in Figure 1.

Sardinia. The opposite phenomenon (high *dF* values, see Table 1) implies high emigration rates of the descendants of the original population, like in Tuscany (in particular LI-Livorno 77.23%, GR-Grosseto 72.94%, AR-Arezzo 69.82%, PI-Pisa 67.33%, SI-Siena 67.23%) and some provinces located in the valley of the river Po in northern Italy (e.g., AT-Asti 73.19%, CR-Cremona 72.48%, MN-Mantova 69.57%). Nevertheless, the high values of *dF* do not tell anything about the range of emigration distances over the time, as emigration movements could have been very local and directed to neighboring provinces or of a wider range

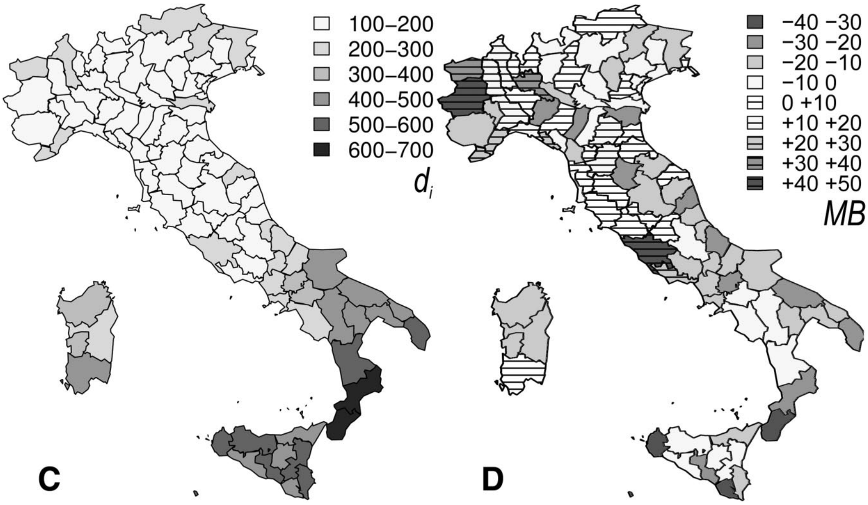

Figure 4. (*continued*)

and directed to very distant regions. To provide a better insight, the mean emigration distance (d_i) from each province is reported in Table 1 and visualized in Figure 4. At first glance, it appears that emigrants from southern Italy and Sicily migrated much farther than emigrants from other areas of the country. A more refined analysis is obtained by dissecting the whole migratory outflow from each province by distance classes (0−100 km; 101−200 km; 201−300 km; 301−400 km; 401−500 km; 501−600 km; 601−700 km; 701−800 km; 801−900 km; 901−1,000 km; 1,001−1,100 km; 1,101−1,200 km—data not presented). By transforming, province by province, the number of migrants in each distance class into the elements of a vector (12 elements in total, one for each distance class), we can summarize their variability in a Principal Component Analysis (PCA) plot (Figure 5). This analysis just gives general trends as the geographic location of the different provinces in the Italian peninsula influences the respective upper bounds of emigration distance. For example, from the very North, it is possible to migrate to the very South by traveling 1,200 km, but from a province located in the center of the peninsula, the maximum displacement is about 600 km. Another limit of the analysis is related to provinces located in Sardinia as some emigration-distance classes remain empty because they correspond to the sea. Also, we were unable to take into account international migrations as we did not analyze foreign databases of surnames, meaning that emigration concerning provinces located at the northern borders of Italy is likely to be underestimated. Even with such proviso, the plot of Figure 5 suggests the existence of three main emigration patterns that are described below.

(1) Isolation-by-Distance. Such pattern is typical of provinces from which emigration took place on a local scale (most frequent distance class: 0−100 km)

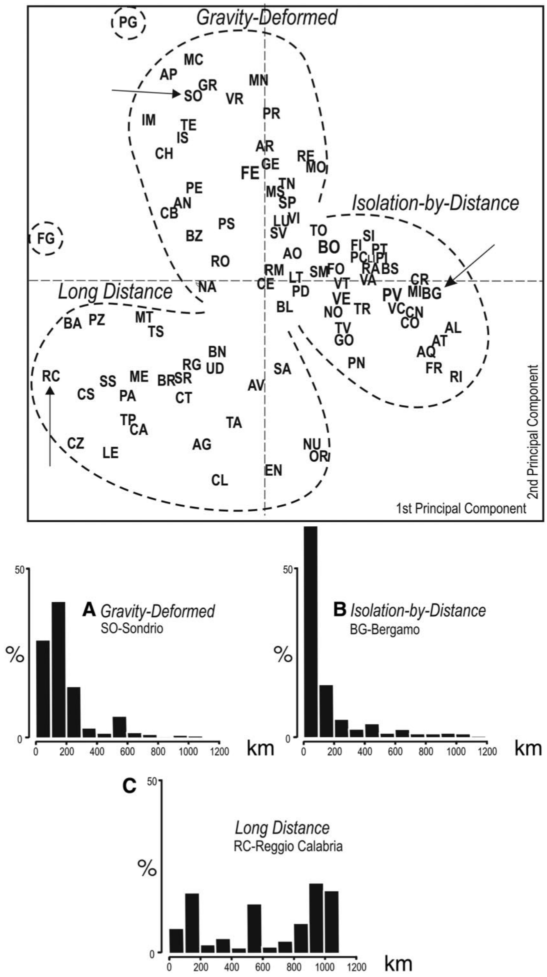

Figure 5. By comparing the present-day distribution of Italian surnames to their inferred origins (see Table 2), we can dissect emigration from each Italian province by distance classes. To this end, we ranked emigration movements by distance class (0−100 km, 101−200 km, etc.) and input such vectors in a Principal Component Analysis (PCA) plotted in the upper part of the figure. Three major clusters corresponding to three patterns are visible: *Gravity-Deformed* (A) *Isolation-by-Distance* (B), and *Long Distance* (C). At the bottom of the figure we have represented three provinces that well represent such patterns, respectively SO-Sondrio, Bg-Bergamo, and RC-Reggio Calabria. From the provinces belonging to the (A) cluster, emigration took place to areas more distant than the immediate vicinity, similarly to orbit deviations of a satellite caused by another planet. From the provinces of the (B) cluster, emigration was mainly directed to very close areas. The last cluster (C) corresponds to long-distance emigration and concerns southern provinces. This pattern is likely to be recent and related to the mechanization of transports. More details are reported in the text as further interesting geographical patterns are found within clusters. Fully spelled names of provinces are reported in Table 1.

with a general motif corresponding to a negative exponential decrease of emigrants with the increase of the distance as it is expected under a diffusionist isolation-by-distance hypothesis. Such pattern exclusively corresponds to provinces located in Northern and Central Italy, that is, to the regions called Piedmont and Lombardy, to a part of Triveneto, to the central part of Tuscany, and to provinces located between Latium and the Adriatic coast (see Figure 5, chart B).

(2) Gravity Deformed. This pattern is similar to the previous one, but here the second or the third distance class of emigration (101−200 km, 201−300 km) prevails over the first one (0−100 km), suggesting that emigrants moved a little further than their immediate neighborhood because they were attracted by an interesting destination (e.g., a rich city). Such "gravity-deformed" cluster largely corresponds to almost all the provinces of Northern and Central Italy that do not belong to the first group (see Figure 5, chart A).

(3) Long-Distance Emigration. This pattern is characterized by emigration movements that totally contrast with the isolation-by-distance scenario as the majority of emigrants moved to very distant provinces. Besides two northeastern areas (like TS-Trieste and UD-Udine), this pattern concerns all the provinces south of Latium, including Sardinia and Sicily (Figure 5, chart C).

Immigration. Immigration rates toward a province (*dB*) are reported in Table 1 and plotted in Figure 4. The lowest values, meaning that very few immigrants moved to a given province, concern southern Italy (LE-Lecce 19.00%, RC-Reggio Calabria 23.59%, RA-Ragusa 25.79%, BA-Bari 26.84%, TP-Trapani 28.89%) and northeasterner areas near the Alps like BZ-Bolzano (30.14%) and UD-Udine (32.42%). Conversely, the provinces that attracted the highest number of immigrants are the capital city (RM-Rome 82.02%), the two major industrial metropolises of the North (MI-Milan 75.07% and TO-Turin 73.51%), the area encompassing the latter two cities (VA-Varese 70.78%, AL-Alessandria 71.65%, VC-Vercelli 69.60%, NO-Novara 67.40%, PV-Pavia 68.36%), and, finally, the Tyrrhenian coast (GR-Grosseto 79.86%, IM-Imperia 77.00%, SV-Savona 76.33%, Livorno 75.69%, PI-Pisa 72.28%, SP-La Spezia 71.64%, SI-Siena 68.69%). The high *dB* of the province of Latina (70.95%) just mirrors the colonization of newly reclaimed lands from 1932 onward and, therefore, does not deserve a special discussion.

Synthesis of Migratory Flows in Italy. By subtracting *dB* from *dF,* it is possible to compute the overall migratory balance (*MB*) that accounts for all the migrations that occurred in the last four centuries in Italy. Provinces with a positive *MB* gained surname diversity (and population) from outside over this long time (immigration exceeded emigration), whereas negative *MB* values indicate the opposite phenomenon. *MB* values are summarized in Table 1 and geographically plotted in Figure 4. The provinces of RM-Rome (+46.03), TO-Turin (+41.81), and MI-Milan (+36.81)

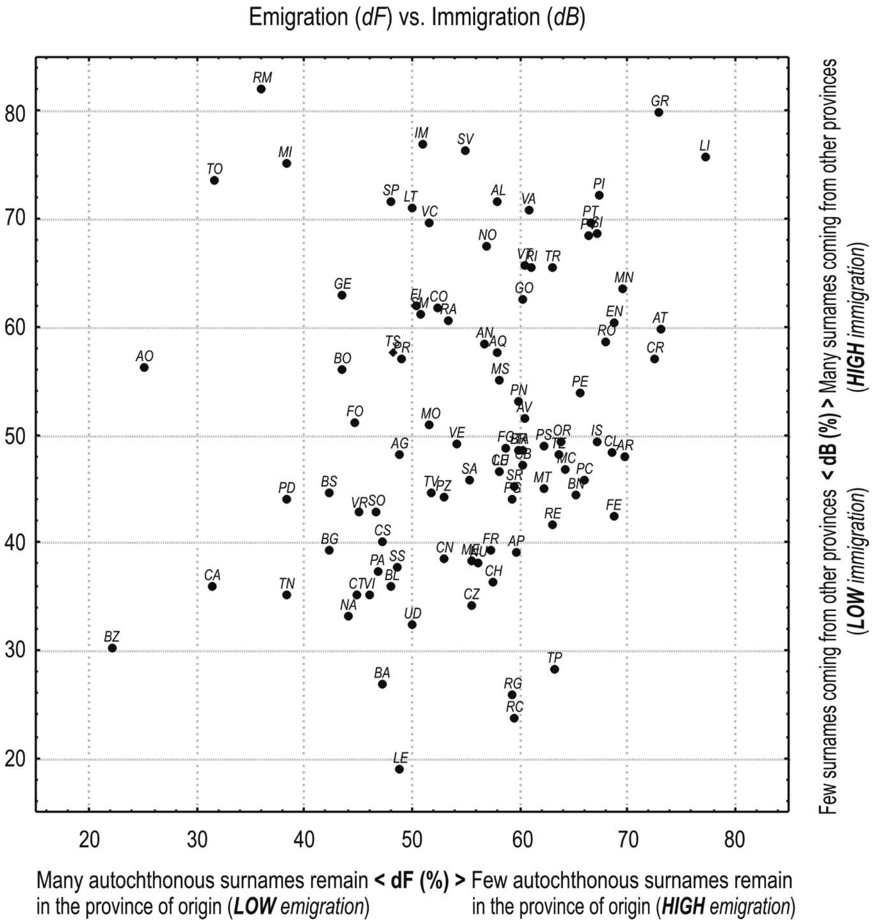

Figure 6. Bidimensional plot of *dF* versus *dB* values, by province, according to Table 1. A separate visual analysis of *dF* and *dB* is provided in Figure 4. By interpreting the plot according to Table 2, it is possible to identify *autochthonous*, *isolated*, *attractive,* and *"corridor"* areas. Fully spelled names of provinces are reported in Table 1.

gained much population. Other provinces with positive MB are mainly located in the Northwest: most notably AO-Aosta (+31.23), Liguria (IM-Imperia +26.05, SP-La Spezia +23.62, SA-Savona +21.46, GE-Genova +19.43), FI-Florence (+11.56), and BO-Bologna (+12.41). Negative *MB* values are mostly found in southern Italy: RC-Reggio Calabria (–35.97) and TP-Trapani (–34.95) being the lowest. Other areas characterized by generally negative values are the Adriatic coast (CH-Chieti –21.21, AP-Ascoli Piceno –20.70, MC-Macerata –17.21) and part of the river Po valley (FE-Ferrara –26.47, RE-Reggio Emilia –21.33, PC-Piacenza –20.24).

Another way to capture the essence of Italian migration flows is to look at Table 2 and Figure 6, where provinces are plotted on two axes that correspond to the values of *dB* and *dF*. As we noted, different areas can be classified as (1) *unattractive* (many emigrants/few immigrants), (2) *attractive* (few emigrants/many immigrants), (3) *corridor* (many emigrants/many immigrants), and (4) *self-sufficient/isolated* (few emigrants/few immigrants). If a biaxial *dB* versus *dF* plot (Figure 6) shows a rather continuous topology that makes difficult the use of schematic categories, we can classify Rome, Milan, and Turin as *attractive* areas, being the provinces located between the last two cities passageway areas. This is not unexpected as Milan and Turin are located rather close to each other, and migrations from one city to the other have probably been very frequent, especially in modern times, given that their industrial areas overlap. Similarly, the entire coastline from Rome to France can be defined as a *corridor*. Clearly, *unattractive* areas are located in the very South (TR-Trapani, RG-Ragusa, RC-Reggio Calabria), while *self-sufficient/isolated* provinces correspond to the Trentino-Alto Adige (TN-Trento, BZ-Bolzano), a region located in the very North in a mountainous area, and to the regional capital of Sardinia CA-Cagliari. To conclude, we note that the group that geographically makes more sense, and has the highest number of provinces belonging to it, is the *corridor* cluster. We find the plot of Figure 6 more satisfactory than Figure 4 because, while not hiding the complexity of the results, it provides a higher level of synthesis.

Discussion

Advantages and Limitations of the General Method. The purposes of this article are (1) to infer the geographic origin of a vast majority of Italian surnames; (2) to compare SOM-based results with those independently obtained by some of us (AL, OF, GZ) through a supervised method; (3) to distinguish polyphyletic and monophyletic surnames, and finally (4) to depict migrations that occurred in Italy since the times of surname introduction. Let us start from the second point. As we have shown, the overlap of results obtained with the General Method and those obtained with the frequency-based method is 97%. We already pointed out that the SOMs-based General Method, being an unsupervised heuristic method, is occasionally prone to different errors:

1) Polyphyletic family names can be recognized as monophyletic. An example is related to surnames given to abandoned children in specific areas (like "Esposito" in Naples or "Martinetti" in Milan) that can be grouped with monophyletic surnames specific of Naples and Milan even though they concern unrelated individuals.

2) Monophyletic surnames characterized by low frequencies and a sparse geographic distribution can be clustered in the SOMs output among polyphyletic family names. To minimize the error, as mentioned in the methodological section, we excluded the surnames whose absolute frequency in the database was lower than 20 occurrences. Anyway, following a conservative approach, we considered "dubious" all the clusters of surnames characterized by frequency peaks in different regions. They correspond to a very limited number of individuals (2.15% of the total) whose surname has an equal probability of being monophyletic or polyphyletic. We are

reasonably confident that their exclusion does not significantly bias any of the results discussed in this article as the distribution of such surnames is quite random. Concenring monophyletic family names, drift and founder effect phenomena can interfere with the attribution of a correct geographic origin to a surname. When families split into two or more branches that establish in different places, it can happen that the original branch dies out and that the geographic origin of that surname is assigned to the area where the branch survives.

To come back to the validation of the General Method, we can consider that the 3% discrepancy between our unsupervised results and those emerging from the supervised approach accounts for the sources of error listed above. Such discrepancy is negligible, and we conclude that the General Method outputs correct and reliable results. The General Method can be generalized to other areas without the need of further validations.

Concerning Italy, population mobility is unquestionably the main force that shaped the current surname distribution. Drift and founder effect phenomena, disturbing factors that are typical at micro-geographic scales or when the population size is low (Darlu et al. 2001), are not detectable. The explanation probably relies on our methodology: instead of smaller operational units, we have analyzed whole provinces (in general rather populous), and the SOMs analysis was preceded by a weighting procedure taking into account the size of each province. The latter caution has minimized drift and founder effects and decreased the probability of erroneous geographic attributions. Nevertheless, a weakness related to the use of whole provinces is that surnames whose geographic origin is located at the borders of two (or three) neighboring ones cannot be confidently attributed to one of them (13.45% of the total number of surnames, that is 7942). If our choice to randomly attribute a half (or a third) of them to each neighboring province still provides reasonably good evidence of their original location, it interferes with the exact estimation of the migratory flows visually described in Figure 5.

Polyphyletism. Polyphyletic surnames are confusing as they lead to overestimated levels of co-ancestry and hamper studies meant to associate specific Y-chromosome lineages to given surnames. We show that polyphyletism is a major phenomenon in Italy, as more than half (53.98%) of the whole population sample bears polyphyletic surnames. This rate is 2-fold when compared to the one measured in The Netherlands with similar methods (Manni et al. 2005). Another peculiarity of Italy is that some of the most frequent polyphyletic surnames can have a regional diffusion. The regional polyphyletism of Italy implies that even surnames with multiple origins can convey a certain geographic signal, thus mirroring the existence of distinct cultural areas that may be related, or reflected, by the different dialects spoken in the country and to its political history. As already mentioned, surnames are a specific part of language, and a same surname with a general meaning like "peasant" can be spelled "Massai" in the South and "Campagnolo" in the North, both being polyphyletic ones. This phenomenon is apparent in Sardinia, a region that has the highest proportion of polyphyletic surnames of all Italy and, interestingly, some of the most divergent dialects. Nevertheless, linguistic differences are insufficient to explain the observed regionalism of polyphyletic surnames, as many of them would be spelled almost in the same way all over the Italian peninsula, what very often does not happen. The fact that their distribution appears to be regional points to more complex socio-demographic and historical dynamics that go beyond the purposes of this paper (for a more general overview of them, see Darlu et al. 2012).

Mobility of Italian Populations since the Late Middle Ages: Geographic Overview. Our results, on average, suggest that less than a half of the Italians ($dF = 55.22\% \pm 10.63\%$) still live where their male ancestors were established at the time of surname introduction, that is at least four centuries ago. More in detail, large migratory flows from the South and, to a lesser extent, from the Northeast, were directed to the city of Rome and to the surroundings of Turin and Milan (check Table 1 and Figures 4 and 6). A large part of such migrations took place after the establishment of a unified Italian state (1861), when Rome became the capital city and industrialization started. Further, our results highlight major emigration from mountainous areas toward the plains with some exceptions shown by the negative migratory balance observed in a part of the Adriatic coast (CH-Chieti, AP-Ascoli Piceno, etc.). Negative values, unexpectedly, also concern the Po valley (FE-Ferrara, PC-Piacenza, RE-Reggio Emilia) and may be explained by demographic phenomena that occurred before the unification of Italy (1861).

Interestingly, regions like Tuscany and Latium show *dB* values as high as 79.86% (GR-Grosseto), meaning that a great part of their present-day inhabitants does not bear autochtonous surnames. The reverse situation can be found in the northeastern macroregion called Triveneto (formed by three regions named Veneto, Trentino Alto Adige, and Friuli Venezia Giulia; see Figure 1). As far as surnames are concerned, Triveneto appears to be among the most autochthonous areas of the country. In fact, those northeastern provinces were characterized by high-emigration and low-immigration rates with a migratory balance (*MB*, Table 1 and Figure 4) that is generally negative. We mentioned Triveneto because this region is characterized by a strong cultural cohesion mirrored by a certain self-assessed "identity" that, partly, relies on its well-preserved and lively dialects. According to a same self-assessed identity, a large proportion of the current population of Latium and Tuscany could have been expected to be largely autochthonous. For example, Tuscany is historically well-known for having been peopled by the Etruscans (an important pre-Roman Italian people) and has a special cultural prominence because its dialect was adopted in the 19th century as the official language of Italy. Despite such facts, our surname analyses deny any major "autochthony" to the population of Tuscany as shown by its high *dF* and *dB* values (Table 1; Figure 4, Figure 6). Tuscany, together with northern Latium and Liguria, has been a corridor area (Figure 6). We are stressing this finding because it has a special relevance on the genetic side.

Similarly to Triveneto, southern regions experienced massive emigration flows (see the negative *MB* values in Table 1 and Figure 4) which have not been counterbalanced by any significant immigration event—a result expected, considering the persistent economical weaknesses of this part of Italy. Very low *dB* values (Table 1; Figure 4) indicate that southern Italy is the most "surname-autochthonous" area of the whole country, and a contemporary random surname sampling of its population would largely mirror the Y-chromosome variability of the populations here established four centuries ago, if not before. For example, *dB* values tell that a random individual sampled in the province of RA-Ragusa, RC-Reggio-Calabria, or LE-Lecce would bear one of the surnames autochthonous of the region with a probability of 74%, 76%, and 81%, respectively.

Mobility of Italian Populations since the Late Middle Ages: General Patterns. The emigration patterns to other provinces that are mentioned in the Results section (*Isolation-by-Distance*, *Gravity Deformed*, and *Long Distance*—see Figure 5) represent a sandwich of phenomena that took place over four centuries with an unknown tempo, and unfortunately do not provide evidence for back-migration phenomena. We suggest that the *Isolation by Distance* patterns are the result of population diffusion that took place over a long time and largely correspond to the short migrations associated to marriages (Darlu et al. 2012). Differently, the *Gravity-Deformed* and *Long-Distance* patterns are probably related to the 20th century's social changes of the country (urbanization, industrialization, abandonment of mountainous and rural areas) and to the advent of mechanized transportation. Such documented massive migratory flows overlap with the migrations of previous times that very likely were shaped by Isolation-by-Distance phenomena (IBD).

Perspectives of Investigation

A deeper level of dissection of Italian surname structures is still possible as there are, on average, more than 3 clusters (cells) in Figure 2 corresponding to each province. We analyzed the data set at a provincial level (95 units plus the Republic of San Marino), but the number of operational units could be multiplied by a factor of 80 by addressing a smaller administrative division, that is the more than 8000 Italian communes (*Comuni*). This task is challenging both in terms of computational power and in terms of synthesis, as a number of microregional bottlenecks and founder effect phenomena will become apparent, together with an increased effect of geography. Once this tremendous task is completed, it will be possible to identify numerous isolates that, genetically tested, can lead to a more accurate molecular cartography of the Y-chromosome variability of the country. It goes without saying that, to provide a satisfactory discussion and interpretation of highlighted patterns, a strong level of integration between many anthropological disciplines is required.

In the same vein, we have great expectations from the results of a future Y-chromosome DNA study of the Italian population conducted on autochthonous surnames. The bearers of any surname typical of the area they inhabit have high chances to be a representative subset of the Middle Ages and Renaissance Italian population thus showing that the Y-chromosome variability of four centuries ago was dissimilar from the one observed today. As an example, the population of some parts of Tuscany is believed to descend from the Etruscans, an important pre-Roman people that was established in the region in the Iron Age. This is why Guimaraes et al. (2009) expected to find a strong Etruscan mitochondrial DNA signature in contemporary Tuscans but did not. By referring to ancient Medioeval DNA samples, they interpreted the lack of genealogical continuity as the result of *extensive demographic change occurred before* AD 1000. In agreement with such authors, our analyses on patrilineal surnames show that few inhabitants of Tuscany are "autochthonous" of the region. We suggest that the *extensive demographic change* mentioned by Guimaraes et al. (2009) about the female line, later (after the introduction of surnames) has concerned the male line as well, thus weakening the chances to detect any signature of a far past, including the Etruscans. The example above is reported only to highlight the wide variety of research questions the General Method may help to address.

To conclude, we note that the historical root of surnames does not allow conclusions prior to their origin in the late Middle Ages. Nevertheless and besides the potential application of ancient DNA methods, it is true that any Y-chromosome genetic structure inferred by using a sampling scheme derived from the results of the General Method will reveal, better than other methods, the genetic variability of the past.

Acknowledgments The lists of the surnames autochthonous of each province can be made available to third parties, for scientific purposes, upon request to the corresponding author. F. Manni would like to acknowledge the Wenner Gren Foundation (grant n° 7247) that financed a seminal and similar research and dedicates this article to Professor Italo Barrai (University of Ferrara, Italy). A. Lisa, O. Fiorani, and G. Zei were supported by a grant CARIPLO 2010–0253. A. Boattini and D. Pettener were supported by grants PRIN 2007 and 2009 to D. P. To avoid any possible conflict of interests, the editorial process has been independently secured by Dr. Pierre Darlu (CNRS, France).

Received 4 April 2012; revision accepted for publication 25 June 2012.

Literature Cited

Angius, A., P. M. Melis, L. Morelli et al. 2001. Archival, demographic, and genetic studies define a Sardinian sub-isolate as a suitable model for mapping complex traits. *Hum. Genet.* 109:198–209.

Bedoya, G., P. Montoya, J. García et al. 2006. Admixture dynamics in Hispanics: A shift in the nuclear genetic ancestry of a South American population isolate. *Proc. Natl. Acad. Sci. USA* 103:7234–7239.

Boattini, A., D. Luiselli, M. Sazzini et al. 2011. Linking Italy and the Balkans. A Y-chromosome perspective form the Arbereshe of Calabria. *Ann. Hum. Biol*. 38:59–68.

Boattini, A., M. E. Pedrosi, and D. Luiselli. 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.

Bodmer, W. F., and L. L. Cavalli-Sforza. 1968. A migration matrix model for the study of random genetic drift. *Genetics* 59:565–592.

Bowden, G. R., P. Balaresque, T. E. King et al. 2008. Excavating past population structures by surname-based sampling: The genetic legacy of the Vikings in northwest England. *Mol. Biol. Evol.* 25:301–309.

Cavalli-Sforza, L. L., A. Moroni, and G. Zei. 2004. *Consanguinity, Inbreeding, and Genetic Drift in Italy*. Princeton and Oxford: Princeton University Press.

Cheshire, J., P. Mateos, and P. A. Longley. 2011. Delineating Europe's cultural regions: Population structure and surname clustering. *Hum. Biol.* 83:573–598.

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

Darlu, P., G. Bloothoft, A. Boattini et al. 2012. The family name as socio-cultural feature and genetic metaphor: From concepts to methods. *Hum. Biol*. 84:169–214.

Darlu, P., A. Degioanni, and L. Jakobi. 2001. Les cloisonnement dans les Pyrénées occidentales. Évolution, du XIXe siècle à nos jours. In *Le Patronyme: Histoire, Anthropologie, Société,* G. Brunet, P. Darlu, and G. Zei, eds. Le Patronyme: Histoire, Anthropologie, Société. Paris: CNRS Editions, 173–187.

De Felice, E. 1980. *I Cognomi Italiani.* Bologna: Società Editrice il Mulino.

De Felice, E. 1982. *Dizionario Dei Cognomi Italiani.* Milano: Mondadori.

Graf, O. M., M. Zlojutro, R. Rubicz et al. 2010. Surname distributions and their association with Y-chromosome markers in the Aleutian Islands. *Hum. Biol.* 82:745–757.

Guimaraes, S., S. Ghirotto, A. Benazzo et al. 2009. Genealogical discontinuities among Etruscan, Medieval, and contemporary Tuscans. *Mol. Biol. Evol.* 26:2157–2166.

Hill, E. W., M. A. Jobling, and D. G. Bradley. 2000. Y-chromosome variation and Irish origins. *Nature* 404:351–352.

Jobling, M. A., and C. Tyler-Smith. 2003. The human Y chromosome: An evolutionary marker comes of age. *Nat. Rev. Genet.* 4:598–612.

Jobling, M. A. 2001. In the name of the father: Surnames and genetics. *Trends Genet.* 17:353–357.

Kaski, S. 1997. Data exploration using self-organizing-maps. *Acta Polytechn. Scand.* 82:1–57.

King, T. E., S. J. Ballereau, K. E. Schürer, and M. A. Jobling. 2006. Genetic signatures of co-ancestry within surnames. *Curr. Biol.* 16:384–388.

King, T. E., and M. A. Jobling. 2009a. Founders, drift, and infidelity: The relationship between Y chromosome diversity and patrilineal surnames. *Mol. Biol. Evol.* 26:1093–102.

King, T. E., and M. A. Jobling. 2009b. What's in a name? Y chromosomes, surnames and the genetic genealogy revolution. *Trends Genet.* 25:351–360.

Kohonen, T. 1982. Self-organized formation of topologically correct feature maps. *Biol. Cybern.* 43:59–69.

Kohonen, T. 1984. *Self-organization and Associative Memory*. Berlin: Springer.

Manni, F., W. J. Heeringa, and J. Nerbonne. 2006. To what extent are surnames words? Comparing the geographic patterns of surname and dialect variation in the Netherlands. Numero special de LLC Literary and Linguistic Computing "Progress in Dialectometry: Toward Explanation" 21:507–527.

Manni, F., W. Heeringa, B. Toupance et al. 2008. Do surname differences mirror dialect variation? *Hum. Biol.* 80:41–64.

Manni, F., and B. Toupance. 2010. Autochthony and HLA frequencies in The Netherlands: When surnames are useless markers. *Hum. Biol.* 82:457–467.

Manni, F., B. Toupance, A. Sabbagh et al. 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.

Martínez-González, L. J., E. Martínez-Espín, J. C. Álvarez et al. 2012. Surname and Y chromosome in Southern Europe: A case study with Colom/Colombo. *Eur. J. Hum. Genet.* 20:211–6.

McEvoy, B., and D. G. Bradley. 2006. Y-chromosomes and the extent of patrilineal ancestry in Irish surnames. *Hum. Genet.* 119:212–9.

McEvoy, B., C. Brady, L. T. Moore et al. 2006. The scale and nature of Viking settlement in Ireland from Y-chromosome admixture analysis. *Eur. J. Hum. Genet.* 14:1288–1294.

Mocci, E., M. P. Concas, M. Fanciulli et al. 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.

Moore, L. T., B. McEvoy, E. Cape et al. 2006. A Y-chromosome signature of hegemony in Gaelic Ireland. *Am. J. Hum. Genet.* 78:334–338.

R Development Core Team. 2008. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. www.R-project.org.

Rodríguez Díaz, R., and M. J. Blanco Villegas. 2010. Genetic structure of a rural region in Spain: Distribution of surnames and gene flow. *Hum. Biol.* 82:301–314.

Shlush, L. I., D. M. Behar, G. Yudkovsky et al. 2008. The Druze: A population genetic refugium of the Near East. *PLoS One* 3:e2105.

Sykes, B., and C. Irven. 2000. Surnames and the Y chromosome. *Am. J. Hum. Genet.* 66:1417–9.

Traglia, M., C. Sala, C. Masciullo et al. 2009. Heritability and demographic analyses in the large isolated population of Val Borbera suggest advantages in mapping complex traits genes. *PLoS One* 4:e7554.

Wehrens, R., and L. M. C. Buydens. 2007. Self- and Super-organising Maps in R: The kohonen package. *J. Stat. Softw.* 21(5).

Winney, B., A. Boumertit, T. Day et al. 2011. People of the British Isles: Preliminary analysis of genotypes and surnames in a UK-control population. *Eur. J. Hum. Genet.* 20:203–210.

Zei, G., G. Barbujani, A. Lisa et al. 1993. Barriers to gene flow estimated by surname distribution in Italy. *Ann. Hum. Genet.* 57:123–140.

Zei, G., and R. Guglielmino Matessi. 1983. Surnames in Sardinia. 1. Fit of frequency distributions for neutral alleles and genetic population structure. *Ann. Hum. Genet.* 47:329–352.

Zei, G., A. Lisa, O. Fiorani et al. 2003. From surnames to the history of Y chromosomes: The Sardinian population as a paradigm. *Eur. J. Hum. Genet.* 11:802–807.